Beyond Images: Fetal Ultrasound Video Understanding for Detecting Congenital Heart Diseases

Divyanshu Mishra

St. Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas Term 2025

हार नहीं मानूंगा, रार नहीं ठानूंगा,
काल के कपाल पर लिखता मिटाता हूँ
गीत नया गाता हूँ

– अटल बिहारी वाजपेयी

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Alison Noble, for her exceptional guidance and support throughout my PhD. When I began this journey, I approached research with the mindset of an engineer, focused primarily on technical challenges. Under her mentorship, I learned to think like a scientist: to formulate meaningful research questions, to pursue challenging problems with curiosity and persistence, and to appreciate the broader context of my work even while delving deeply into specific details. She consistently inspired me to tackle ambitious problems and provided unwavering support throughout the process. She encouraged me to pursue research that makes a real-world impact and to think independently, to take the road less traveled rather than simply follow the crowd, especially in the current fast-paced AI landscape.

Her supervision style provided me with great intellectual freedom; she never imposed strict deadlines but instead encouraged exploration and creativity in finding the best solutions. Our one-to-one meetings were not only about research progress but also about personal well-being. She often reminded me of the importance of balance, of taking time to enjoy life, cultivating hobbies, and not being disheartened when research did not go as planned. I am also grateful to her for allowing me to attend numerous conferences and summer schools, which not only broadened my research perspective but also gave me the wonderful opportunity to travel and connect with colleagues across the research community. Her encouragement and perspective have profoundly shaped both my academic and personal growth, and I will carry the lessons learned from her guidance wherever my path leads next.

I am deeply thankful to Professor He Zhao, who, during his time as a postdoctoral researcher, played a key role in shaping my early research journey. He guided me through my first paper submission when I was still unfamiliar with the many intricacies of academic research and continued to support me on several projects thereafter. I fondly remember the long brainstorming sessions in his office and our many hours spent debugging code together. His technical insight, patience, and enthusiasm for research greatly influenced how I approached problem solving and how I learned to present research with clarity and precision.

My sincere thanks also go to Professor Yuki Asano for his excellent technical guidance and insightful discussions during my research. His exceptional intuition

in technical subjects, ability to distill complex ideas with clarity, and preference for clean, elegant solutions over unnecessary complexity have greatly influenced my growth as a researcher. I truly appreciated his constructive feedback and the enthusiasm with which he approached every discussion. His mentorship has had a lasting impact on how I think about and approach technical challenges.

I am grateful to my clinical collaborators, Dr. Olga Patey and Professor Aris T. Papageorghiou, for their invaluable clinical guidance and support. They patiently explained key clinical concepts, helped me understand the underlying clinical needs, and provided insightful feedback that shaped my understanding of the project. Their perspective was crucial in helping me connect my technical work with the real-world clinical need.

I owe special gratitude to Dr. Koundinya Desiraju, my longtime mentor, who first saw in me the spark for research and encouraged me to apply for a PhD. He has always motivated me to work on impactful clinical problems, and his guidance both in research and in life has played a significant role in shaping the person I am today.

I am fortunate to have had the friendship and collaboration of Pramit, whose constant encouragement and support have meant so much throughout my PhD. We have shared the journey through all the rejections and acceptances together, and his optimism and determination have been a great source of motivation.

I am also deeply thankful to have met Yash at ICVSS, with whom I share a rare connection where we often find ourselves thinking the same thing and need only a glance at each other before bursting into laughter. Our joyful and often hilarious moments, shared trips, and countless anime discussions have kept the child in me alive, providing much-needed lightness and balance throughout this demanding PhD journey.

To all the amazing friends I met during my time in Oxford, thank you for making this journey so much more enjoyable and memorable. Your companionship, laughter, endless yapping, and willingness to talk about anything other than work kept me grounded through the ups and downs of research life. I am grateful to Alex, Felix, Jong, Mohammad (Bhai sahab), Ziyun, Angus, Joshua, Gargi, Harry, Furat, Ritika, Rotem, Sindhu, Anthony, Hermione, Elizabeth (Lizz) and many others for the countless coffee breaks, books discussions, travels, movie nights, hikes, and all the moments in between that made Oxford feel like home. Your friendship filled this journey with joy and unforgettable memories.

I am equally grateful to my friend Medhavi (+Atom), who moved to Melbourne for her master's. Despite the time differences and her own first-time-abroad struggles, she was always there for me, supporting me, listening to me, and calling me (often at the weirdest possible hours) just to check on me. From across the seas, she patiently listened to my endless rants and troubles, showing that distance, time zones, and sleep schedules don't stand a chance against genuine friendship.

I am also deeply grateful to my brother Varun, who, even from across continents and time zones, was always by my side through the highs and lows of my PhD. He has been a constant pillar in my life and one of the first people I would call no matter what happened. We started our journeys around the same time, me in academia and him in industry, and he stayed closely involved in every decision I made, from important life choices like whether to pursue a PhD to equally critical ones like which anime I should watch next. His support, humor, and belief in me kept me going when things felt overwhelming.

I wish to remember my maternal grandfather, the late Mr. Shyam Swarup Dwivedi, who instilled in me the values of hard work, punctuality and organization. His words on striving to work hard, to be a good person before a successful one, and his unwavering belief in me from the very beginning have stayed with me throughout my life. The walks we shared since I first learned to walk, and his conversations filled with wisdom and warmth, have profoundly shaped the person I am today. I owe much of my success to him.

I also would like to acknowledge my younger sister, Nandani, who is one of the kindest, most aware, and cheerful people I know. Her constant positivity, motivation, and the way she has cared for our parents while I was away pursuing my PhD mean more to me than words can express. I am deeply grateful for her love and support.

Finally, I would like to thank my parents, Mrs. Divya Mishra and Mr. Subhash Chand Mishra, who have supported me throughout my life from my very first steps to completing my PhD. I am forever grateful for their unconditional love, guidance, and the values they have instilled in me. Everything I have achieved is a reflection of their love, support, and sacrifices.

Abstract

Congenital heart disease (CHD) is the leading cause of infant mortality among congenital anomalies, yet many cases are missed during routine prenatal ultrasound screening. Routine fetal scans are primarily designed for general anatomical assessment and are not specifically focused on detailed evaluation of the fetal heart. As a result, imaging of fetal cardiac structures is often limited in both coverage and the level of detail required for reliable diagnosis. Detection is further challenged by the subtle presentation of many defects and the reliance on sonographers who are not specifically trained in fetal cardiology, making diagnosis highly operator-dependent.

This thesis introduces a video-based approach that learns directly from continuous fetal ultrasound videos, enabling temporal modeling of the beating heart rather than relying on isolated static frames. By leveraging the spatiotemporal information inherent in ultrasound videos, the approach aims to streamline workflow and assist sonographers in more effective CHD detection.

An unsupervised dual-conditioned diffusion model (DCDM) is developed to identify standard heart planes in free-hand ultrasound scans, reducing manual review effort without requiring labeled abnormal data. By conditioning jointly on image features and class information, DCDM enhances robustness to anatomical variability and reliably separates cardiac from non-cardiac views. Since free-hand scans are primarily used for general anatomical assessment and often focus on a single "best" frame per anatomy, they miss the continuous motion of the fetal heart. To address this limitation, the thesis adopts continuous "heart sweeps" for comprehensive capture of dynamic cardiac features. A novel task, visual query-based video clip localization (VQ-VCL), is formulated to efficiently retrieve diagnostically relevant views from these sweeps. We show that Transformer-based models, including STAN-LOC, TIER-LOC, and MCAT, enable user-guided, accurate retrieval through advanced spatiotemporal and class-specific modeling.

Building on these advances in unsupervised cardiac modeling, the thesis further addresses data scarcity, privacy constraints, and cross-domain generalization through large-scale self-supervised learning. We propose Sparse Tube Ultrasound Distillation (STUD), together with a divergence-guided model merging strategy, for decentralized anomaly detection across clinical sites without sharing patient data. Extending beyond fetal ultrasound, an approach called DISCOVR (Distilled Image Supervision

for Cross-Modal Video Representation) is introduced as a video foundation model for general echocardiography. DISCOVR employs a dual-branch self-supervised architecture that aligns spatial semantics from images with temporal dynamics in videos via online semantic cluster distillation. This cross-modal framework is shown to produce anatomically grounded and temporally coherent video representations, enabling robust transfer across fetal, pediatric, and adult populations. DISCOVR demonstrates state-of-the-art performance in anomaly detection, representation learning, and segmentation, establishing a scalable foundation for future video-based analysis of echocardiography.

Overall, this thesis presents a unified progression from unsupervised cardiac frame detection to transformer-based localization and self-supervised modeling of cardiac motion. By reframing CHD detection as a video understanding problem, it establishes new methodological foundations for automated fetal cardiac analysis. Collectively, DCDM, STAN-LOC, TIER-LOC, MCAT, STUD, and DISCOVR demonstrate that modeling the temporal and spatial richness of ultrasound data can enhance the accuracy, efficiency, and consistency of prenatal CHD screening models, marking a significant step toward intelligent, video-based diagnostic systems suitable for clinical settings.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Background

Congenital heart disease (CHD) is the most prevalent congenital malformation, affecting approximately 1% of all live births and contributing substantially to infant mortality [1, 2]. Routine prenatal ultrasound screening is typically performed between 18 and 20 weeks of gestation to evaluate essential cardiac views, such as the four chamber (4CH) and outflow tracts [3]. However, detection rates remain suboptimal due to operator expertise, image quality, and fetal positioning. Even in advanced healthcare systems, up to 51% of CHD cases are estimated to go undetected during initial screenings [4].

A critical step in prenatal CHD detection is heart standard plane identification, which means locating ultrasound frames where all cardiac landmarks are clearly visible. However, manually detecting these standard planes is both time consuming and highly dependent on sonographer expertise.

1. In the thesis, we first use data from the *PULSE* [5] project, which consists of free-hand fetal ultrasound videos spanning a range of anatomies (head, abdomen, femur, heart). By modeling standard plane detection as an *out-of-distribution (OOD) detection* task, we introduce a dual-conditioned diffusion model (DCDM) to automatically isolate heart frames from non-heart frames, reducing the need for exhaustive manual review. However, free-hand scanning

is primarily performed for general anatomical assessment rather than targeted cardiac evaluation. As a result, it often captures only a single "best" frame per anatomy and provides limited temporal coverage of the fetal heart, potentially overlooking the dynamic motion patterns that are critical for detecting subtle anomalies.

2. To address this shortcoming, the *Clinical Artificial Intelligence Fetal Echocardiography (CAIFE)* project introduces *heart sweeps*, continuous scans designed to both capture spatial and temporal features of the fetal heart. Although these sweeps ensure a more comprehensive coverage, manually detecting standard planes within them is a challenge.
   To streamline the process, we propose **Visual-Query-based Video Clip Localization (VQ-VCL)**. After acquiring a sweep, the sonographer provides a "visual query" image representing a desired standard cardiac view. Based on this query, our VQ-VCL model automatically identifies and retrieves the corresponding spatio-temporal clip of the standard cardiac view from the input video. This clip can then be used by the sonographer to efficiently analyze the cardiac view of interest without the need to spend time searching for it, streamlining workflow allowing and potentially allowing them to consult with more patients.

3. Finally, to further simplify sonographer workflow, we introduce a **self-supervised anomaly detection** technique based on a *normality modeling* approach. Trained exclusively on a large normal heart sweep dataset, the derived model learns spatio-temporal patterns characteristic of healthy fetal hearts and flags out-of-distribution (abnormal) instances in a zero-shot manner. This approach leverages readily available normal heart data for training and requires abnormal cases only at test time, reducing the need for labeled data and making the process scalable.

An overview of how the thesis contributions differ from and build on existing approaches is presented in Fig. 1.1, and these contributions are detailed further in the following section.

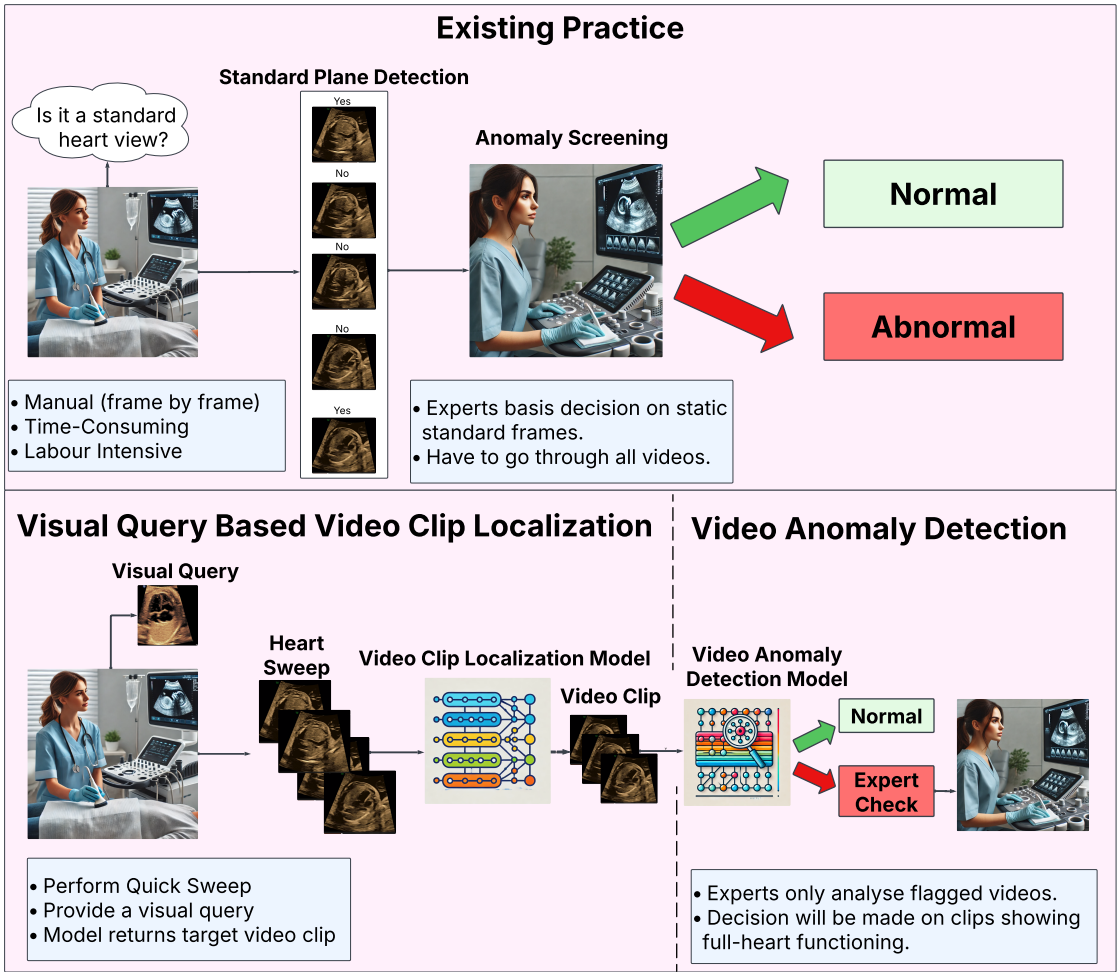


Figure 1.1: Schematic showing how the different topics investigated in the thesis fit together to address an unmet clinical need in ultrasound-based fetal health screening.

1.2 Thesis Outline

This thesis explores video understanding in medical image analysis and computer vision, addressing problems such as video localization, self-supervised video representation learning, video anomaly detection, and out-of-distribution detection, with the aim of assisting experts in detecting congenital heart diseases (CHDs). The work is structured as follows:

- **Chapter 1: Introduction:** This current chapter provides an overview of the thesis and highlights the main contributions of the work.
- **Chapter 2: Literature Review**

  This chapter defines congenital Heart Diseases (CHD) and video understanding, discusses the preliminaries, and reviews related prior research in this area.

- **Chapter 3: Datasets** The chapter describes the different datasets used in our experiments throughout the doctoral thesis.

- **Chapters 4-9 each describe original technical contributions of the thesis.**

- **Chapter 4: Unsupervised Out-of-Distribution Detection using Diffusion Models in Free-hand Ultrasound Videos:** This chapter presents dual-conditioned diffusion models (DCDM), as published at MICCAI 2023 [6], for out-of-distribution (OOD) detection in fetal ultrasound videos. The approach targets challenging scenarios characterised by high intra-class variability and subtle differences between in-distribution (ID) and OOD samples, as encountered when distinguishing heart views from other anatomies.

  DCDM conditions on both ID class information and latent image features to enable reconstruction-based OOD detection. By constraining the generative manifold, reconstructed samples remain structurally and semantically consistent with in-distribution data. Experimental results show improvements over reference techniques, including a 12% increase in accuracy, a 22% increase in precision, and an 8% improvement in F1 score.

- **Chapter 5: Visual Query-based Video-Clip Localization for a Single Heart View** This chapter presents the task of visual query-based video clip localization for medical video understanding, as published at MICCAI 2024 [7]. The task addresses the need to accurately detect standard frame clips in fetal ultrasound videos, which is a critical step for reliable clinical assessment and diagnosis. Identifying these standard clips supports consistent evaluation of fetal development and detection of abnormalities. STAN-LOC is introduced as a method that incorporates three main components. A Query-Aware Spatio-Temporal Fusion Transformer combines information from the visual query and the input video to generate features that capture spatio-temporal relationships within the data. A Multi-Anchor, View-Aware Contrastive loss is used to reduce the impact of noise arising from manual annotations, particularly at event boundaries and in videos containing highly similar objects. During

inference, a query selection algorithm is applied to identify the most suitable visual query for a given video and reduce sensitivity to query quality. STAN-LOC is applied to detecting standard-frame clips in fetal heart ultrasound sweeps using four-chamber view queries, and its performance is also evaluated on the PULSE dataset for retrieving the standard transventricular plane in fetal head videos. Results show that STAN-LOC outperforms the previous state of the art by 22% in mean temporal Intersection over Union (mtIoU).

- **Chapter 6: Multi-Tier Transformer for Visual Query-based Video-Clip Localization** In this chapter, we extend STAN-LOC [7] to the task of multi heart view clip localization, as published in the Medical Image Analysis journal [8]. Building on prior work in visual query-based video understanding, we propose TIER-LOC, a model that extracts and integrates features from multiple levels, referred to as Tiers, ranging from coarse to fine detail. This multi Tier formulation supports improved detection of subtle differences between heart views and better adaptation to variations in scale and resolution. TIER-LOC is composed of three main components. A Multi-Tier Spatio-Temporal Transformer fuses spatio-temporal features from multiple Tiers of both the video frames and the visual query. A Multi-Tier Dual Anchor Contrastive Loss is used to address annotation noise, particularly at event boundaries and in videos containing highly similar anatomical structures. In addition, a Temporal Uncertainty-Aware Localization Loss reduces sensitivity to imprecise event boundaries by relaxing strict localization constraints during training. The effectiveness of TIER-LOC is evaluated on three datasets, including two ultrasound video datasets and an open-source egocentric video dataset. The method accurately localizes standard-frame clips across multiple anatomical views in fetal ultrasound heart sweeps and shows robust performance on the large-scale PULSE dataset as well as a dataset derived from Ego4D. Across these benchmarks, TIER-LOC achieves performance gains of 7%, 4%, and 4% over state-of-the-art models.

- **Chapter 7: Class-Specific Token Learning for Visual Query-based Video-Clip Localization** Chapter 7 introduces the Multi-Tier Class-Aware

Token Transformer (MCAT), published in AAAI 2025 [9], for visual query-based video clip localization (VQ-VCL) in fetal ultrasound videos. The method addresses limitations of existing VQ-VCL approaches that rely on shared or generic embeddings to represent different anatomical classes and often fail to capture subtle anatomy-specific features required for accurate and fine-grained localization in clinical settings. MCAT incorporates a class-specific token learning mechanism within a multi-tier transformer framework. This design enables the model to disentangle features unique to each anatomical class and to activate only the token corresponding to the queried anatomy, resulting in more focused and efficient feature representations. The multi-tier architecture, with tier-specific embeddings and cross-attention mechanisms, further supports representation of anatomical variation across different spatial scales. MCAT is evaluated on two fetal ultrasound video datasets and a VQ-VCL dataset based on Ego4D. Results show that the method outperforms SOTA VQ-VCL approaches using shared embeddings, achieving 10% and 13% higher mIoU on the ultrasound datasets and a 5.35% improvement on Ego4D, while reducing token usage by 96%.

- **Chapter 8: Self-Supervised Video Representation Learning for Video Anomaly Detection**

  This chapter focuses on Sparse Tube Ultrasound Distillation (STUD), as published in MICCAI 2025 [10], a privacy-preserving, zero-shot detection framework for congenital heart disease (CHD) in fetal ultrasound videos. STUD is designed to overcome key challenges in deep learning-based CHD detection, including the scarcity of labeled abnormal data, strict privacy regulations, and the high cost and practical difficulty of collecting and curating large-scale ultrasound video datasets for rare conditions. Traditional approaches often rely on aggregating large datasets from multiple sites, a process that is constrained by privacy concerns and data governance policies. STUD reformulates CHD detection as a video normality modeling problem and incorporates a model merging strategy for decentralized collaboration without data sharing. Each hospital independently trains a sparse video

tube-based self-supervised video anomaly detection (VAD) model on normal fetal heart ultrasound clips using a self-distillation loss, enabling learning of healthy case distributions. By operating on sparse spatio-temporal video tubes, STUD significantly reduces the number of processed tokens compared to dense video representations, resulting in an order-of-magnitude improvement in computational efficiency. Knowledge from site-specific models is then combined using the Divergence Vector-Guided Model Merging approach, DivMerge, which aggregates model weights without exchanging sensitive data. This strategy preserves privacy while maintaining rich, domain-agnostic spatio-temporal representations and mitigates performance degradation caused by inter-site domain shifts. The effectiveness of STUD is evaluated on real-world fetal ultrasound data collected from five hospital sites. At inference time, anomaly detection is performed in a zero-shot manner using a k-Nearest Neighbours classifier on the learned representations, without any fine-tuning on abnormal cases. The merged model substantially outperforms individual site-specific models, achieving a 23.77% increase in accuracy and a 30.13% improvement in F1-score on external test sets.

- **Chapter 9: Self-Supervised Learning of Echocardiographic Video Representations via Online Cluster Distillation**

  This chapter focuses on DISCOVR (Distilled Image Supervision for Cross Modal Video Representation), a self-supervised dual branch framework for cardiac ultrasound video representation learning accepted to NeurIPS 2025. While self-supervised learning (SSL) has achieved strong performance in natural image and video domains, its application to echocardiography remains challenging due to the subtlety of cardiac structures, complex temporal dynamics, and the lack of domain-specific pre-trained models. Existing SSL approaches, including contrastive, masked modeling, and clustering-based methods, often struggle in this setting because of high inter-sample similarity, sensitivity to low PSNR inputs typical of ultrasound, and augmentation strategies that can distort clinically relevant features. DISCOVR combines a clustering-based video encoder that models temporal dynamics with an

online image encoder designed to capture fine-grained spatial semantics. The two branches are connected through a semantic cluster distillation loss, which transfers evolving anatomical knowledge from the image encoder to the video encoder through cross-modal cluster alignment. This design enables the video representations to remain temporally coherent while being enriched with detailed semantic information about anatomical structures. The framework is trained without labels, pretrained models, or aggressive data augmentations, relying instead on online self-distillation to learn anatomically meaningful representations directly from ultrasound data. DISCOVR is trained exclusively on normal echocardiography videos, learning characteristic patterns of healthy cardiac motion and structure, with pathological cases identified as deviations from this learned normality. DISCOVR is evaluated across six echocardiography datasets spanning fetal, pediatric, and adult populations. Experimental results show that DISCOVR outperforms both domain-specific video anomaly detection methods and state-of-the-art video-based SSL baselines in zero-shot and linear probing scenarios, while also achieving superior segmentation transfer.

- **Chapter 10: Conclusions** In this concluding chapter, we highlight the collective contributions of thesis which covers: unsupervised out-of-distribution detection, visual query-based video clip localization, and self-supervised video representation learning for advancing CHD screening in fetal ultrasound. We also reflect on the key limitations encountered during this work and outline promising directions for future research, providing an integrated perspective on the overall impact and future potential research questions that can be investigated as a result of our research.

1.3 Thesis Contributions

This thesis makes the following contributions to advance video understanding and its application for congenital heart disease (CHD) screening in fetal ultrasound videos:

- **Unsupervised Out-of-Distribution Detection using Diffusion Models in Free-hand Ultrasound Videos:** We introduce dual-conditioned diffusion models (DCDM) for out-of-distribution detection in fetal ultrasound videos. DCDM leverages both class-conditioning and latent feature conditioning within a diffusion model to robustly distinguish standard heart views from other anatomies, without the need for labeled OOD data.

- **Visual Query-based Video-Clip Localization for a Single Heart View:** We define and formulate the novel task of visual query-based video clip localization (VQ-VCL) for medical video understanding. We propose STAN-LOC as the first solution for this task, combining query-aware spatio-temporal fusion, a contrastive loss tailored to handle annotation noise, and an inference-time query selection algorithm for accurate retrieval of standard ultrasound clips.

- **Multi-Tier Transformer for Visual Query-based Video-Clip Localization:** We extend VQ-VCL localization to multiple heart views by introducing TIER-LOC, a multi-tier transformer model. TIER-LOC extracts and fuses features across several scales, incorporates a robust loss design, and generalizes the VQ-VCL framework for fine-grained retrieval across varying anatomical scenarios.

- **Class-Specific Token Learning for Visual Query-based Video-Clip Localization:** We present MCAT, a multi-tier class-aware token transformer that disentangles representations for different anatomical classes and selectively activates class-specific tokens during inference. This approach enhances both the accuracy and computational efficiency of VQ-VCL models by moving beyond generic, shared embeddings.

- **Self-Supervised Video Representation Learning for Video Anomaly Detection:** We develop Sparse Tube Ultrasound Distillation (STUD) together with a divergence-guided model merging framework (DivMerge) for privacy-preserving anomaly detection in ultrasound. This methodology enables decentralized collaborative model training, allowing zero-shot anomaly detection without the need for sharing raw patient data.

- **Self-Supervised Learning of Echocardiographic Video Representations via Online Cluster Distillation:** We introduce DISCOVR, a dual-branch self-supervised learning architecture for ultrasound video. DISCOVR combines clustering-based temporal modeling and image-based semantic encoding, utilizing a distillation mechanism to achieve semantically rich and temporally coherent representations for a range of downstream tasks.

1.4 Research Publications

I am first author on the following peer-reviewed works related to video understanding, all of which have been accepted or published in leading peer-reviewed venues in the field.

1. **Chapter 4: Mishra, D.**, Zhao, H., Saha, P., Papageorghiou, A.T. and Noble, J.A., 2023, October. Dual conditioned diffusion models for out-of-distribution detection: Application to fetal ultrasound videos. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 216-226). Cham: Springer Nature Switzerland.
2. **Chapter 5: Mishra, D.**, Saha, P., Zhao, H., Patey, O., Papageorghiou, A.T. and Noble, J.A., 2024, October. STAN-LOC: Visual Query-Based Video Clip Localization for Fetal Ultrasound Sweep Videos. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 742-752). Cham: Springer Nature Switzerland.
3. **Chapter 6: Mishra, D.**, Saha, P., Zhao, H., Hernandez-Cruz, N., Patey, O., Papageorghiou, A.T. and Noble, J.A., 2025. TIER-LOC: Visual Query-based Video Clip Localization in fetal ultrasound videos with a multi-tier transformer. Medical Image Analysis, p.103611.
4. **Chapter 7: Mishra, D.**, Saha, P., Zhao, H., Hernandez-Cruz, N., Patey, O., Papageorghiou, A. and Noble, J.A., 2025, April. MCAT: Visual Query-Based Localization of Standard Anatomical Clips in Fetal Ultrasound Videos Using Multi-Tier Class-Aware Token Transformer. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 27, pp. 28267-28275).

5. **Chapter 8:** Saha, P. *, **Mishra, D. ***, Hernandez-Cruz, N., Patey, O., Papageorghiou, A.T., Asano, Y.M. and Noble, J.A., 2025, September. Self-supervised normality learning and divergence vector-guided model merging for zero-shot congenital heart disease detection in fetal ultrasound videos. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 560-571). Cham: Springer Nature Switzerland. *Equal Contribution.

6. **Chapter 9: Mishra, D.**, Salehi, M., Saha, P., Patey, O., Papageorghiou, A.T., Asano, Y.M. and Noble, J.A., 2025. Self-supervised Learning of Echocardiographic Video Representations via Online Cluster Distillation. arXiv preprint arXiv:2506.11777. (Accepted in NeurIPS 2025)

2
Background

2.1 Introduction

My thesis focuses on developing and applying deep learning models to identify congenital heart diseases (CHD) directly from fetal ultrasound videos. Section 2.2 provides an overview of CHD, including its prevalence, clinical significance, and the central role of ultrasound in its detection. This section also reviews current clinical workflows and highlights the challenges encountered in routine cardiac anomaly screening. Section 2.3 presents a detailed survey of existing automated CHD detection methods, with particular attention to how prior work has addressed segmentation, landmark localization, and the limitations arising from relying primarily on static images or limited frames. The section also identifies key gaps in the literature, such as underutilization of temporal video information and the limited use of frameworks that can generalize across populations and clinical sites. In the next chapter, an introduction to state-of-the-art video understanding techniques is provided, detailing how recent advancements in deep learning are shifting from traditional static-frame analysis to comprehensive modeling of temporal dynamics in medical videos. This sets the foundation for the novel video-based methods introduced and evaluated throughout this thesis.

2.2 Congenital Heart Disease

Congenital heart disease (CHD) refers to abnormalities present in the human heart since birth. CHD is the sixth major cause of infant death globally, fifth in middle-income countries and second in high-income countries [11]. CHD contributes to 50% of all infant deaths related to malformations. In 2019, 3.12 million babies were born with some form of CHD (2305.2 per 100,000 live births), and 217,000(6.95%) died, which included 150,000 (4.8%) infants[12]. CHD detection rates have increased in the last decade due to using technologies like ultrasound. Major types of CHD can be detected using ultrasound screening.

Obstetric ultrasound (US) has become the standard procedure in clinical practice. It serves as a crucial tool in confirming pregnancy and determining gestational age, identifying multiple pregnancies, detecting congenital anomalies, assessing placental health, monitoring fetal position, tracking fetal growth, evaluating amniotic fluid levels, and assisting in conducting various diagnostic tests.

During routine clinical practice, a second-trimester anomaly scan is typically performed between 18 and 20 weeks of pregnancy. This scan is conducted to screen for any fetal anomalies. In the case of assessing the fetal heart, various views such as the four-chamber (4CH) view, three-vessel view (3VV), three-vessel trachea view (3VT), left ventricle outflow tract (LVOT), and right ventricle outflow tract (RVOT) are examined for potential anomalies. If a cardiac anomaly is suspected during the anomaly scan, the patient is then referred for fetal echocardiography, which provides a more detailed evaluation of the fetal heart. However, it is important to note that approximately 51% of congenital heart diseases (CHDs)[4] may be missed during the initial screening, leading to potential adverse outcomes.

2.2.1 CHD Detection Literature

An *et al.* [13] proposed an instance segmentation model using Mask-RCNN [14] to segment only the four cardiac chambers. They modified the object proposal mechanism of Mask-RCNN to group nearby proposals and assign them to the same class. This strategy helped in retrieving missing parts and obtaining better segmentation of the chambers (mean Dice 0.72). However,their model was only compared with

U-Net[15]. Although video data was available, the authors only utilized end-systolic and end-diastolic frames, neglecting temporal information. Despite the dataset containing abnormal heart conditions, the proposed segmentation model was not evaluated for heart condition detection. Finally, only four landmarks were annotated in the 4CH view, which is insufficient to detect the majority of CHDs.

Wong *et al.* [16] designed two view specific(3VT and 4CH) and one combined view(4CH + 3VT) semantic segmentation networks to accurately identify 14 landmarks in the 4CH and 3VT views of the fetal heart. To overcome the difficulty of handling objects of different sizes, they employed an exponential log loss approach. This approach combines exponential dice loss and exponential cross-entropy loss, providing a weighted combination that addresses the issue at hand. Specifically, the authors recognized that simple dice loss is not ideal for small structures, as misclassifying a few pixels can significantly impact the coefficient. To tackle this, they introduced exponential dice loss inspired by focal loss and which focuses more on misclassified examples. Despite achieving a dice score of 79%, the multiview model developed by the authors encounters an issue. While the model incorporates structures from both the 3VT and 4CH views, it only takes input from a single view. This limitation leads to missing labels, causing the dice score to abruptly drop to zero and resulting in training instability.

Pu *et al.* [17] introduce MobileUNet-FPN, where they segment 13 key anatomical structures in a 4CH view. Their network is a combination of a U-Net [15], MobileNet[18] and Feature Pyramid Network (FPN)[19] and aims to utilize multi-scale features to segment the fetal heart. Despite its complex design, the segmentation model was found to only outperform baselines in detecting 6/13 anatomical landmarks and to only shows a small improvement of 0.39%. in IOU compared to baselines.

Patra *et al.* [20] build a spatio-temporal model to classify and localize fetal heart frames in fetal ultrasound videos. The paper introduces two approaches to encode the temporal information known as Direct Temporal Encoding(DTE) and Hierarchical Temporal Encoding(HTE) to improve the localisation(79.68% IOU) and detection performance(83.58% accuracy).The work is aimed to guide the sonographer to detect and localize heart views but offers no solution for detecting CHDs.

Chotzoglou *et al.* [21] took an unsupervised anomaly detection approach where they directly applied $\boldsymbol{\alpha - GAN}$ [22] to model normal 4CH heart images and showed anomaly detection performance by detecting Hypoplastic Left Heart Syndrome (HLHS).

Sengan *et al.* [23] developed ARVNet, a U-Net [15] inspired model which segments 3VT, 3VT, LVOT, RVOT and 4CH view but only use 4CH view to detect fetal cardiac rhabdomyoma.

Sapitri *et al.* [24] used different YOLO models to detect 9 anatomical landmarks in the 4CH view of the fetal heart. The paper just compared the performance of different YOLO architectures on the task and used the dataset for benchmarking.

Xu *et al.* [25] propose an approach to segment seven anatomical structures in the apical 4CH view of the fetal heart. They introduce two key components: the Dilated Convolutional Chain (DCC) and the W-Net module. The DCC module consists of stacked dilated CNN layers with increasing dilation rates. Its purpose is to gather both global and local information from the 4CH view, aiding in the segmentation process. On the other hand, the W-Net module employs two U-Net networks in a sequential manner, enabling a repetitive process of encoding and decoding to enhance the accuracy of the segmentation maps. The work focuses on developing a frame-level segmentation model for a normal fetal heart and uses the heart dataset as a benchmark to compare their model with existing approaches rather than developing solutions for detecting congential heart diseases.

Tan *et al.* [26] introduce a pipeline to automatically detect Hypoplastic Left Heart Syndrome (HLHS) using 4CH, RVOT and LVOT views of the fetal heart. The pipeline comprises of two main parts. The first part curates the data into the above mentioned classes using the predictions from SonoNet [27] while the second part trains a multi-task model to predict the heart view and if the view is normal heart or HLHS. The use of SonoNet introduces potential bias in the trained model as SonoNet is only trained on healthy heart cases. Moreover, to counter the model's performance drop on frames with low-diagnostic information, they add cardiac view preserving perturbation to the image and exclude predictions where perturbations change the classifier's prediction. This leads to removal of 27% images from the high-quality set(images having high-diagnostic quality according to the clinician) and

39% images in the low- quality set (images having low-diagnostic quality according to the clinician).

Dong *et al.* [28] develop a three-stage quality assessment network for 4CH planes in fetal cardiography. The first stage comprises a view classifier that classifies whether a frame is a 4CH view. The second stage consists of a multi-task classifier model trained to classify the image into different levels of zoom and gain. Finally, the image is fed into an object detection system known as ARVBNET, which is essentially a Single Shot Detector (SSD) equipped that is enhanced with Aggregated Residual Visual Block (ARVB) blocks. These ARVB blocks are composed of grouped and dilated CNN layers. ARVBNET detects the presence of left atrial pulmonary vein angle (PVA), apex cordis, moderator band and multiple ribs. The algorithm uses a predefined point system to score the detection results of each stage and quantify the scores to decide whether the given frame is standard or non-standard.

Lu *et al.* [29] extended the YOLOX [30] model by incorporating an additional segmentation branch. Moreover, to improve segmentation performance, they modified the existing Non-Maximum Suppression(NMS) mechanism by making it depend on confidence in segmentation performance, object localisation, and object classification performance. However, the motivation to design a model doing both object detection and instance segmentation tasks is not well justified, and no comparison with the original YOLOX is made. Furthermore, in clinical settings, acquiring both bounding box and segmentation annotations may not be feasible. This limitation reduces the practicality and applicability of the proposed approach in real-world clinical scenarios. Finally, no clinical application of the developed model was discussed or demonstrated.

Arnaout *et al.* [31] propose a multi-stage pipeline to detect complex CHDs. In the first stage, they train a view classifier to classify the input images into one of the 5 heart views (3VV, 3VT, LVOT, A4C, ABDO) or NT (non-target). Once classified, each heart view is passed to a per-view diagnostic classifier that predicts if the view is normal or abnormal. The score of each per-view diagnostic classifier (except ABDO) is then fed to a composite classifier which gives the final diagnosis (normal or abnormal). Moreover, the A4C image is additionally fed to a segmentation model to estimate biometry parameters cardiothoracic ratio (CTR), cardiac axis (CA)

and fractional area change (FAC). The method provides an end-to-end approach to detecting CHDs. However, the approach comprises 8 models, which might be computationally expensive to use in real time. Additionally, the method requires 5 perfectly captured standard heart views for each patient, which might not be available.

Nurmaini *et al.* [32] train a Mask-RCNN [14] for instance segmentation of 6 landmarks in the 4CH view. However, they only utilize one landmark, which represents if there is a hole in the septum to detect septal defects, while segmentation maps for other landmarks are not utilized. Moreover, as they train a Mask-RCNN model, it requires both bounding-box and segmentation annotations.

Gong *et al.* [33] propose a complex approach to detect fetal congenital heart disease using a One-Class classification strategy. The method begins by training three auto-encoders to identify 4CH views from other views, referred to as the DAnomaly module. These identified 4CH frames are then utilized for training an unsupervised Wasserstein GAN with gradient-penalty (WGAN-GP) [34] model. The WGAN-GP is similar to a GAN but employs Wasserstein distance as a loss function, which provides enhanced stability compared to the conventional adversarial GAN loss. The objective of training the WGAN-GP model is to learn low-level features in 4CH frames which can then be transferred to the final classifier. Hence, the authors utilize the trained discriminator of the WGAN-GP model for transfer learning by adding a global-average pooling layer and some fully connected layers, resulting in a DGACNN classifier. DGACNN is then fine-tuned on four-chamber end-systole frames to classify frames as normal or abnormal. The final model, incorporating DAnomaly screening, achieves only a 1.5% improvement in accuracy over baseline but requires training three computationally expensive autoencoder models.

Nurmaini *et al.* [35] propose a two-stage pipeline for septal defect detection in a 4CH view. The first stage comprises a pre-trained U-Net [15] to segment five anatomical 4CH view landmarks, while in the second stage, the resulting mask is fed to Faster-RCNN [36] for detecting septal defects. The authors report a mAP of 87.80% for detecting septal defects; however, no comparisons with relevant baselines like Mask-RCNN were made.

Qiao *et al.* [37] propose a YOLO-v4 [38] based model FLDS for localization of 4

anatomical landmarks in the 4CH view. They develop a multi-stage residual hybrid attention module (MRHAM) to focus the model on spatial and content information of the fetal cardiac chambers. This approach reports a mAP of 95.3% in four chamber localization, however they have only compared with variants of a YOLO-v4 architecture. Additionally, the authors have only localized 4 anatomical landmarks in the 4CH view and have not discussed clinical application of the proposed work. Komatsu *et al.* [39] propose SONO, a model to detect structural abnormalities in fetal US videos. They annotate 18 anatomical landmarks present in the 4CH and 3VT view of the fetal heart and train a YOLOv2 model to detect them. Further, they utilize the detection probabilities of 8 selected anatomical landmarks in 4CH and 4 in 3VT to calculate an abnormality score. The abnormality score is simply one minus the mean probability of detection of the selected landmarks for each view. A sample having an abnormality score below a pre-defined threshold was termed abnormal. Finally, the authors provide a bar code-like timeline visualization which represents the position and detection probability of each anatomical landmark in a US-screening video. The SONO model utilized YOLOv2 and had an AUC of 78.70% and 89.10% for 4CH and 3VT. Comparisons with existing object detection based methods are missing.

After carefully examining the existing literature, we have identified several shortcomings of the current congenital Heart Disease (CHD) detection approaches. These limitations include:

- Most of the approaches (except Patra *et al.* [20]) have introduced models that ignore the temporal information in fetal echocardiography videos. However, as the heart is beating, the anatomical landmarks within each view change structurally depending on the phase in the cardiac cycle; hence, approaches treating each frame separately are not enough to capture all the variation. Moreover, the heart comprises multiple views (SITUS, 4CH, LVOT, 3VV, 3VT), and temporal modelling can help model the interactions and transitions between them, thereby improving CHD detection.

- Many approaches emphasize the significance of detecting congenital heart disease; however, they primarily utilize medical data to benchmark their

models for tasks such as segmentation and object detection without leveraging the model's outputs to aid clinical diagnosis.

- All approaches assume that label data is readily available and do not utilize limited label data modelling approaches like semi-supervised learning. However, it is very hard to get expert annotations in practical scenarios, especially for medical data.

2.3 Video Understanding

Deep-learning-based video understanding is the field of deep learning that focuses on extracting meaningful information from video data. Meaningful information can be present in the form of actions performed in the video (Video Action Recognition/Classification), location of the actions in the video (Temporal Action Localization), information about the objects present in the video (Video Object Detection/Segmentation) or retrieving frames/clips from a video based on a query (Video Question Answering/ Video-based Image Retrieval). A key distinction between video and image understanding lies in incorporating temporal modelling. Temporal information is vital since objects and their interactions with the environment often span multiple frames, establishing dependencies over time. This sets video understanding apart from image understanding, where the focus is primarily on analyzing static (or spatial only) visual content.

2.3.1 Video Object Detection

Video Object Detection (VOD) is an extension of object detection techniques used in still images [36],[40] , specifically designed for analyzing video. However, applying still image object detectors to video often encounters challenges in consistently detecting objects due to their appearance changes over time. Videos typically contain valuable temporal information, wherein the same object can appear in multiple frames for a certain duration. To enhance accuracy, researchers have proposed integrating temporal information into object detectors. In recent years, a prevalent approach has been feature refinement, which incorporates spatio-temporal information [41]–[42]. This technique utilizes aggregated informative features from

neighbouring frames to compensate for degraded features such as motion blur, camera defocus, and large pose variations caused by fast motion in the target frame. On the other hand, methods like FGFA [41], and MA-Net[43] calculate optical flow between frames to incorporate motion information, others such as TSSD [44] and convLSTM[45] employ recurrent neural networks to propagate features from neighbouring frames. In recent studies, SELSA [46], MEGA [42], and RDN[47], a new approach has emerged to improve video object detection that incorporates distant frames in addition to neighbouring frames. This is in response to the limitations of relying solely on neighbouring frames, as it often leads to detection issues caused by deterioration in object appearance due to fast motion. However, as these methods perform object-level aggregation of region proposals, they heavily rely on the accuracy of the frame-level object detector. Additionally, the memory requirement becomes a crucial factor when leveraging distant frames, particularly when employing a static sliding window or external memory that is updated randomly [47] or in a specific order [42], [48]. As Transformer-based models exhibit improved performance in image object detection, researchers have started expanding their applications to the domain of videos [49–51]. The TransVOD families [49, 51] have introduced a temporal Transformer to the original Deformable-DETR [52], enabling the fusion of spatial and temporal information to address the challenge of feature degradation. Likewise, PTSEFormer [50] has introduced progressive feature aggregation modules to enhance the performance of existing Transformer-based image object detectors. Building on TransVOD[51], FAQ [53] proposes a plug-play novel query aggregation module (VQA) for initialisation and aggregation of queries in video-based transformers and surpasses TransVOD by 2.7%.

2.3.2 Temporal Action Detection

Temporal Action Detection (TAD) involves identifying and categorising all actions within a video comprising multiple actions (untrimmed). Existing methods can be broadly classified into two- and one-stage methods. Two-stage methods [54–58] divide the detection process into proposal generation and proposal classification. Previous works [59–64] have mainly focused on the proposal generation phase. Specifically, some approaches [59, 62, 63] predict the likelihood of action boundaries

and densely match the start and end time instances based on prediction scores. Anchor-based methods [60, 61], on the other hand, classify actions within predefined multiscale windows with regular temporal intervals. However, two-stage methods suffer from high complexity and cannot be trained in an end-to-end manner. On the other hand, several recent works have directed their attention towards single-stage TAL, aiming to localize actions in a single shot without relying on action proposals. Many of these approaches adopt an anchor-based methodology, where anchor windows are sampled from sliding windows. In their pioneering work, Lin et al. [65] introduced the first single-stage TAL framework, drawing inspiration from a single-stage object detector [66]. In a different vein, Buch et al. [67] introduced a recurrent memory module specifically tailored for single-stage TAL. Long et al. [68] proposed a novel approach that utilizes Gaussian kernels to dynamically optimize the scale of each anchor, leveraging a 1D convolutional network. Yang et al. [69] also employed convolutional networks for single-stage TAL to explore the potential of combining anchor-based and anchor-free models. Lin et al. [70] recently presented an anchor-free single-stage model by incorporating a saliency-based refinement module within the convolutional network's design. Several recent studies [71–76], have utilized Transformers for single-stage TAD to enhance detection performance. Specifically, these works utilize a CNN-based feature extractor to extract features from all the frames and then feed them to a Transformer that utilizes them to predict the action and regress the location in the input video.

2.3.3 Visual Query 2D Localization

Visual Query 2D Localization (VQ2D) is the task of identifying the most recent time a queried object, represented by a cropped image, was observed in a video. To address this task, Grauman *et al.*[77] introduced a Siamese-RCNN network that utilizes Faster-RCNN [36] to generate region-proposals (bounding boxes that might contain an object) and extract features inside the proposals. The proposal's features are compared with the query image using a Siamese head, and the most similar proposal is selected. CocoFormer [78] addresses several concerns regarding the training methods employed by Siam-RCNN. These concerns include biased data sampling, where the target object is consistently available during training, and inadequate

modelling of global context in the query-proposal classifier. To tackle these issues, CocoFormer follows a similar approach to [77], extracting region proposals from the input video and features from the query image. However, an additional step is introduced in CocoFormer, where the features of the query image are passed through a conditional projection layer. This transformation is then applied to the proposal features, resulting in query-aware proposals. The query-aware proposals are subsequently input to a multi-head attention mechanism, enabling the model to leverage the global context of the corresponding frame and make accurate predictions about the target object. Furthermore, the authors incorporate negative query-frame training pairs that include objects in the background to address sampling bias. This helps create a more balanced and representative training dataset. MINOTAUR [79] proposes a generic multi-task model based on TubeDETR [80] that can solve the task of VQ2D and Video Grounding (given a video and text caption, localize the action temporally and spatially). Their approach involves utilizing a Modality-specific Query Encoder, which separately encodes the text and image queries, while the input video is encoded using a visual backbone. The encoded query representations and the video encoding are then fed into a transformer-based Video-Query encoder, facilitating the capture of interactions across multiple modalities. The output of the Video-Query encoder is passed through a Space-Time Decoder, which leverages the video-query features to learn modality-specific time embeddings. Finally, the prediction heads utilize these time embeddings to make predictions for foreground identification, start/end time logits, and bounding boxes for every frame within the video sequence.

2.3.4 Self-Supervised Representation Learning

Self-supervised learning (SSL) aims to learn transferable visual features without manual labels by solving proxy tasks defined on raw data. In images, *contrastive* methods such as SimCLR [81] and MoCo (including MoCo-v2 and MoCo-v3) [82–84] learn view-invariant embeddings by pulling together different augmentations of the same image and pushing apart others. NNCLR [85] improves stability by retrieving nearest neighbors as additional positives. Non-contrastive and redundancy-reduction approaches like BYOL [86] and SimSiam [87],remove explicit negatives

via asymmetric student–teacher or predictor designs with stop-gradient, while Barlow Twins [88] and VICReg [89] regularize cross-correlations and per-dimension variance to avoid collapse. Prototype-based methods such as DeepCluster [90], SwAV [91], SeLa-v2 [92], and DINO/DINOv2 [93, 94] align features to evolving cluster assignments and benefit from multi-crop training. Masked prediction forms a second major family: MAE [95] reconstructs pixels from heavily masked patches using an asymmetric encoder–decoder; BEiT/BEiT-v2 [96, 97] and iBOT [98] replace raw-pixel targets with tokenizer or teacher-driven targets; MaskFeat [99] reconstructs hand-crafted features; and data2vec [100, 101] learns modality-agnostic latent targets. I-JEPA [102] predicts the embedding of masked image regions directly from a context block, reducing reliance on heavy augmentations and pixel-space decoding.

In videos, SSL must capture motion and temporal redundancy in addition to appearance. Early temporal pretexts include speed prediction and order reasoning exemplified by SpeedNet [103]. Contrastive video SSL extends image recipes with temporal consistency, for example CVRL [104], while cross-modal alignment leverages audio–visual correspondence in XDC [105] and AVID [106]. Masked video modeling has emerged as a strong and scalable paradigm. VideoMAE [107] shows that random tube masking with high ratios and lightweight decoders is effective and data-efficient; VideoMAE-v2 [108] improves scaling with dual masking. Joint image–video pretraining is feasible with OmniMAE [109]. Motion-aware or semantic targets further enhance temporal learning: MGMAE [110] biases masks toward dynamic regions; MVD [111] predicts teacher features for masked tokens; and SIGMA [112] introduces Sinkhorn-guided clustering to inject higher-level semantics during reconstruction. Token-efficiency methods such as EVEREST [113] remove redundant spatiotemporal tokens to enable longer clips at similar compute; related token-selection/merging strategies (e.g., TokenLearner and ToMe-ViT [114, 115]) serve a similar goal when paired with SSL. Predictive joint-embedding ideas have also been extended to video: V-JEPA [116] and V-JEPA-2 [117] learn to forecast latent representations of masked spatiotemporal regions from observed context without negatives or pixel-space losses.

Clinical video domains adapt these ideas to low SNR, speckle, subtle anatomy, and periodic motion. In ultrasound and echocardiography, USCL [118] demonstrated in-domain contrastive pretraining from ultrasound videos. EchoCLR [119]

introduced patient-aware positives and a frame-reordering pretext tailored to echocardiogram videos. Masked modeling has been customized for sonography and long clinical videos: EDMAE [120] decouples masking and decoding for pediatric echo view recognition; deblurring-MAE [121] addresses speckle and blur in ultrasound images; UltraMAE [122] incorporates multi-modal inputs for ultrasound pretraining; and SurgMAE [123] adapts masked video pretraining to very long surgical procedures. Beyond cardiac imaging, Endo-FM [124] pretrains an endoscopy foundation model at scale. These studies generally report improved transfer under label scarcity and domain shift when masking, targets, and augmentations are made clinically safe and temporally aware.

3 Datasets

In this chapter, we discuss the various ultrasound and computer vision datasets, both private and public, that were utilized for the different contributions throughout this PhD. Table 3.1 provides a consolidated overview of all datasets used in this thesis.

Table 3.1: Summary of datasets used in this thesis.

Dataset	Population	Videos	Public	Chapters
PULSE	Fetal	357	No	4, 5, 6
CAIFE	Fetal	30,000+	No	4, 5, 6, 7, 8, 9
EchoNet-Dynamic	Adult	10,030	Yes	9
EchoNet-Pediatric	Pediatric	10,755	Yes	9
RVENet	Mixed	3,576	Yes	9

3.1 PULSE Heart Data

Perception Ultrasound by Learning Sonographer Experience (PULSE) [5] (Ethics Reference 18/WS/0051) is a prospective data collection study where full-length routine obstetric ultrasound scan videos while tracking the actions of the sonographer are recorded (Fig. 3.1). The videos collected are free-hand, meaning the sonographers have no fixed trajectory and can scan the various anatomies in any direction, but are guided by clinical guidelines on standard plane capture. In a typical scan, the sonographer looks for a list of standardized anatomy views, freezes the video

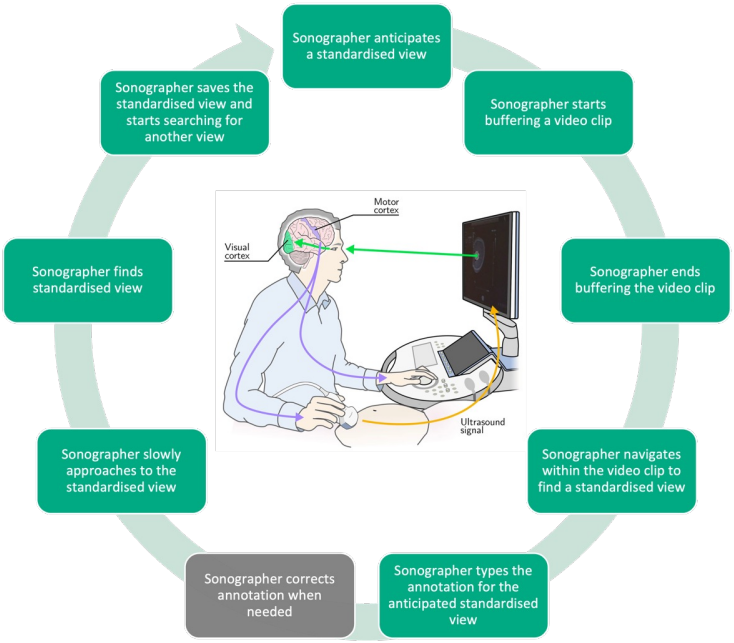


Figure 3.1: PULSE Sonographer Setup and Workflow.

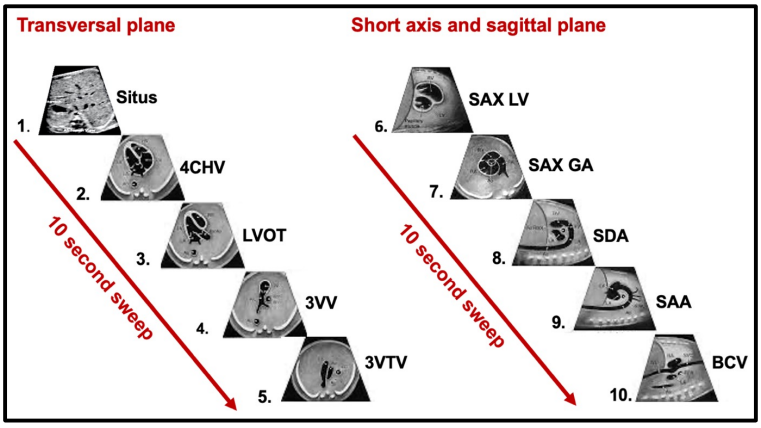


Figure 3.2: Figure depicting the Transversal and Short axis sweeps collected in CAIFE project.

once the standard view is located and performs manual biometry and/or annotation on the located view before looking for the next view. The average duration of a full scan is 20-45 minutes and comprises 13 anatomical views. We used 357 patient scans from second-trimester videos and extracted standard frames of 6 heart views (SITUS, 3VV, 3VT, 4CH, LVOT, and RVOT) to create the PULSE heart data.

3.2 CAIFE Heart Sweeps

The *Clinical Artificial Intelligence in Fetal Echocardiography (CAIFE)* study [125], part of the COCHE Centre grant, is an international, multicenter, multidisciplinary initiative designed to curate large fetal cardiac ultrasound datasets and develop artificial intelligence (AI) models for prenatal congenital heart disease (CHD) detection using routine clinical images and standardized cine sweeps.

CAIFE and scanning protocol

The CAIFE study protocol combines routine clinical imaging with standardized research recordings and follows international guidelines for fetal cardiac screening and echocardiography [126–128]. Prospective data collection is embedded within routine obstetric and fetal cardiology appointments between 16 and 40 weeks' gestation, extending the standard visit by approximately ten minutes for additional imaging and documentation. Retrospective data are drawn from scans performed as part of regular clinical care in hospital ultrasound departments, including both general obstetric screening sessions and dedicated fetal cardiology clinics.

Clinical views

Retrospective and prospective data collection includes fetal cardiac still images and short video clips from routine clinical obstetric and fetal cardiology scans, encompassing the following standardized views: cardiac situs (Situs); four-chamber view (4CHV); left ventricular outflow tract (LVOT); three-vessel view (3VV); three-vessel trachea view (3VTV); short-axis view of the left ventricle (SAX LV); short-axis view of the great arteries (SAX GA); sagittal ductal arch (SDA); sagittal aortic arch (SAA); and sagittal bicaval view (SBV). Pulsed-wave Doppler recordings are also collected across all fetal cardiac valves, alongside standard measurements of chambers, valves, and ventricular walls obtained during the fetal cardiology scan. Where available, additional clinical parameters are also recorded [126–128].

Research sweeps (CAIFE)

In the prospective arm, standardized CAIFE cine sweeps are acquired by moving the ultrasound probe across the fetal heart in a sequential, consistent, and controlled manner to ensure comprehensive visualisation of key cardiac structures. Two sweeps are obtained:

1. **Transverse sweep (T-sweep):** a smooth transverse sweep progressing through, in order, the Situs, 4CHV, LVOT, 3VV, and 3VTV views (see Figure 3.2, steps 1–5).

2. **Short-axis/sagittal sweep (S-sweep):** a sweep progressing through SAX LV, SAX GA, SDA, SAA, and SBV views (see Figure 3.2, steps 6–10).

Both T- and S-sweeps are recorded first in two-dimensional B-mode, followed by color Doppler, and finally in dual (split) mode combining B-mode and color on the same image. Each cine sweep lasts up to 10 seconds, and operators may perform up to five attempts in each plane to secure at least one diagnostically informative recording. Data are exported as DICOM files, de-identified locally, and securely transferred to the central CAIFE repository for curation and analysis [128].

Participating clinical sites and ultrasound practice

CAIFE spans five hospitals with obstetric ultrasound and specialist fetal cardiology services:

- **John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust (Oxford, UK)** — a large tertiary teaching hospital with 832 beds. Prospective data were collected in the Maternal Ultrasound Department and Fetal Medicine Unit by research sonographers and fetal cardiologists, while retrospective data were retrieved from ViewPoint and Medcon archives by the clinical research team.

- **Royal Brompton & Harefield Hospitals, Guy's & St Thomas' NHS Foundation Trust (London, UK)** — the UK's largest specialist heart and lung center. Prospective fetal cardiology data were obtained during routine clinical care by fetal cardiologists, and retrospective datasets were extracted from AFGA archives.

- **St George's University Hospitals NHS Foundation Trust (London, UK)** — one of the largest teaching hospitals in Europe, with over 1,300 beds. Prospective fetal cardiology data were collected during routine clinical scans in the Fetal Medicine Unit, and retrospective data were obtained from ViewPoint archives.

- **Chelsea & Westminster Hospital NHS Foundation Trust (London, UK)** — a 430-bed teaching hospital where both prospective and retrospective data were collected by fetal medicine and cardiology specialists during routine clinical care.

- **Gold Coast University Hospital (Southport, Australia)** — a major public and teaching hospital serving the Queensland region. Prospective data were collected by obstetricians and sonographers, and retrospective data were obtained by the clinical research fellow.

Across these sites, multivendor ultrasound systems are in routine use (e.g., GE Voluson, Siemens, Aloka/Hitachi, Toshiba/Canon, Fujifilm). Probe settings and presets vary by platform and operator; clinicians and sonographers typically use at least two probes across gestation (higher nominal frequencies in early gestation; lower in mid/late gestation).

Study governance and contributors

The study was conceived and supervised by J. Alison Noble, with clinical oversight from Aris T. Papageorghiou. Data management was coordinated by Netzahualcoyotl Hernandez-Cruz, and clinical data collection was led by Olga Patey in collaboration with site-based fetal cardiology and ultrasound teams across participating hospitals.

3.2.1 CAIFE Sweep and View-Specific Frame Annotation

For the CAIFE study, each research specific 10 second ultrasound sweep, performed in both the transversal (T sweep) and sagittal/short axis (S sweep) planes, was meticulously annotated on a frame by frame basis with respect to both anatomical

view and frame type. Every frame was first mapped to a specific cardiac view according to established clinical and sonographic guidelines, such as situs, four chamber view (4CHV), left ventricular outflow tract (LVOT), three vessel view (3VV), three vessel trachea view (3VTV), short axis views, and sagittal aortic arch (SAA) [126, 127]. Within each annotated view, frames were further classified as *standard* (displaying the canonical plane as recommended by guidelines), *parastandard* (off axis but containing most of the relevant structures), or *transitional* (captured while moving between parastandard planes). For every frame, detailed annotations captured the presence and appearance of specific cardiac anatomical features relevant to that view, including chamber morphology, septal integrity, valve structure and function, and the spatial relationships of the great vessels, to ensure fine grained, clinically meaningful data curation [125]. This comprehensive annotation scheme enables AI models developed under the CAIFE framework to learn both the structural signatures and the spatiotemporal evolution of fetal cardiac anatomy within continuous ultrasound sweeps.

3.2.2 Normal and Abnormal Scan Annotation

In addition to frame-level annotations, all ultrasound scans in the CAIFE dataset were systematically labeled as either *normal* or *abnormal* following expert clinical review. A scan was annotated as normal if there was no evidence of structural, functional, or rhythm abnormalities, while abnormal scans exhibited one or more confirmed congenital heart defects (CHD), based on clinical diagnosis and detailed visual examination [125]. For abnormal scans, the CHD subtype was also recorded according to a well-defined list of cardiac anomalies. These binary diagnostic labels, combined with the detailed view and frame-level annotations, provide a robust foundation for training and evaluating AI models aimed at distinguishing normal fetal heart anatomy from a spectrum of congenital heart diseases.

3.3 VQ-VCL datasets

3.3.1 Single View VQ-VCL Retrieval Datasets

For the single-view VQ-VCL retrieval task, we curated VQ-VCL datasets from CAIFE [125] and PULSE [5]. The CAIFE dataset comprises transverse fetal cardiac

sweep videos acquired according to the CAIFE protocol, progressing through situ views, four-chambers (4CH), left ventricular outflow tract (LVOT), three-vessel view (3VV) and ending with the view of the three-vessel trachea (3VT). Within this dataset, the retrieval task was defined as identifying and retrieving standard 4CH clips.

The PULSE dataset consists of fetal head sweep videos, from which clips corresponding to the transventricular (TV) and transcerebellar (TC) views were extracted. Here, the retrieval task focused on extracting TV frames using TV visual queries.

For CAIFE, 96 videos were used for training and 10 videos for testing, with visual queries sourced from 11 separate videos, as illustrated in Fig. 3.3(a). For PULSE, the training set comprised 159 videos, the test set included 23 videos, and 8 visual queries were drawn from 8 separate videos, as shown in Fig. 3.3(b).

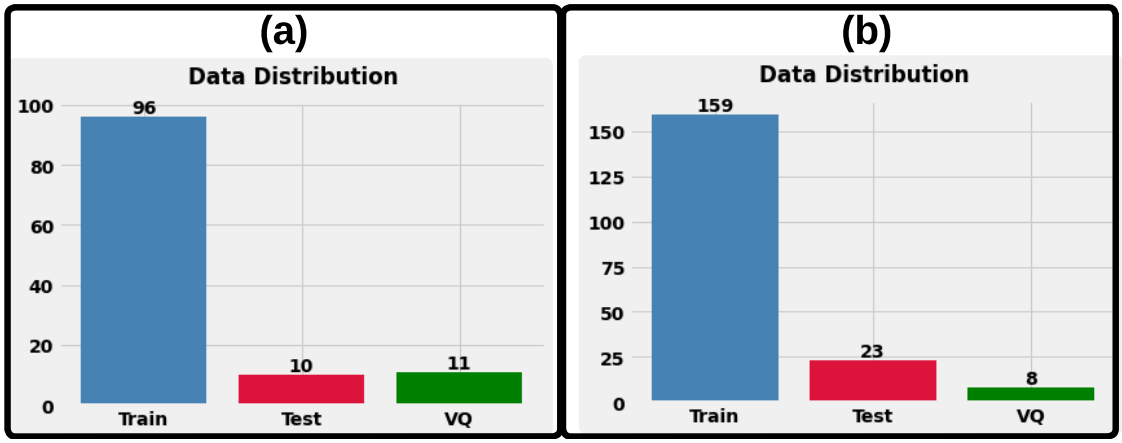


Figure 3.3: (a) Heart sweep data distribution for standard 4CH clip retrieval and (b) PULSE data distribution for TV-frame retrieval in fetal head video clips.

3.3.2 Multi-View VQ-VCL Retrieval Datasets

To extend our single-view VQ-VCL retrieval task [7] to a multi-view setting, we curated multi-view VQ-VCL datasets from CAIFE [125] and PULSE [5].

The CAIFE multi-view dataset consists of transverse fetal cardiac sweep videos acquired according to the CAIFE protocol, spanning multiple standard cardiac views including situs, four-chamber (4CH), left ventricular outflow tract (LVOT), three-vessel view (3VV), and three-vessel trachea view (3VT). In this setting, 200

videos were used for training and 47 videos were used for testing, with visual queries sourced from 12 separate held-out videos, as shown in Fig. 3.4(a).

The PULSE multi-view dataset comprises fetal anomaly scan videos containing multiple clinically relevant anatomical views. For this dataset, 200 videos were used for training and 30 videos were used for testing, with visual queries extracted from the same 30 videos, as illustrated in Fig. 3.4(b).

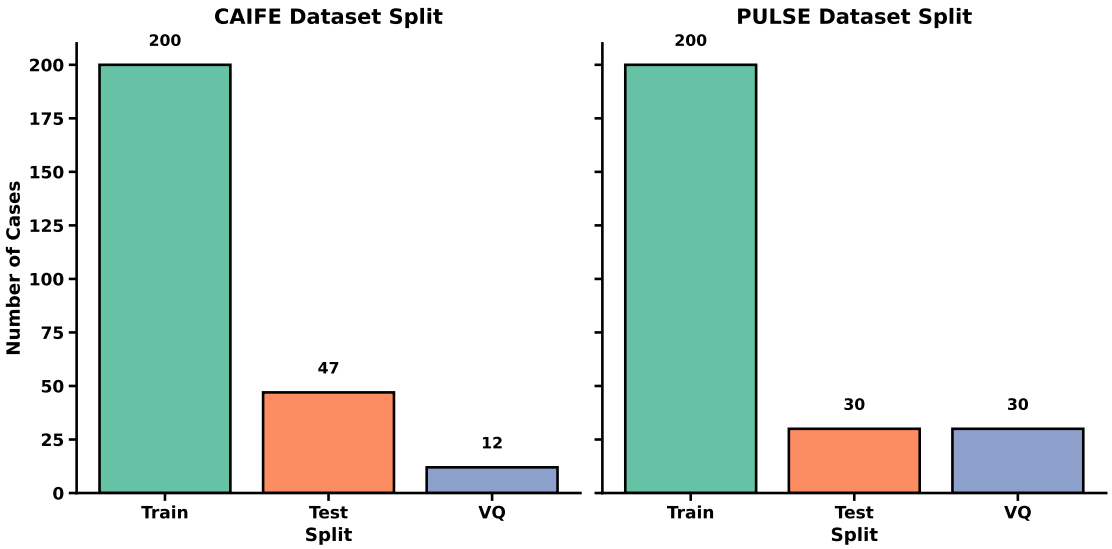


Figure 3.4: Dataset split and visual query distribution for the multi-view VQ-VCL retrieval task. (a) CAIFE multi-view dataset showing the number of training videos, test videos, and visual query source videos. (b) PULSE multi-view dataset illustrating the corresponding distribution of training videos, test videos, and visual query sources.

3.4 Multi-Site CHD Detection Datasets

In our paper, *Self-supervised Normality Learning and Divergence Vector-guided Model Merging for Zero-shot Congenital Heart Disease Detection in Fetal Ultrasound Videos* [10], we utilized datasets collected from five different CAIFE clinical sites for training, internal validation, and external validation. The following subsections describe the data collected from each participating site.

3.4.1 Site01

This dataset comprises retrospective and prospective CAIFE sweeps and cineloops collected at John Radcliffe Hospital, Oxford University Hospitals NHS Foundation

Trust (Oxford, UK). A total of 8,878 normal videos were used for model training, while 604 videos (462 normal and 142 abnormal) were used for validation and 667 videos (536 normal and 131 abnormal) were used for testing, as shown in Fig. 3.5.

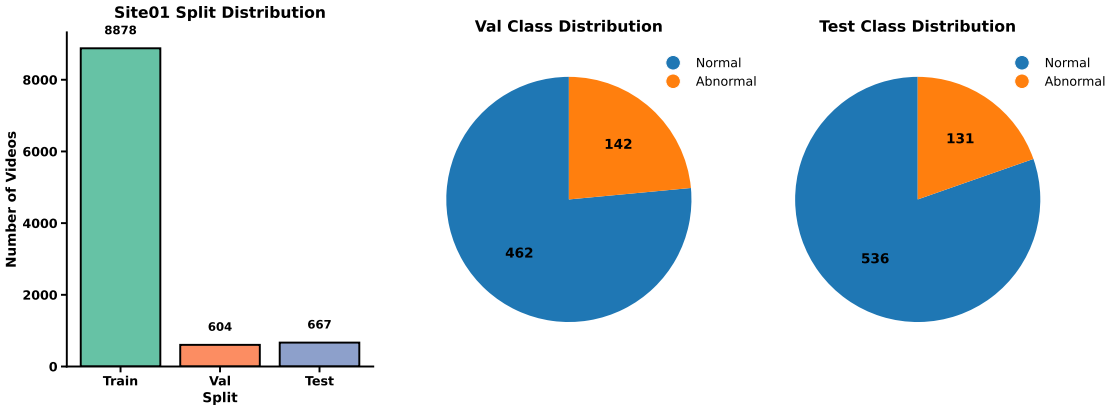


Figure 3.5: Dataset distribution for Site01 dataset.

3.4.2 Site02

The Site02 dataset was collected from St George's University Hospitals NHS Foundation Trust (London, UK) and used for training, validation, and testing following the same protocol as Site01. The training set consists of 16,074 videos containing exclusively normal cases. For evaluation, the validation set includes 2,218 videos (864 normal and 1,354 abnormal), while the test set includes 2,254 videos (915 normal and 1,339 abnormal), as illustrated in Fig. 3.6.

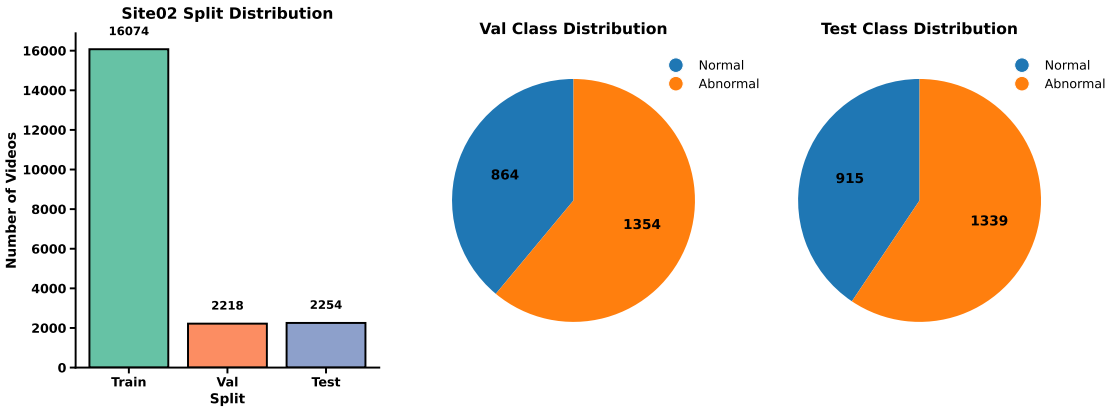


Figure 3.6: Dataset distribution for Site02 dataset.

3.4.3 Site03

The Site03 dataset was collected from Chelsea & Westminster Hospital NHS Foundation Trust (London, UK) and was used exclusively for external validation. Although the dataset distribution is shown in Fig. 3.7, only the test set was used for external evaluation. The dataset comprises 377 normal videos in the training split, 28 videos in the validation split (21 normal and 7 abnormal), and 29 videos in the test split (21 normal and 8 abnormal).

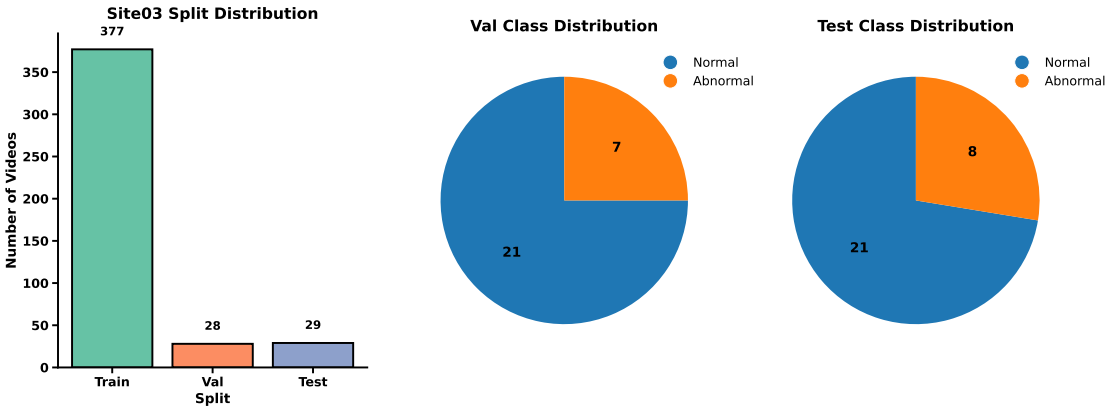


Figure 3.7: Dataset distribution for Site03 dataset.

3.4.4 Site04

The Site04 dataset was collected from Gold Coast University Hospital (Southport, Australia) and was used exclusively for external validation. While the full dataset distribution is illustrated in Fig. 3.8, only the test set was used for external evaluation. The dataset includes 93 normal videos in the training split, 17 videos in the validation split (5 normal and 12 abnormal), and 18 videos in the test split (6 normal and 12 abnormal).

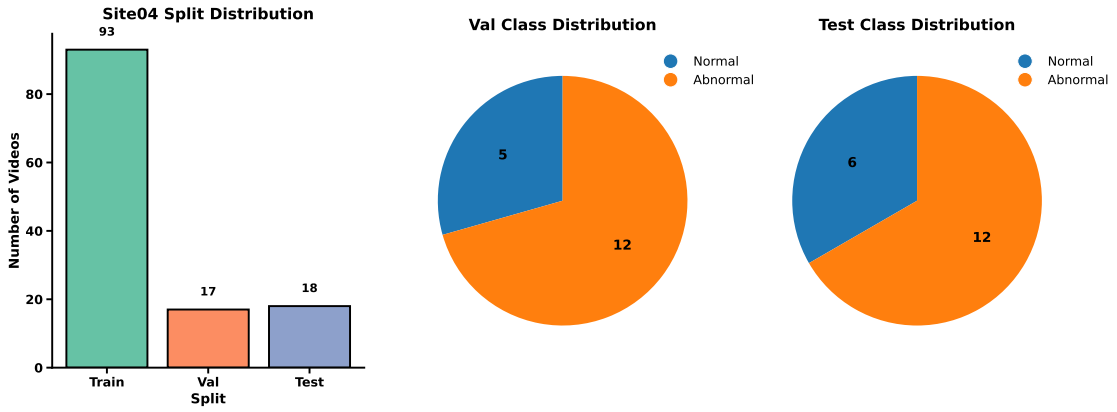


Figure 3.8: Dataset distribution for Site04 dataset.

3.4.5 Site05

The Site05 dataset was collected from Royal Brompton & Harefield Hospitals, Guy's & St Thomas' NHS Foundation Trust (London, UK) and followed the same training, validation, and testing design as Site01 and Site02. The training set comprises 1,573 videos consisting exclusively of normal cases. In contrast, both the validation and test sets exhibit a pronounced class imbalance, with a substantially higher proportion of abnormal cases. Specifically, the validation set includes 810 videos (87 normal and 723 abnormal), while the test set includes 816 videos (80 normal and 736 abnormal), as illustrated in Fig. 3.9.

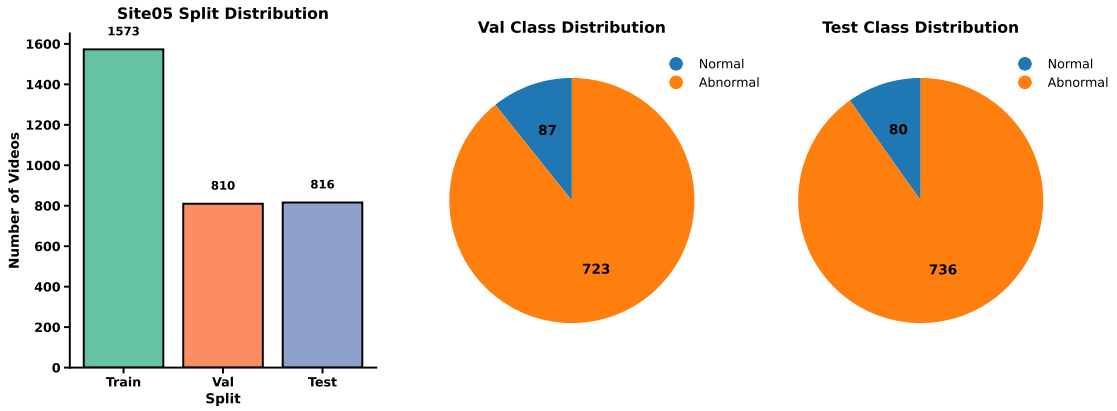


Figure 3.9: Dataset distribution for Site05 dataset.

3.5 FetalEcho1

FetalEcho1 and FetalEcho2 (described in the following section) are curated subsets of CAIFE data used specifically for the self-supervised learning experiments in Chapter 9.

FetalEcho1 is a subset dataset comprising only cases collected at John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust (Oxford, UK), both retrospectively and prospectively. For model training, we used a combination of retrospective and prospective CAIFE sweeps, with training performed exclusively on normal cases. In contrast, validation and testing used only prospective sweeps that strictly followed the CAIFE protocol and comprised both normal and abnormal cases. The dataset included 8,273 videos for training and 414 videos for validation (202 normal and 212 abnormal) and testing (151 normal and 166 abnormal), as shown in Fig. 3.10.

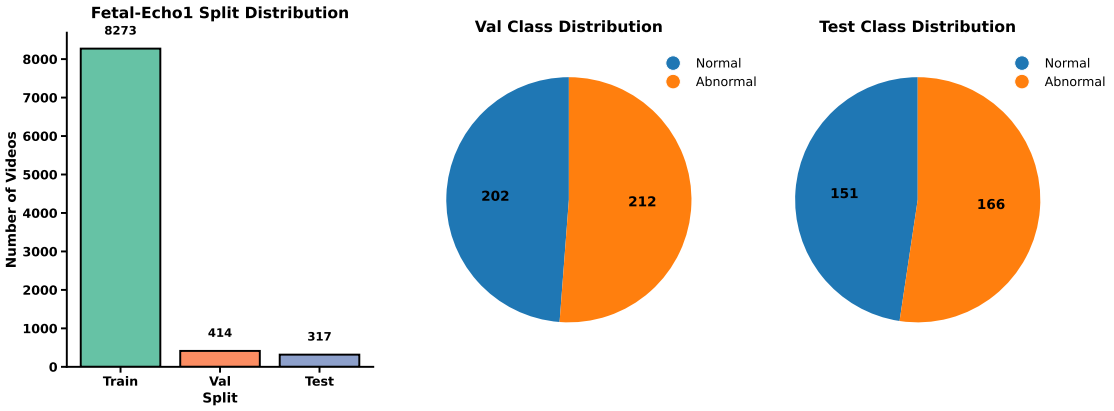


Figure 3.10: Dataset distribution for Fetal-Echo1 dataset.

3.6 FetalEcho2

FetalEcho2 was collected from St George's University Hospitals NHS Foundation Trust (London, UK). Training data exclusively comprised of normal cases, while the validation and test sets comprised both normal and abnormal cases. The dataset included 4,154 videos for training, 320 videos for validation (150 normal and 170 abnormal), and 305 videos for testing (173 normal and 132 abnormal) as shown in Fig. 3.11.

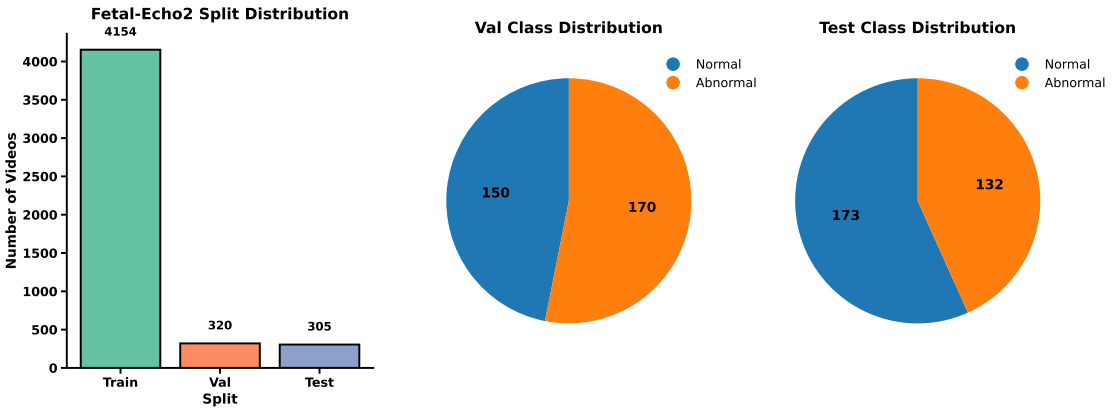


Figure 3.11: Dataset distribution for Fetal-Echo2 dataset.

3.7 EchoNet Dynamic Anomaly

EchoNet Dynamic [129] is a large-scale public dataset of adult apical four-chamber (4CH) echocardiography videos, each annotated with ejection fraction (EF), ventricular volumes, and expert ventricle tracings. To enable research on automated cardiac anomaly detection, we curated the **EchoNet Dynamic Anomaly** dataset from the original collection. In this new dataset, each video is labeled as *normal* or *abnormal* based on ejection fraction: samples with EF $< 45\%$ or EF $> 75\%$ are considered abnormal, reflecting clinically significant cardiac dysfunction. The training set comprises 7,378 videos containing only normal cases, while the validation set includes 1,326 videos (661 normal and 665 abnormal) and the test set includes 1,326 videos (642 normal and 684 abnormal), as shown in Fig. 3.12.

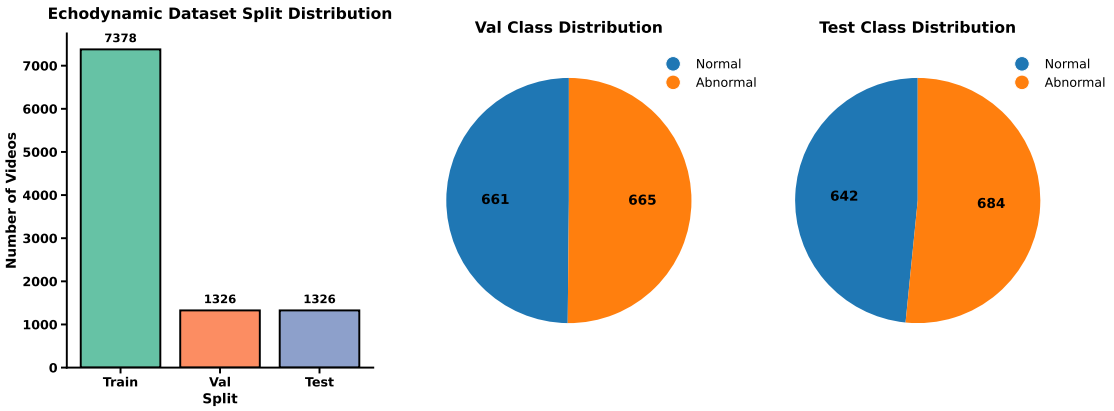


Figure 3.12: Dataset distribution for Echo-Dynamic dataset.

3.8 EchoNet Pediatric LVH Anomaly

EchoNet Pediatric [130] (EchoPediatric LVH) comprises parasternal long-axis echocardiography videos from pediatric patients, annotated with key measurements such as EF and ventricular volumes. Recognizing the unique challenges in pediatric cardiac assessment, we curated the **EchoNet Pediatric LVH Anomaly** dataset from the original resource. Videos are classified as *normal* or *abnormal* using the same ejection fraction thresholds (<45% or >75%) to identify potential cardiac dysfunction in children. The training set contains 7,837 videos consisting exclusively of normal cases, whereas the validation set comprises 1,592 videos (679 normal and 913 abnormal) and the test set comprises 1,326 videos (705 normal and 887 abnormal), as illustrated in Fig. 3.13.

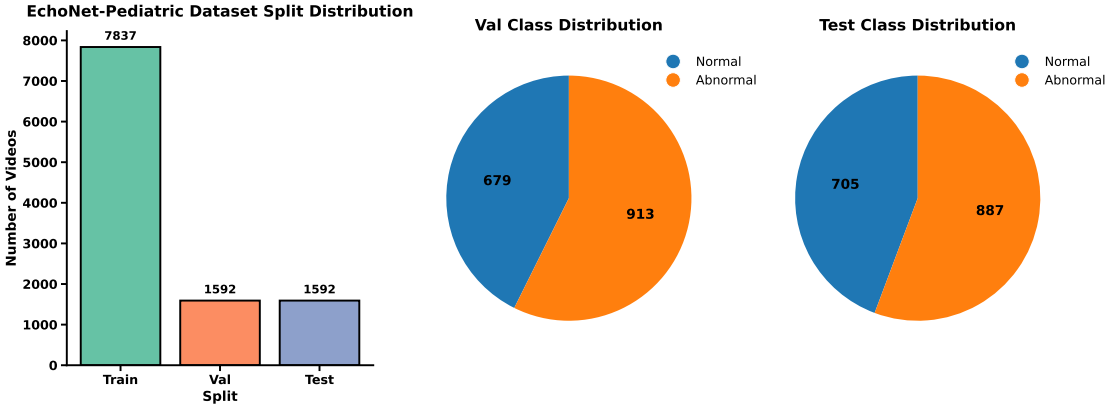


Figure 3.13: Dataset distribution for Echo-Pediatric dataset.

3.9 RVENet Anomaly

RVENet [131] is a specialized dataset with right ventricular echocardiography videos from pediatric and adult patients; it serves as a crucial source for studying RV function. For anomaly detection research, we curated the **RVENet Anomaly** dataset from RVENet. Here, each video is designated as *normal* or *abnormal* based on ejection fraction values, with EF $<$ 45% or EF $>$ 75% flagged as abnormal, supporting the identification of right ventricular dysfunction using widely accepted clinical cutoffs. The training set includes 2,516 videos consisting exclusively of normal cases, while the validation set contains 487 videos (258 normal and 229

abnormal) and the test set contains 573 videos (279 normal and 294 abnormal), as illustrated in Fig. 3.14.

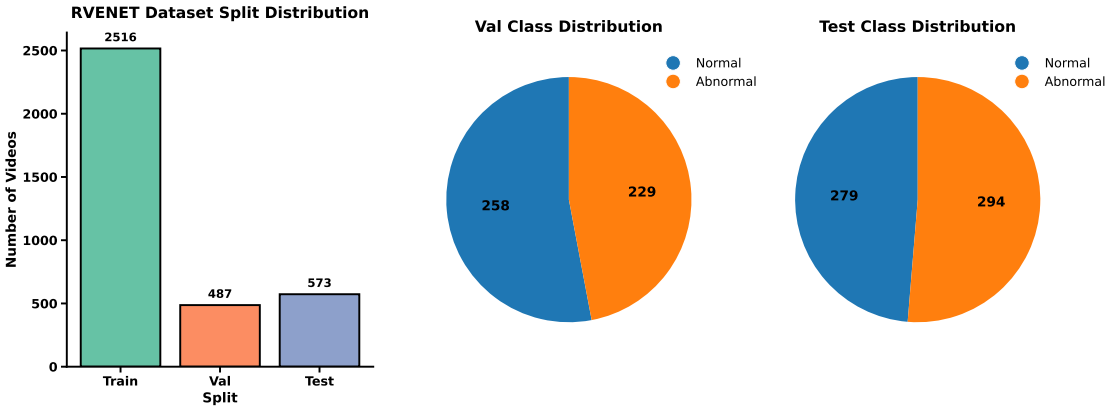


Figure 3.14: Dataset distribution for RVENET dataset.

4

Dual Conditioned Diffusion Models for Out-Of-Distribution Detection: Application to Fetal Ultrasound Videos

Background: In standard clinical practice, fetal ultrasound scans are performed to capture approximately 15 different anatomical views in a single session. If we want to analyze a single anatomy in detail, for example, the heart, it becomes time-consuming and inefficient to manually go through every frame in these long video recordings. This task is complicated by the visual similarity between heart views and other anatomies, such as the abdomen or thorax, as well as substantial variation both within and between standard heart views. This makes the automated separation of heart frames from other regions both important and challenging. To address this, we approach the problem as out-of-distribution (OOD) detection: we treat the target anatomy, in this case heart views, as in-distribution and all other anatomies as OOD. We present dual conditioned diffusion models (DCDM) which guide the generative process using both in-distribution (ID) class information and latent image features, enabling accurate separation of ID and OOD frames based on reconstruction, without needing any labeled OOD data. Although we focus on the identification of the heart views in this work, the approach can be applied to any anatomy of interest. In our experiments, DCDM outperforms previous methods, showing a 12% improvement in accuracy, 22% higher precision, and an

8% increase in F1 score, highlighting its value for efficient and automated clip identification in routine fetal ultrasound.

Authors: Divyanshu Mishra, He Zhao, Pramit Saha, Aris T. Papageorghiou, J.Alison Noble

Published in Conference: Mishra, D., Zhao, H., Saha, P., Papageorghiou, A.T. and Noble, J.A., 2023, October. Dual conditioned diffusion models for out-of-distribution detection: Application to fetal ultrasound videos. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 216-226). Cham: Springer Nature Switzerland.

Author Contribution: I was the lead technical author of the paper, responsible for formulating the problem statement, proposing the solution, designing the codebase, conducting the experiments, and preparing the original manuscript draft. He Zhao contributed to technical discussions, while Pramit Saha contributed to the overall discussion. J. Alison Noble provided both technical guidance and overall supervision. All authors reviewed and approved the final version of the manuscript.

Abstract

Out-of-distribution (OOD) detection is essential to improve the reliability of machine learning models by detecting samples that do not belong to the training distribution. Detecting OOD samples effectively in certain tasks can pose a challenge because of the substantial heterogeneity within the in-distribution (ID), and the high structural similarity between ID and OOD classes. For instance, when detecting heart views in fetal ultrasound videos there is a high structural similarity between the heart and other anatomies such as the abdomen, and large in-distribution variance as a heart has 5 distinct views and structural variations within each view. To detect OOD samples in this context, the resulting model should generalise to the intra-anatomy variations while rejecting similar OOD samples. In this paper, we introduce dual-conditioned diffusion models (DCDM) where we condition the model on in-distribution class information and latent features of the input image for reconstruction-based OOD detection. This constrains the generative manifold of the model to generate images structurally and semantically similar to those within the in-distribution. The proposed model outperforms reference methods with a 12% improvement in accuracy, 22% higher precision, and an 8% better F1 score.

4.1 Introduction

Existing out-of-distribution (OOD) detection methods work well when the in-distribution (ID) classes have low heterogeneity (low variance) but fail when in-distribution classes have high heterogeneity [132] or high spatial similarity between ID and OOD classes [133]. Fetal ultrasound (US) anatomy detection is one such application where both the challenges co-exist.

In this paper, we propose a Dual-Conditioned Diffusion Model (DCDM) to detect OOD samples when in-distribution data has high variance and test the performance by detecting heart views in fetal US videos as an example application. Specifically, an ultrasound (US) typically comprises 13 anatomies and their views. However,analysis models are usually developed for anatomy-specific tasks. Hence, to separate heart views from other 12 anatomies (head, abdomen, femur etc) we develop an OOD detection algorithm. Our in-distribution data comprises five structurally different heart views captured across different cardiac cycles of a beating heart during obstetric US scanning. We develop a diffusion-based model for reconstruction-based OOD detection, which extends [134] with a novel dual conditioning mechanism that alleviates the influence of high inter- and intra-class variation within different classes by leveraging in-distribution class conditioning (IDCC) and latent image feature conditioning (LIFC). These conditioning mechanisms allow our model to generate images similar to the input image for in-distribution data. The primary contributions of our paper are summarized as follows:

1. We introduce a novel conditioned diffusion model for OOD detection and demonstrate that the dual conditioning mechanism is effective in tackling challenging scenarios where in-distribution data comprises multiple heterogeneous classes and there is a high spatial similarity between ID and OOD classes.

2. Two original conditions are proposed for the diffusion model, which are in-distribution class conditioning (IDCC) and latent image feature conditioning (LIFC). IDCC is proposed to handle high inter-class variance within in-distribution classes and high spatial similarity between ID and OOD classes. LIFC is introduced to counter the intra-class variance within each class.

3. We demonstrate in our experiments that DCDM can detect and separate heart views from other anatomies in fetal ultrasound videos without needing any labeled data for OOD classes. Extensive experiments and ablations demonstrate superior performance over existing OOD detection methods. Our approach is not fetal ultrasound specific and could be applied to other OOD applications.

4.2 Related Work

OOD detection [135] involves identifying samples that do not belong to the training distribution. Such models can be categorized into: (a) unsupervised OOD detection [132] and (b) supervised OOD detection. [136–138]. Unsupervised OOD detection methods can again be divided into two main categories: (i) likelihood-based approaches [139–141], and (ii) reconstruction-based [142–144]. Likelihood-based approaches suffer from several issues, including assigning higher likelihood to OOD samples [145, 146], susceptibility to adversarial attacks [147], and calibration issues [148]. Current reconstruction-based approaches are sensitive to dimensions of the bottleneck layer and require rigorous tuning specific to the dataset and task [149]. Additionally, models trained using a generator-discriminator architecture and optimizing adversarial losses can be highly unstable and challenging to train [150, 151]. Finally, reconstruction-based methods often rely on highly compressed latent representations, which can lead to loss of important low-level detail. This can be problematic when discriminating between classes with high spatial similarity. Recently, diffusion models have been introduced to address these limitations on tasks such as image synthesis [152], and OOD detection [149].

Denoising Diffusion Probabilistic Models (DDPMs) [134] are generative models that work by gradually adding noise to an input image through a forward diffusion process followed by gradually removing noise using a trained neural network in the backward diffusion process [153].

To guide the generative process of a diffusion model (DM), previous works [154–156] condition the DDPMs on task-specific conditioning. In image-to-image translation tasks like super-resolution, colorization, *etc.*, previous papers [155]

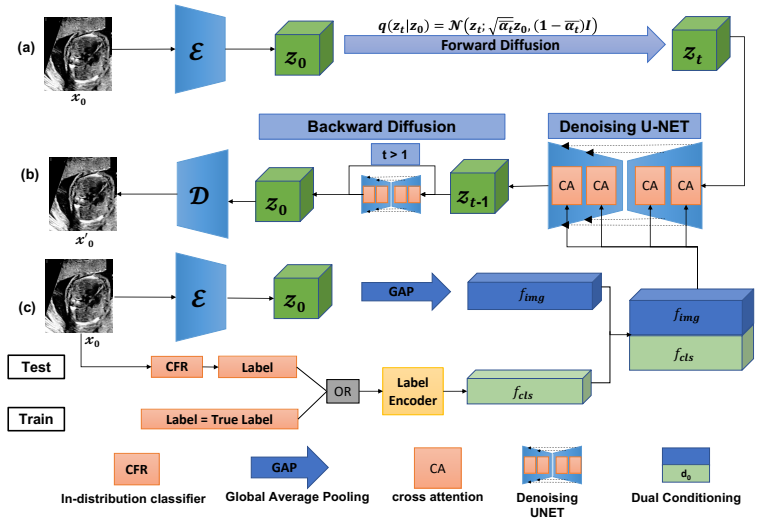


Figure 4.1: DCDM architecture where **(*a*)** the input image x_0 is mapped to the latent vector z_0 using a pretrained encoder $\mathcal{E}$ and forward diffusion is applied, **(*b*)** the backward diffusion process denoises the latent vector z_t and the final denoised latent vector z_0 is mapped to pixel space by the decoder $\mathcal{D}$ **(*c*)** the dual-conditioning mechanism. We obtain f_{img} by passing the input image x_0 through the encoder $\mathcal{E}$. f_{cls} is obtained using the true label during training or predicted class label during testing.

condition the model by concatenating a resized or grayscale version of the input image to the noised image. This concatenation is unsuitable for reconstruction-based OOD detection as the model will generate similar images for ID and OOD samples. In the context of OOD detection using DMs, previous works [149] have trained unconditional DDPMs and, during inference, sampled using a Pseudo Linear Multi Step (PLMS) [157] sampler for varying noise levels. However, their approach generates 5500 samples to detect OOD samples for each input image which is time-consuming and impractical for settings where shorter inference times are needed. AnoDDPM [158] utilizes simplex noise rather than Gaussian noise to corrupt the image (t=250 rather than t=1000) for anomaly detection. However, this approach requires data specific tuning,and is outperformed by [149].

4.3 Methods

4.3.1 Dual Conditioned Diffusion Models

Diffusion models are generative models that rely on two Markov processes known as forward and backward diffusion [134]. To improve efficiency during training

and inference, forward and backward diffusion is applied to the latent space [156]. Autoencoder (AE = $\mathcal{E}$ + $\mathcal{D}$) is pretrained separately on ID heart data and can successfully reconstruct the input heart images (SSIM=0.956). The latent variable z_0 is obtained by passing an input image x_0 through a pretrained encoder $\mathcal{E}$. Given the latent vector z_0 and a fixed variance schedule [134] $\{\beta_t \in (0,1)\}_{t=1}^{T}$, the forward diffusion process, defined by Eqn. 4.1, gradually adds Gaussian noise to z_0 to give a noised latent vector z_t where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{4.1}$$

In backward diffusion, we aim to reverse the forward diffusion process and predict z_{t-1} given z_t. To predict $(z_{t-1}|z_t)$,we train a denoising U-Net [134] denoted as $\boldsymbol{\epsilon}_\theta\left(z_t, t, d_0\right)$ that takes the current timestep t, noised latent vector z_t and the dual conditioning embedding vector d_0 as input and predicts the noise at timestep t as shown in Eqn. 4.2.

$$\boldsymbol{z}_{t-1} = \mathcal{N}(\boldsymbol{z}_{t-1}; \frac{1}{\sqrt{\alpha_t}}\Big(\boldsymbol{z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, d_0)\Big), (1-\bar{\alpha}_t)\mathbf{I}) \tag{4.2}$$

The dual embedding vector d_0 is obtained by combining IDCC (f_{cls}) and LIFC (f_{img}) vectors, which we explain in Section 3.2. The output z_{t-1} is again input to ϵ_θ. This process is repeated until z_0 is obtained.

The final model optimisation objective is given by Eqn. 4.3 where ϵ is the original noise added during the forward diffusion process.

$$\mathcal{L}_{DCDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, d_0\right)\|_2^2\right] \tag{4.3}$$

Once we obtain z_0 from the backward diffusion process, it is passed on to the decoder $\mathcal{D}$ and mapped back to the pixel space to give generated image x_0'.

4.3.2 Dual Conditioning Mechanism

Image features and in-distribution class information are utilized in our proposed dual conditioning mechanism. This guides the DCDM to generate images that are spatially and semantically similar to the input image for in-distribution samples and dissimilar for OOD samples.

Latent Image Feature Conditioning (LIFC): The image conditioning dictates the desired appearance of generated images in terms of shape and texture. In our model, we use the features extracted by a pretrained encoder for conditioning. Empirically, we use the same encoder $\mathcal{E}$ as our feature extractor to obtain latent feature vector z_0 as shown in Fig. 4.1. Specifically, the input image of dimension $224 \times 224 \times 3$ is passed through the encoder $\mathcal{E}$ and a feature map with the size of $7 \times 7 \times 128$ is obtained which is followed by global average pooling (GAP) resulting in a feature vector (f_{img}) with dimension 128.

In-Distribution Class Conditioning (IDCC): Given an in-distribution dataset comprising n heterogeneous classes, conditioning the model only on image-level features is insufficient. Therefore we introduce an in-distribution class conditioning (IDCC) that informs the DCDM of the class of the input image and enables it to generate samples belonging to the same class for ID. A label encoder generates a unique class conditional embedding (f_{cls}) of dimension 128 for each class label. The class label is assigned based on the ground truth label during the training phase and to the classifier's prediction during inference, as depicted in Fig. 4.1. In practice, we train a CNN classifier, freeze its weight and use it as our in-distribution classifier (CFR), as discussed in section 3.3.

Cross Attention Guidance: To integrate the dual-conditioning guidance into the diffusion model, we use a cross-attention [159] mechanism inside the denoising U-Net rather than just concatenation [155] as it is more effective [160–162] and allows condition diffusion models on various input modalities [156]. Our LIFC and IDCC are first concatenated to give a feature vector with a dimension of 256. This acts as a side input to each UNet block. The features from the UNet block and the conditional features are fused by cross-attention and serve as input to the following UNet block as shown in Fig. 4.1.For more details,regarding cross-attention block refer to Rombach *et al* [156].

4.3.3 In-Distribution Classifier

The in-distribution classifier (CFR) serves two main functions. First, it provides labels for the class conditioning during inference; second, it is utilized as a feature extractor for calculating the OOD score.

Inference Class Guidance. IDCC requires in-distribution class information to generate the class conditional embedding. However, class information is only available during training. To obtain class information during inference, we separately train a ConvNext CNN based classifier (accuracy = 88%) on the in-distribution data and use its predictions as the class information. During inference, the input image x_0 is passed through the classifier, and the predicted label is used to generate the class embedding by feeding to the label encoder as shown in Fig. 4.1. Moreover, as the classifier is only trained on in-distribution data, it classifies an OOD sample to an in-distribution class. The classifier's prediction is utilized by the DCDM and it tries to generate an image belonging to in-distribution class for the OOD samples. This reduces the structural and semantic similarity between the input and the generated image, as demonstrated by our qualitative results (Fig. 4.2).

Feature-Based OOD Detection To evaluate the performance of the DCDM, the cosine similarity between features of the input image x_0 and the generated image x_0' from the in-distribution classifier is calculated and is referred as an OOD score where f_0 and f_0' are the features of x_0 and x_0', respectively:

$$\text{OOD score} = sim(f_0, f_0') \ = \frac{f_0 \cdot f_0'}{\|f_0\|_2 \, \|f_0'\|_2}, \tag{4.4}$$

An input image x_0 is classified as in-distribution (ID) or OOD based on Eqn. 4.5 where τ is a pre-defined threshold and y_{pred} is the prediction of our feature-based OOD detection algorithm.

$$y_{pred} = \begin{cases} 0(ID) & if \text{ OOD score} > \tau \\ 1 \text{ (OOD)} & otherwise \end{cases} \tag{4.5}$$

4.4 Experiments and Results

Dataset and Implementation

For our experiments, we utilized a fetal ultrasound dataset of 359 subject videos that were collected as part of the PULSE project [5]. The in-distribution dataset consisted of 5 standard heart views (3VT, 3VV, LVOT, RVOT, and 4CH), while the out-of-distribution dataset comprised of three non-heart anatomies - fetal head,

abdomen, and femur. The original images were of size 1008×784 pixels and were resized to 224×224 pixels.

To train the models, we randomly sampled 5000 fetal heart images and used 500 images for evaluating image generation performance. To test the performance of our final model and compare it with other methods, we used an held-out dataset of 7471 images, comprising 4309 images of different heart views and 3162 images (about 1000 for each anatomy) of out-of-distribution classes. Further details about the dataset are given in **Supp. Fig. 4.5 and 4.6**.

All models were trained using PyTorch version 1.12 with a Tesla V100 32 GB GPU. During training, we used T=1000 for noising the input image and a linearly increasing noise schedule that varied from 0.0015 to 0.0195. To generate samples from our trained model, we used DDIM [163] sampling with T=100. All baseline models were trained and evaluated using the original implementation.

4.4.1 Results

We evaluated the performance of the dual-conditioned diffusion models (DCDMs) for OOD detection by comparing them with two current state-of-the-art unsupervised reconstruction-based approaches and one likelihood-based approach. The first baseline is Deep-MCDD [164], a likelihood-based OOD detection method that proposes a Gaussian discriminant-based objective to learn class conditional distributions. The second baseline is ALOCC [132] a GAN-based model that uses the confidence of the discriminator on reconstructed samples to detect OOD samples. The third baseline is the method of Graham *et al.* [149], where they use DDPM [134] to generate multiple images at varying noise levels for each input. They then compute the MSE and LPIPS metrics for each image compared to the input, convert them to Z-scores, and finally average them to obtain the OOD score.

Quantitative Results The performance of the DCDM, along with comparisons with the other approaches, are shown in Table 4.1. The GAN-based method ALOCC [132] has the lowest AUC of 57.22%, which is improved to 63.86% by the method of Graham *et al.* and further improved to 64.58% by likelihood-based Deep-MCDD. DCDM outperforms all the reference methods by 20%, 14% and 13%, respectively and has an AUC of 77.60%. High precision is essential for OOD

Table 4.1: Quantitative comparison of our model (DCDM) with reference methods

Method	AUC(%)	F1-Score(%)	Accuracy(%)	Precision(%)
Deep-MCDD [164]	64.58	66.23	60.41	51.82
ALOCC [132]	57.22	59.34	52.28	45.63
Graham et al. [149]	63.86	63.55	60.15	50.89
DCDM(Ours)	**77.60**	**74.29**	**77.95**	**73.34**

detection as this can reduce false positives and increase trust in the model. DCDM exhibits a precision that is 22% higher than the reference methods while still having an 8% improvement in F1-Score.

Qualitative Results Qualitative results are shown in Fig. 4.2. Visual comparisons show ALOCC generates images structurally similar to input images for in-distribution and OOD samples. This makes it harder for the ALOCC model to detect OOD samples. The model of Graham *et al.* generates any random heart view for a given image as a DDPM is unconditional, and our in-distribution data contains multiple heart views. For example, given a 4CH view as input, the model generates an entirely different heart view. However, unlike ALOCC, the Graham *et al.* model generates heart views for OOD samples, improving OOD detection performance. DCDM generates images with high spatial similarity to the input image and belonging to the same heart view for ID samples while structurally diverse heart views for OOD samples. In Fig 4.2 (c) for OOD sample, even-though the confidence is high (0.68), the gap between ID and OOD classes is wide enough to separate the two.Additional qualitative results can be observed in **Supp. Fig. 4.7**

4.4.2 Ablation Study

Ablation experiments were performed to study the impact of various conditioning mechanisms on the model performance both qualitatively and quantitatively. When analyzed quantitatively, as shown in Table 4.2, the unconditional model has the lowest AUC of 69.61%. Incorporating the IDCC guidance or LIFC separately, improves performance with an AUC of 75.27% and 77.40%, respectively. The best results are achieved when both mechanisms are used (DCDM), resulting in an 11% improvement in the AUC score relative to the unconditional model. Although there is a small margin of performance improvement between the combined model (DCDM)

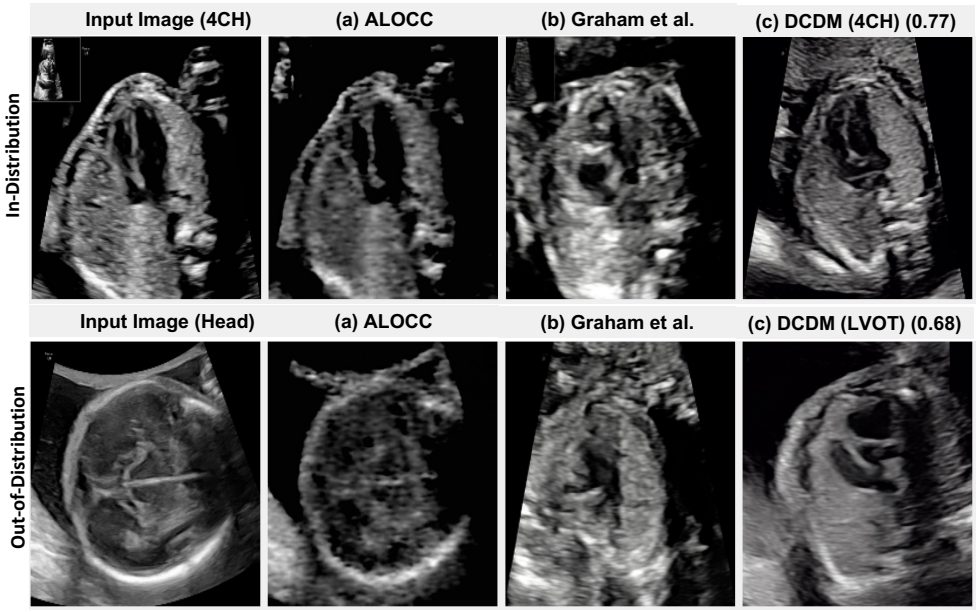


Figure 4.2: C
oQualitative comparison of our method with **(*a*)** ALOCC generates similar images to the input for ID and OOD samples **(*b*)** Graham *et al.* generates any random heart view for a given input image **(*c*)** Our model generates images that are similar to the input image for ID and dissimilar for OOD samples. Classes predicted by CFR and the OOD score ($\tau = 0.73$) are mentioned in brackets.

and the LIFC model in terms of AUC, the precision improves by 3%, demonstrating the combined model is more precise and hence the best model for OOD detection.

As shown in Fig. 4.3, the unconditional diffusion model generates a random heart view for a given input for both in-distribution and OOD samples. The IDCC guides the model to generate a heart view according to the in-distribution classifier (CFR) prediction which leads to the generation of similar samples for in-distribution input while dissimilar samples for OOD input. On the other hand, LIFC generates an image with similar spatial information. However, heart views are still generated for OOD samples as the model was only trained on them. When dual-conditioning (DC) is used, the model generates images that are closer aligned to the input image for in-distribution input and high-fidelity heart views for OOD than those generated by a model conditioned on either IDCC or LIFC alone. **Supp. Fig. 4.4** presents further qualitative ablations.

Table 4.2: Ablation study of different conditioning mechanisms of DCDM.

Method	Accuracy (%)	Precision (%)	AUC (%)
Unconditional	68.16	58.44	69.61
In-Distribution Class Conditioning	74.39	66.12	75.27
Latent Image Feature Conditioning	77.02	70.02	77.40
Dual Conditioning	**77.95**	**73.34**	**77.60**

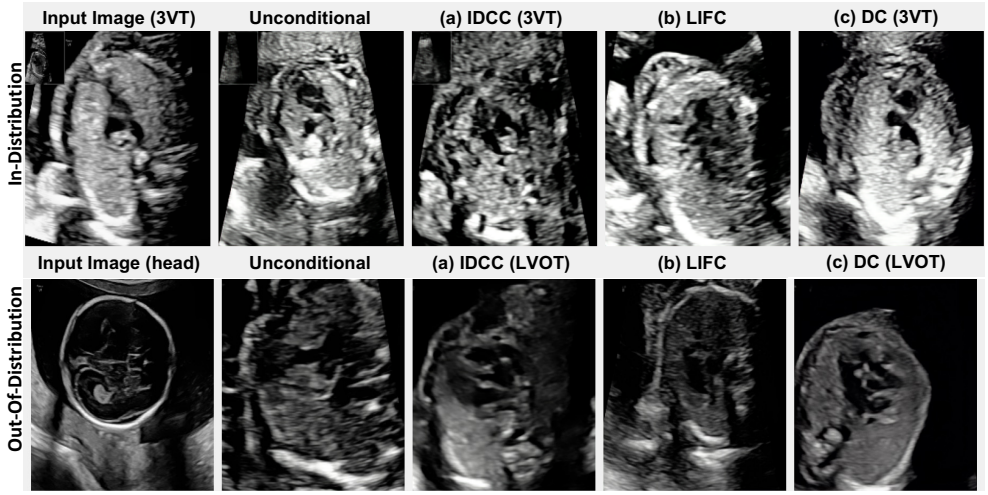


Figure 4.3: Qualitative ablation study showing the effect of **(a)** IDCC, **(b)** LIFC and, **(c)** DC on generative results of DM. Brackets in IDCC, DC show labels predicted by CFR.

4.5 Conclusion

We introduce novel dual-conditioned diffusion model for OOD detection in fetal ultrasound videos and demonstrate how the proposed dual-conditioning mechanisms can manipulate the generative space of a diffusion model. Specifically, we show how our dual-conditioning mechanism can tackle scenarios where the in-distribution data has high inter- (using IDCC) and intra- (using LIFC) class variations and guide a diffusion model to generate similar images to the input for in-distribution input and dissimilar images for OOD input images. Our approach does not require labeled data for OOD classes and is especially applicable to challenging scenarios where the in-distribution data comprises more than one class and there is high similarity between the in-distribution and OOD classes.

Supplementary Material

Figure 4.4: Qualitative ablation: for each row from left to right—input image (ground-truth label), unconditional, IDCC (prediction), LIFC, DC (prediction).

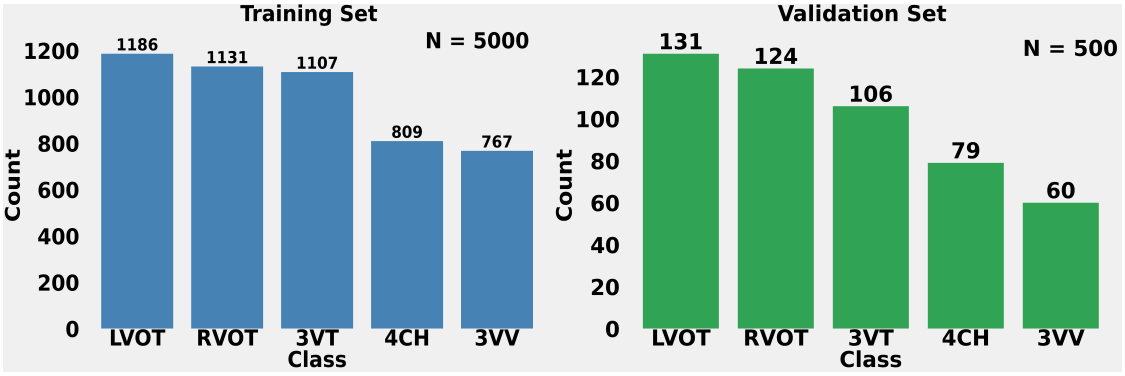


Figure 4.5: Dataset distribution by heart class in the training (left) and validation (right) sets.

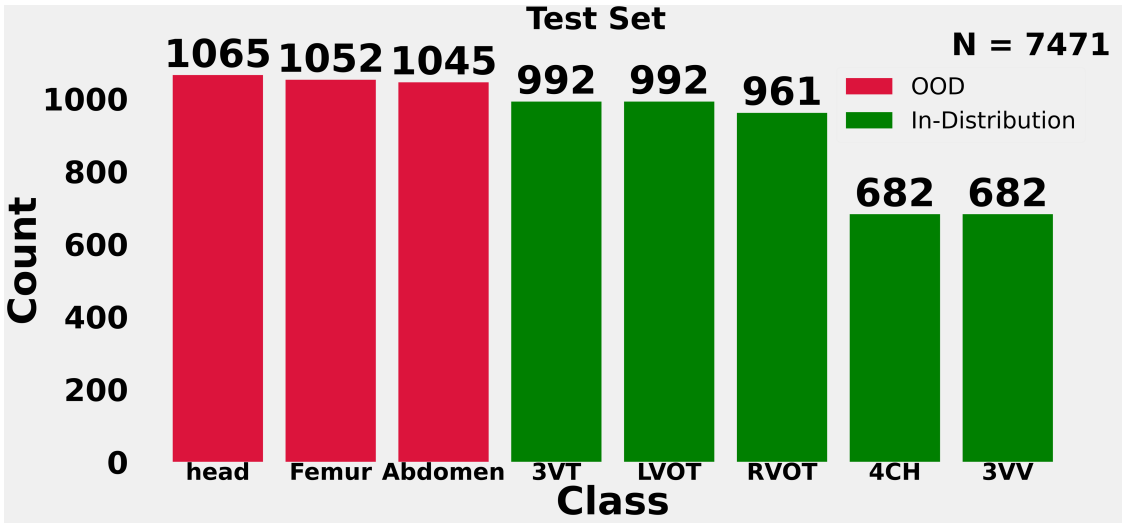


Figure 4.6: Dataset distribution of the test set.

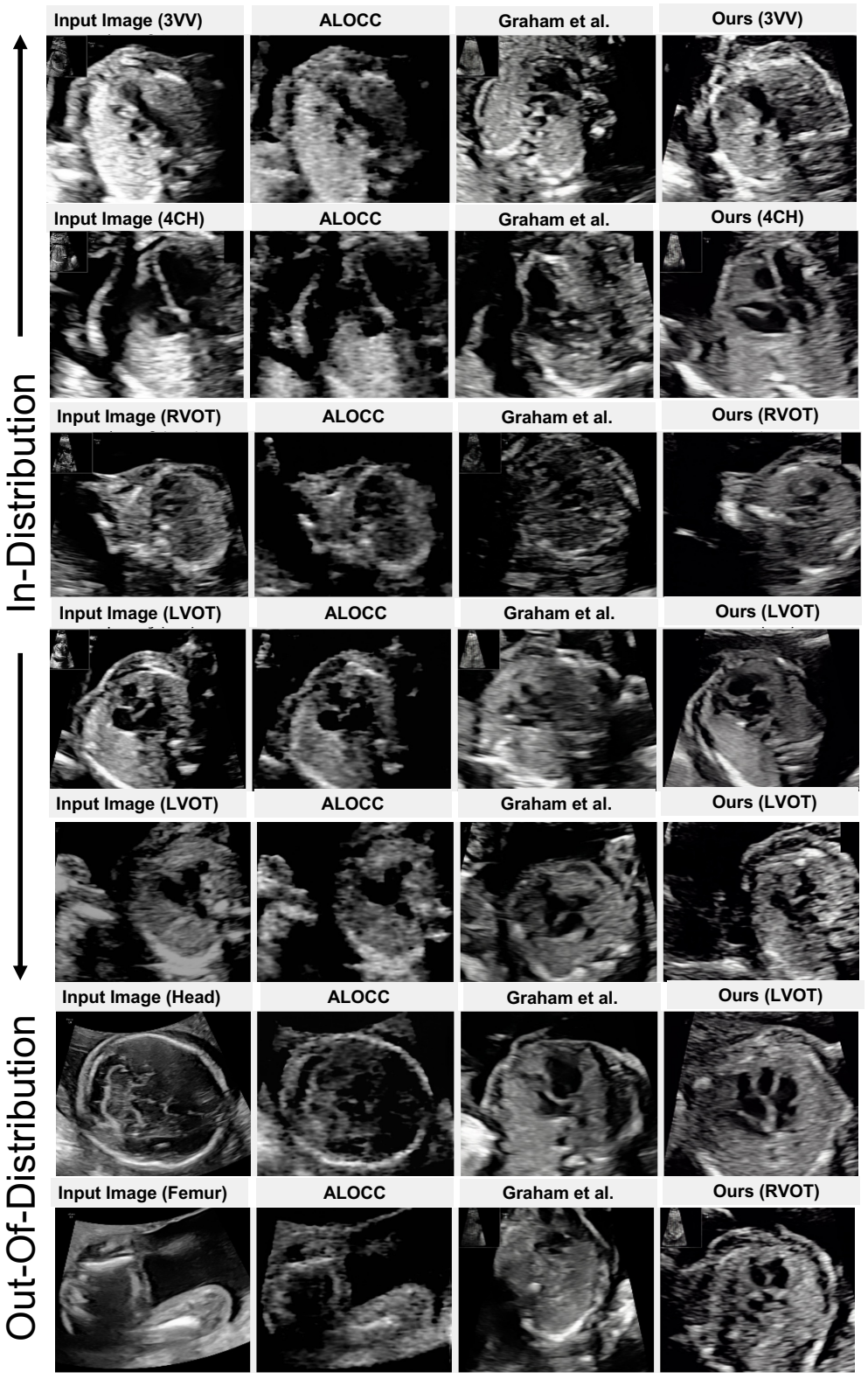


Figure 4.7: Qualitative results for in-distribution and out-of-distribution examples. For each row, from left to right: input image (true label), ALOCC, Graham et al., and our method (prediction). Our model generates images structurally and semantically similar to the input for ID samples (heart views) but dissimilar for OOD samples (head and femur).

5

STAN-LOC: Visual Query-based Video Clip Localization for Fetal Ultrasound Sweep Videos

Background: Building upon DCDM for out-of-distribution (OOD) detection in fetal ultrasound, we next address a key limitation of standard clinical practice and existing automated approaches for fetal cardiac assessment. Free-hand scanning during routine fetal ultrasound generally targets a single "best" frame for each anatomical view, emphasizing high-quality static images. However, this approach is not tailored to capture the nuanced spatiotemporal dynamics of the beating heart that are crucial for identifying subtle congenital abnormalities. While DCDM can automatically isolate heart frames and reduce the need for exhaustive manual review, free-hand ultrasound scans inherently provide only a limited glimpse of the continuous spatial and temporal relationships between different cardiac views.

To overcome this limitation, the Clinical Artificial Intelligence Fetal Echocardiography (CAIFE) project introduces heart sweeps, which are continuous ultrasound acquisitions designed to capture dynamic spatial and temporal features of the fetal heart in a single scan. Though these sweeps promise more comprehensive coverage, the task of manually identifying standard planes within lengthy video sequences becomes increasingly challenging due to the high structural similarity across views, annotation inconsistencies, and subtle transitions at anatomical boundaries.

To address these challenges, we introduce the novel task of visual query based video clip localization (VQ-VCL) for medical video analysis and propose STAN-LOC as the first solution. The VQ-VCL task is designed to operate on ultrasound sweep videos acquired as part of the CAIFE protocol, in which the sonographer performs a brief free-hand sweep to capture the fetal heart in a single continuous acquisition. Given such a sweep, the goal of VQ-VCL is to retrieve diagnostically relevant clips using a visual query. In this work, we assume that the visual query is selected from a predefined standardized atlas comprising representative standard view images collected from multiple patients, rather than being patient specific or selected in real time during scanning. This design reflects a realistic clinical deployment scenario in which sonographers can select the desired anatomical view from an atlas integrated into the imaging system, ensuring consistency and generalisability across patients and operators. While alternative sources of visual queries, such as patient specific prior scans or real time manual frame selection by the sonographer, are conceptually compatible with the proposed framework, they are not explored in this thesis.

Since the atlas is constructed primarily from scans exhibiting normal cardiac anatomy, retrieval performance may be affected when target videos contain structural abnormalities that deviate from typical anatomical appearances. Such variability can reduce visual similarity between the query and target frames, motivating the need for query robust feature representations and inference time query selection strategies, as incorporated in STAN-LOC. STAN LOC integrates visual query and video streams using a query aware spatio temporal transformer and employs a multi anchor, view aware contrastive loss to address label noise and object similarity, along with a robust query selection strategy. On fetal heart ultrasound sweeps from the PULSE dataset, STAN LOC surpasses existing methods, delivering a 22% improvement in mean temporal Intersection over Union (mtIoU), underscoring its promise for automated and standardized retrieval of diagnostically relevant clips in clinical workflows.

Authors: Divyanshu Mishra, Pramit Saha, He Zhao, Olga Patey, Aris T. Papageorghiou, J. Alison Noble

Published in Conference: Mishra, D., Saha, P., Zhao, H., Patey, O., Papageorghiou, A.T. and Noble, J.A., 2024, October. STAN-LOC: Visual Query-Based Video Clip Localization for Fetal Ultrasound Sweep Videos. In International

Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 742-752). Cham: Springer Nature Switzerland.(MICCAI 2024)

Author Contribution: I was the lead technical author of the paper, responsible for formulating the problem statement, proposing the solution, designing the codebase, conducting the experiments, and preparing the original manuscript draft. He Zhao contributed to technical discussions, while Pramit Saha contributed to the overall discussion. J. Alison Noble provided both technical guidance and overall supervision. All authors reviewed and approved the final version of the manuscript.

Abstract

Detecting standard frame clips in fetal ultrasound videos is crucial for accurate clinical assessment and diagnosis. It enables healthcare professionals to evaluate fetal development, identify abnormalities, and monitor overall health with clarity and standardization. To augment sonographer workflow and to detect standard frame clips, we introduce the task of Visual Query-based Video Clip Localization in medical video understanding. It aims to retrieve a video clip from a given ultrasound sweep that contains frames similar to a given exemplar frame of the required standard anatomical view. To solve the task, we propose STAN-LOC that consists of three main components: (a) a Query-Aware Spatio-Temporal Fusion Transformer that fuses information available in the visual query with the input video. This results in visual query-aware video features which we model temporally to understand spatio-temporal relationship between them. (b) a Multi-Anchor, View-Aware Contrastive loss to reduce the influence of inherent noise in manual annotations especially at event boundaries and in videos featuring highly similar objects. (c) a query selection algorithm during inference that selects the best visual query for a given video to reduce model's sensitivity to the quality of visual queries. We apply STAN-LOC to the task of detecting standard-frame clips in fetal ultrasound heart sweeps given four-chamber view queries. Additionally, we assess the performance of our best model on PULSE [2] data for retrieving standard transventricular plane (TVP) in fetal head videos. STAN-LOC surpasses the state-of-the-art method by 22% in mtIoU.

5.1 Introduction

In a routine pregnancy ultrasound assessment of the fetus, the sonographer scans through different fetal anatomies to evaluate fetal development and identify potential anomalies. For each anatomy, the sonographer reviews each frame meticulously and selects standard frames which are frames that contain all the anatomical landmarks in the correct anatomical orientation, size and position as defined by clinical guidelines (such as ISUOG [3, 165]). This process is time-consuming. Integrating a video-clip localization model has the potential to augment the sonographer's workflow allowing the sonographer to focus on detailed analysis and anomaly detection. However, automatically detecting standard frames is challenging as the frames before/after the standard frames are highly similar, with often small misalignment of anatomical landmarks. Moreover, most of the views have high global structural similarity with only minor local variations, thereby making the detection of their temporal boundaries challenging as shown in Fig. 5.1. As the task is complex, even experts can find it difficult to agree on what they refer to as a standard or non-standard frame as shown in Suppl. Fig. 1 that depicts a study where two cardiologists were asked to annotate the same 10 fetal heart videos. The kappa score [166] between the two experts was only 66% in this case, verifying the complexity of the task. This results in noisy annotations further complicating the issue. Existing works utilizing visual queries mainly comprise image retrieval [167–170] and the recently defined task of visual query-based 2D localization (VQ2D) [78, 171, 172] in the Ego4D [77] dataset. However, both lines of work output a single image and typically utilize coarser-grained datasets. In scenarios like surgical procedure planning, disease screening/diagnosis, and procedure/process demonstration, users often need a video clip of the object rather than just a single image. Retrieving a video clip in our context is more challenging because along with the object, the ultrasound probe is in motion, leading to various deformations, occlusions, and motion blur. These factors deviate the object's appearance from the original query, making it harder for the model to accurately locate all its instances within the video clip.

In this paper, we introduce the Visual Query-based Video Clip Localization (VQ-VCL) task where, given an ultrasound scan, a sonographer provides an exemplar

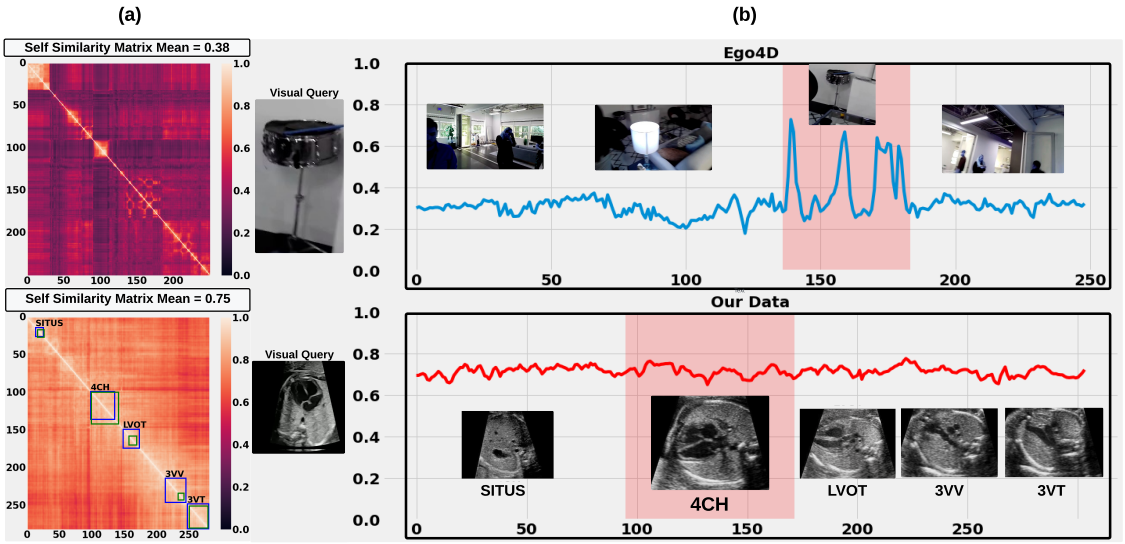


Figure 5.1: (a) Self-similarity matrix for a randomly chosen video from Ego4D (top, mean=0.38) [77] and our clinical video dataset (bottom, mean=0.75), which reveals higher task difficulty for our video clip localization task. The uncertainty in annotations of two expert cardiologists are shown in green and blue boxes, respectively. (b) Cosine similarity of the visual query with the video for both Ego4D (top) and our data (bottom). Compared to Ego4D, our clinical data obtains similar scores along the video emphasizing the challenge whereas Ego4D exhibits high scores only within the region of interest.

frame representing the anatomical view they wish to review. The model's objective is to retrieve a clip from the scan that contains the corresponding standard frames depicting the anatomy. We develop a query-aware spatio-temporal transformer model (STAN-LOC) that retrieves the clip-containing frames similar to the visual query from a given video.

Our contributions are: (a) We introduce the task of Visual Query-based Video Clip Localization (VQ-VCL) and propose a query-aware spatio-temporal transformer model, STAN-LOC, to automate this task. STAN-LOC includes a Query Aware Spatio-Temporal Fusion transformer to model the spatial and temporal relationship between the video and visual query. (b) To deal with noisy labels and challenging event boundaries, we include a Multi-Anchor, View-Aware Contrastive Loss and a Temporal Uncertainty Robust Localization Loss. (c) We propose a VQ selection module to guide the model during inference to select the best query candidate for a given input video. (d) We demonstrate STAN-LOC performance for two different real-world tasks of standard-frame detection with limited data availability and noisy labels.

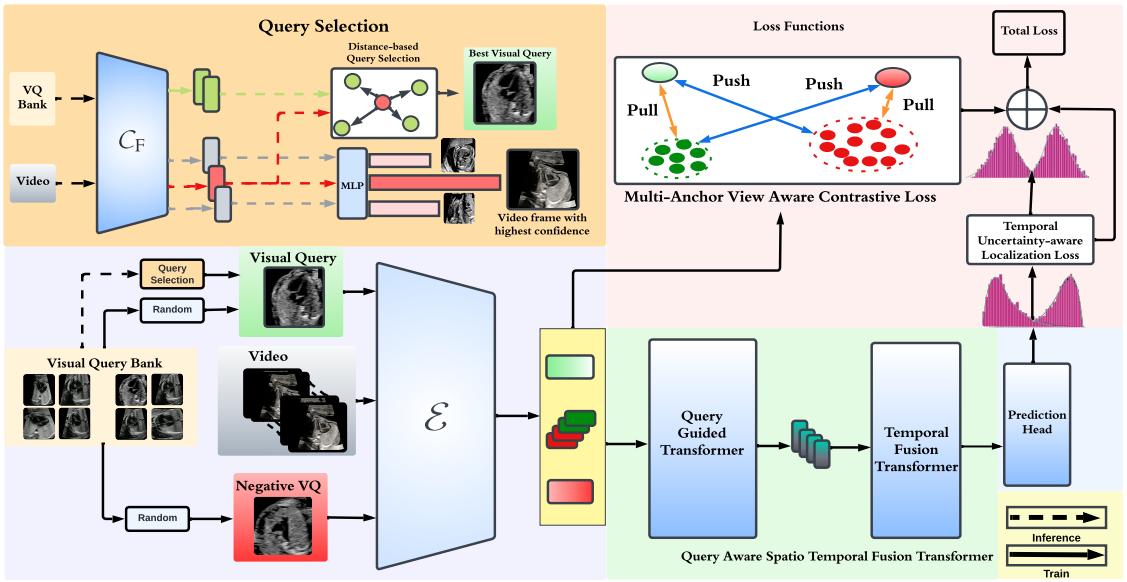


Figure 5.2: Main architecture for STAN-LOC where input video, visual query (VQ), and negative VQ are passed through the backbone to extract features during training. The features from the visual query and the video are passed to the query-aware spatio-temporal fusion transformer and the resultant fused features are fed to the prediction head to predict the distributions of start and end frames. During inference, we select the best VQ from the visual query bank using the Query Selection algorithm.

5.2 Methods

Video Clip Retrieval Task Description: Visual query-based video clip localization (VQ-VCL) is formulated as a temporal localization task. Given a video v and an exemplar query frame q from a separate exemplar database $\mathcal{Q}$, the model is trained to predict the start (t_s) and end (t_e) frames of a clip v_q where $v_q \subset v$ contains frames semantically similar to q.

STAN-LOC Overall Architecture: Our proposed architecture, as depicted in Fig. 5.2, takes a video v and a visual query q as inputs. These inputs are passed through a shared ResNet101 [173] encoder $\mathcal{E}$, resulting in video features $f_v \in \mathbb{R}^{\mathbf{TxHxWxC}}$ and visual query features $f_q \in \mathbb{R}^{\mathbf{HxWxC}}$. The extracted features are then fed to the Query-Aware Spatio-Temporal Fusion Transformer which first fuses the visual query with the video features and then models the resulting features temporally to yield spatio-temporal features. Finally, the spatio-temporal features are passed through an MLP responsible for predicting the distribution of start and end frames. During training, a Multi-Anchor View-Aware Contrastive Loss is

proposed to make the model more sensitive to the query frame, which is further elucidated in Section 5.2.2. At inference, we integrate a query selection algorithm detailed in Section 5.2.3 to choose the most suitable query for the input video, enhancing the model's overall performance.

5.2.1 Query-Aware Spatio-Temporal Fusion Transformer

Query-Guided Spatial Transformer: The design of the encoder to fuse the video and the visual query features is crucial. Previous works for visual grounding [80], and moment retrieval [174] naively concatenate the features from video and query together. This reduces the relevance of visual queries and results in features possessing less information about the visual query [175]. To ensure that the video features (f_v) are contextualised by the information contained within the visual query features (f_q), we designed a Query-Guided Spatial Transformer. Specifically. we introduce cross-attention [159] layers to fuse the video and visual query features. Formally, given the video features f_v and visual query features f_q, we project video features to obtain query Q_v whereas key K_q and value V_q are obtained from the visual query feature f_q. The attention operation [159] is performed between Q_v, K_q and V_q to obtain query-guided feature QV_f, which can be formulated as:

$$QV_f = FFN\left(softmax(\frac{Q_v K_q^T}{\sqrt{d_k}})V_q\right), \tag{5.1}$$

where FFN is a feed-forward network and d_k is the dimensionality of the query and key vectors.

Temporal Fusion Transformer: To incorporate temporal information in the query-aware video features QV_f and fuse the spatio-temporal features, we designed a temporal fusion transformer. Formally, given QV_f, we first add fixed sinusoidal positional encoding to enrich the features with positional information. Then we perform self-attention [159] by projecting the resulting video features to Q_{v_q}, K_{v_q}, and V_{v_q} as shown in Eq. 5.2. This helps in modelling the temporal interactions within the visual query-aware video features and generates spatio-temporal features F_T.

$$F_T = FFN\left(softmax(\frac{Q_{v_q} K_{v_q}^T}{\sqrt{d_k}})V_{v_q}\right) \tag{5.2}$$

5.2.2 Loss Functions

Multi-Anchor, View-Aware Contrastive Loss: In settings with high spatial similarity between the video frames as seen in Fig. 5.1, estimating the correct event boundary is an extremely challenging task. Moreover, as the objects of interest and the data acquisition device are both in motion, object appearance can strongly deviate from the visual query. To mitigate the above issues, we introduce a Multi-Anchor, View-Aware Contrastive Loss. The loss has two components: a) **Positive View-Aware Contrastive Loss ($\mathcal{L}_{PVAC}$)** which aims to bring the visual query features and the ground-truth clip features together while pushing away frames belonging to other classes; b) **Negative View-Aware Contrastive Loss ($\mathcal{L}_{NVAC}$)** which utilizes a negative query (frame belonging to other classes) and aims to push the feature representation of positive frames in the video away from negative frames. Formally, given video-features f_v, visual-query features f_q and negative visual-query features f_{q^-}, we extract the video features belonging to the ground truth clip and consider them as positive features ($f^+_{v_i}$) while the video features of the frames lying outside the ground-truth clip are considered as negative features ($f^-_{v_j}$).

For $\mathcal{L}_{PVAC}$, we consider the visual query features f_q as the anchor and calculate the cosine similarity of f_q with $f^+_{v_i}$ and f_q with $f^-_{v_j}$ where i, j iterate over K_1 positive and K_2 negative features as shown by Eq. 5.3, where $sim(\cdot)$ indicates the cosine similarity function. Finally, we optimize the loss to pull positive features $f^+_{v_i}$ closer to the visual query feature f_q while pushing all K_2 negative $f^-_{v_j}$ away as formulated in Eq. 5.3 where τ^+ is the positive temperature.

$$\mathcal{L}_{PVAC} = -\log \frac{\sum_{i=0}^{K_1} \exp\left(sim(f_q, f^+_{v_i})/\tau^+\right)}{\sum_{j=0}^{K_2} \exp\left(sim(f_q, f^-_{v_j})/\tau^+\right)} \tag{5.3}$$

On the other hand, for $\mathcal{L}_{NVAC}$ the negative visual query features f_{q^-} are considered as the anchor and we calculate the cosine similarity of f_{q^-} with $f^-_{v_i}$ and f_{q^-} with $f^+_{v_j}$ where i, j iterate over K_2 negative and K_1 positive features as shown in Eq. 5.4. Finally, we optimize the loss to pull the negative features $f^-_{v_i}$ closer to the negative visual query feature f_{q^-} while pushing all K_1 positive $f^+_{v_j}$ away as stated in Eq. 5.4 where τ^- is temperature for $\mathcal{L}_{NVAC}$.

$$\mathcal{L}_{NVAC} = -\log \frac{\sum_{i=0}^{K_2} \exp\left(sim(f_{q^-}, f_{v_i}^-)/\tau^-\right)}{\sum_{j=0}^{K_1} \exp\left(sim(f_{q^-}, f_{v_j}^+)/\tau^-\right)} \tag{5.4}$$

The final loss $\mathcal{L}_{MVAC}$ is denoted in Eq. 5.5 where w_p and w_n are tunable weights for $\mathcal{L}_{PVAC}$ and $\mathcal{L}_{NVAC}$ respectively.

$$\mathcal{L}_{MVAC} = w_p * \mathcal{L}_{PVAC} + w_n * \mathcal{L}_{NVAC} \tag{5.5}$$

Temporal Uncertainty Robust Localization Loss: The task of VQ-VCL becomes more challenging when there is a high similarity between the frames belonging to different classes and the event boundaries are unclear. This leads to noisy annotations available to train the model. To reduce sensitivity to noisy annotations, we introduce a Temporal Uncertainty Robust Localization Loss ($\mathcal{L}_{URL}$). Instead of using binary ground truth, we generate two Gaussian distributions $T_s(x)$ and $T_e(x)$ corresponding to the true start frame (t_s) and true end frame (t_e) of the target video clip, with means $\mu_s = t_s$ and $\mu_e = t_e$ and standard deviation ($\sigma = 1$) respectively as shown in Eq. 5.6 . Finally, we optimize the KL-divergence loss between the predicted ($P_s(x)$, $P_e(x)$) and true ($T_s(x)$, $T_e(x)$) start and end distribution and combine them together as shown in Eqs. 5.7 and 5.8 respectively.

$$T_s(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu_s)^2/2\sigma^2},\ T_e(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu_e)^2/2\sigma^2} \tag{5.6}$$

$$KL_s(P_s||T_s) = \sum_x P_s(x)\log(\frac{P_s(x)}{T_s(x)}),\ KL_e(P_e||T_e) = \sum_x P_e(x)\log(\frac{P_e(x)}{T_e(x)}) \tag{5.7}$$

$$\mathcal{L}_{URL} = KL_s + KL_e \tag{5.8}$$

Finally, we combine $\mathcal{L}_{MAC}$ and $\mathcal{L}_{URL}$ to give total loss ($\mathcal{L}$) which we use to train the model as expressed in Eq. 5.9.

$$\mathcal{L} = \mathcal{L}_{MAC} + \mathcal{L}_{URL} \tag{5.9}$$

5.2.3 Inference Query Selection

During inference, a user might provide queries which are low quality or extremely different from the VQ database $\mathcal{Q}$ used in training. To ensure that STAN-LOC is agnostic to the quality of visual queries, we introduce a classifier-based query selection module as shown in Fig. 5.2. The idea of the query selection is to provide a related query according to the input video, where the query frame is dynamically selected during inference. Given a video v, a visual query database $\mathcal{Q}_N$, where N is the number of visual queries, a reference frame F_v^{ref} from the video v is selected by a pre-trained classifier $\mathcal{C}_F$ with the highest confidence. Subsequently, we select M visual queries most similar to our reference frame F_v^{ref} by a distance function D (*e.g.,* Euclidean distance) between corresponding feature vectors. The M visual queries are averaged in the feature space to get the query feature for further retrieval process.

5.3 Experiments and Results

Dataset and Implementation: We evaluate STAN-LOC on two different fetal ultrasound video datasets. The first dataset gathered as part of the CAIFE [125] project, comprises fetal heart sweep videos for standard 4CH clip retrieval. The second dataset is sourced from the PULSE [5] dataset and contains fetal head videos for standard fetal head TV clip retrieval. The fetal heart video sweep dataset comprised 10-second transversal heart sweeps (TS) over the fetal heart. A TS sweep is obtained by scanning from the cardiac situs (Situs) to the four-chamber view (4CH), through the left ventricular outflow tract (LVOT), the three-vessel view (3VV), and finally, the three-vessel trachea view (3VT) of the fetal heart. We utilized 96 videos for training and 10 videos for testing the model. The visual queries for heart data include 609 4CH frames extracted from 11 held-out videos (In-Distribution (ID)) and standard 4CH heart frames extracted from the PULSE [5] data where sonographers freeze the video and only capture standard frames (Out-of-Distribution (OOD)). The fetal head dataset comprises fetal head frames in Transventricular (TV) and Transcerebellar views (TC). The visual queries for this dataset comprise standard TV frames from 8 videos and we utilize 159 videos to train and 23 to test. For all datasets, the visual query and video frames were

resized to 224×224 dimensions. We sampled clips with different start and end frames during training to augment the dataset. Further details are given in Dataset and Training Details section of the supplementary.

Results: STAN-LOC is compared with five different baseline models on two different datasets as shown in Table 5.1, where the models are ResNet3D[176], cosine-similarity supervised 2D CNN, TubeDETR [80],VQLOC [172] and MomentDETR [174], respectively. The chosen comparison metrics are Mean temporal intersection-over-union (mtIoU) and R @ t where R is recall, calculated at temporal IoU thresholds t. ResNet3D [176] exhibits the worst performance, with a mtIoU of 13.89 ± 3.67. Its R @ 0.7 is 0.02 and R @ 0.5 is 0.06 showing the model's inability to model longer-range interactions. TubeDETR [80], performs significantly better than ResNet3D with mtIoU of 27.85 ± 2.70 and R@ 0.5 = 0.22. However, R @ 0.7 of the model is 0.00, implying the model's failure to extract long-range features. Surprisingly, a simple cosine similarity supervised CNN baseline, outperforms the transformer-based TubeDETR with mtIoU of 29.43 ± 5.65 and R @ 0.5 of 0.28. This suggests that the model can learn the spatial correspondence between the video frame and the visual query but struggles with longer interactions (R@0.7=0), possibly due to the absence of temporal information in a 2D-CNN. MomentDETR has the best mtIoU (35.09 ± 3.27) and R@0.3 (0.58) across baselines, however, VQLOC surpasses it in R@0.7 (0.18) and R@0.5 (0.34), demonstrating superior performance in capturing longer interactions. STAN-LOC, with and without query selection, outperforms all baselines with mtIoU of 46.54, 55.04 and 57.42 respectively which is almost 22% more than MomentDETR. Its performance in modelling long-range dependencies is exceptional with R @ 0.7 = 0.50, R @ 0.5 = 0.60 and R @ 0.3 = 0.80 respectively.

Ablation Study:

We performed an ablation study to evaluate the importance of each of the key STAN-LOC components on overall model performance. Refer to Table 5.2. In loss functions ablation, the first row displays the model with only Focal loss [177]. We observe that utilising $\mathcal{L}_{URL}$ instead of Focal loss in STAN-LOC boosts the performance

Table 5.1: Quantitative Results. We test each baseline 5 times with different visual queries and report mean, and standard deviation. For STAN-LOC, we show the performance with and without Query selection (QS) where M is the number of best queries selected.

		Our Data				**PULSE Data** [5]			
Method		**mtIoU**	**R@0.7**	**R@0.5**	**R@0.3**	**mtIoU**	**R@0.7**	**R@0.5**	**R@0.3**
Resnet 3D [176]		13.89 ± 3.67	0.02 ± 0.04	0.06 ± 0.09	0.20 ± 0.10	43.45 ± 2.62	0.18 ± 0.04	0.43 ± 0.04	0.60 ± 0.04
TubeDETR[80]		27.85 ± 2.70	0.00 ± 0.00	0.22 ± 0.08	0.48 ± 0.08	55.91 ± 1.41	0.36 ± 0.04	0.57 ± 0.05	0.77 ± 0.02
Cosine Similarity Sup CNN		29.43 ± 2.38	0.00 ± 0.00	0.28 ± 0.08	0.50 ± 0.00	23.01 ± 0.15	0.17 ± 0.00	0.21 ± 0.02	0.26 ± 0.03
VQLOC [172]		30.87 ± 5.65	0.18 ± 0.04	0.34 ± 0.11	0.44 ± 0.11	42.83 ± 2.57	0.14 ± 0.02	0.34 ± 0.06	0.62 ± 3.64
MomentDETR [174]		35.09 ± 3.27	0.04 ± 0.05	0.32 ± 0.08	0.58 ± 0.11	57.20 ± 0.92	0.26 ± 0.06	0.64 ± 0.06	0.83 ± 0.02
STAN-LOC	W/O QS	46.54 ± 5.53	0.38 ± 0.08	0.50 ± 0.07	0.60 ± 0.10	58.35 ± 2.96	0.51 ± 0.06	0.59 ± 0.09	0.77 ± 0.06
	QS (M=1)	55.04 ± 0.00	0.50 ± 0.00	0.60 ± 0.00	0.70 ± 0.00	$\mathbf{58.67 \pm 0.00}$	0.57 ± 0.00	0.61 ± 0.00	0.83 ± 0.00
	QS (M=5)	$\mathbf{57.42 \pm 0.00}$	$\mathbf{0.50 \pm 0.00}$	$\mathbf{0.60 \pm 0.00}$	$\mathbf{0.80 \pm 0.00}$	58.36 ± 0.00	$\mathbf{0.57 \pm 0.00}$	$\mathbf{0.61 \pm 0.00}$	$\mathbf{0.83 \pm 0.00}$

by 14.22 % mtIoU indicating the importance of soft ground truth for noisy labels. Incorporating $\mathcal{L}_{PVAC}$ to STAN-LOC further improves the performance for mtIoU (+ 9.16%) and recall, demonstrating the importance of positive anchors and their role in pushing positive samples away from negative ones. Further, adding $\mathcal{L}_{NVAC}$ to STAN-LOC boosts the mtIoU to 57.42 showing the importance of a negative anchor and its role in pulling negative samples closer in the feature space and away from positive samples. In Query Selection ablation, we observed variability in performance when selecting random queries during inference with standard deviation (S.D) of 5.53% and 2.90% in mtIoU for ID and OOD VQ databases. We show our Query selection algorithm improves performance significantly. We also ablate different distance functions for query selection and the number of queries selected during inference. We find KL divergence to work well across datasets and visual queries for M=5 to work best. In the Architecture ablation, we observe that both query-guided and temporal fusion transformers are essential for best performance.

5.4 Conclusion

This paper introduces a novel task of Visual Query-based Video-Clip Localization and proposes a video-based transformer model STAN-LOC. STAN-LOC has two architectural components: Query-Guided and temporal-fusion transformers to fuse the query features with the video and further model interactions between these features in the temporal dimension respectively. To deal with noise at temporal class boundaries, a Multi-Anchor View-Aware contrastive loss and Temporal Uncertainty Robust Localization loss are introduced. Finally, to reduce model sensitivity to the

Table 5.2: Ablation study showing effect of loss functions, query selection and architecture components on our model's performance where M is the number of VQ selected.

Loss Functions Ablation							
$\mathcal{L}_{\text{URL}}$	$\mathcal{L}_{\text{PVAC}}$	$\mathcal{L}_{\text{NVAC}}$	**mtIoU**	**R@0.7**	**R@0.5**	**R@ 0.3**	
✗	✗	✗	31.77	0.10	0.20	0.40	
✓	✗	✗	45.97	0.30	0.50	0.70	
✓	✓	✗	55.13	0.40	0.60	0.80	
✓	✓	✓	**57.42**	**0.50**	**0.60**	**0.80**	
Query Selection Ablation							
VQ Database	**QS**	**Distance Function**	**M**	**mtIoU**	**R@0.7**	**R@0.5**	**R@ 0.3**
In Distribution	✗	N/A	Random 5	46.54 ± 5.53	0.38 ± 0.08	0.50 ± 0.07	0.60 ± 0.10
	✓	Euclidean	1	48.27 ± 0.00	0.40 ± 0.00	0.50 ± 0.00	0.70 ± 0.00
	✓	Cosine Similarity	1	51.86 ± 0.00	0.40 ± 0.00	0.50 ± 0.00	0.80 ± 0.00
	✓	KL Divergence	1	55.04 ± 0.00	0.50 ± 0.00	0.60 ± 0.00	0.70 ± 0.00
	✓	KL Divergence	5	$\mathbf{57.42 \pm 0.00}$	$\mathbf{0.50 \pm 0.00}$	$\mathbf{0.60 \pm 0.00}$	$\mathbf{0.80 \pm 0.00}$
Out Of Distribution	✗	N/A	Random 5	42.96 ± 2.90	0.34 ± 0.05	0.50 ± 0.0	0.54 ± 0.05
	✓	KL Divergence	1	52.23 ± 0.00	0.40 ± 0.00	0.50 ± 0.00	$\mathbf{0.80 \pm 0.00}$
	✓	KL Divergence	5	$\mathbf{55.07 \pm 0.00}$	$\mathbf{0.50 \pm 0.00}$	$\mathbf{0.60 \pm 0.00}$	0.70 ± 0.00
Architecture Ablation							
Query-guided Fusion	**Spatio-Temporal**	**mtIoU**	**R @ 0.7**	**R @ 0.5**	**R @ 0.3**		
✗	✓	42.36	0.20	0.40	0.60		
✓	✗	37.53	0.20	0.40	0.50		
✓	✓	**57.42**	**0.50**	**0.60**	**0.80**		

quality of visual queries during inference, a test-time query selection algorithm is introduced to select the best query for the input video. The model is evaluated for two ultrasound video cases, where the video frames are highly similar and a low amount of training data is available. The effectiveness of the approach is demonstrated by comparing it with SOTA baselines and ablating different components.

Supplementary Material

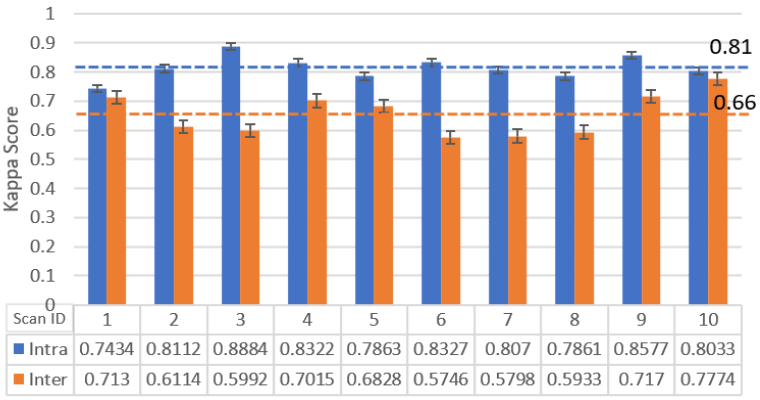
Dataset and Training Details

Table 5.3: Training details.

Component	Value
Framework	PyTorch
PyTorch version	1.8
Optimizer	AdamW
Epochs	200
Number of frames	150
Learning rate	1×10^{-5}
LR scheduler	Step
LR scheduler step size	75
τ^+	0.7
τ^-	0.2
w_1	1
w_2	0.4
Query-guided Transformer layers	1
Spatio-temporal Transformer layers	6
Classifier ($\mathcal{C}_F$)	ConvNeXt-small
Visual encoder ($\mathcal{E}$)	ResNet101
GPU	Tesla V100 32 GB

Table 5.4: Effect of selecting top-K queries and averaging them for ID and OOD VQ banks.

	VQ Data = In-Distribution				**VQ Data = Out-of-Distribution**			
Num Queries	**mtIoU**	R@0.7	R@0.5	R@0.3	**mtIoU**	R@0.7	R@0.5	R@0.3
1	55.04	0.5	0.6	0.7	52.23	0.4	0.5	0.8
5	57.42	0.5	0.6	0.8	**55.75**	**0.5**	**0.6**	0.7
7	**57.42**	**0.5**	**0.6**	**0.8**	51.08	0.4	0.5	**0.8**
9	52.55	0.4	0.5	0.8	51.08	0.4	0.5	0.8



Scan ID	1	2	3	4	5	6	7	8	9	10
Intra	0.7434	0.8112	0.8884	0.8322	0.7863	0.8327	0.807	0.7861	0.8577	0.8033
Inter	0.713	0.6114	0.5992	0.7015	0.6828	0.5746	0.5798	0.5933	0.717	0.7774

Figure 5.3: Inter- and intra-annotation agreement (kappa) for standard frame detection in the Transversal heart sweep (TS). The Cohen's kappa between annotators is approximately 0.66, indicating moderate agreement and the difficulty of the task.

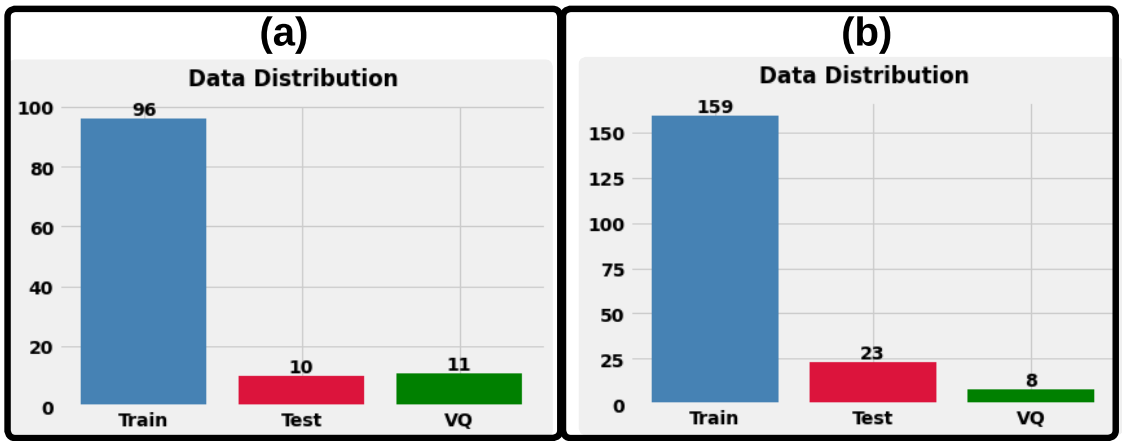


Figure 5.4: (a) Heart sweep data distribution for standard 4CH clip retrieval and (b) PULSE data distribution for TV-frame retrieval in fetal head video clips.

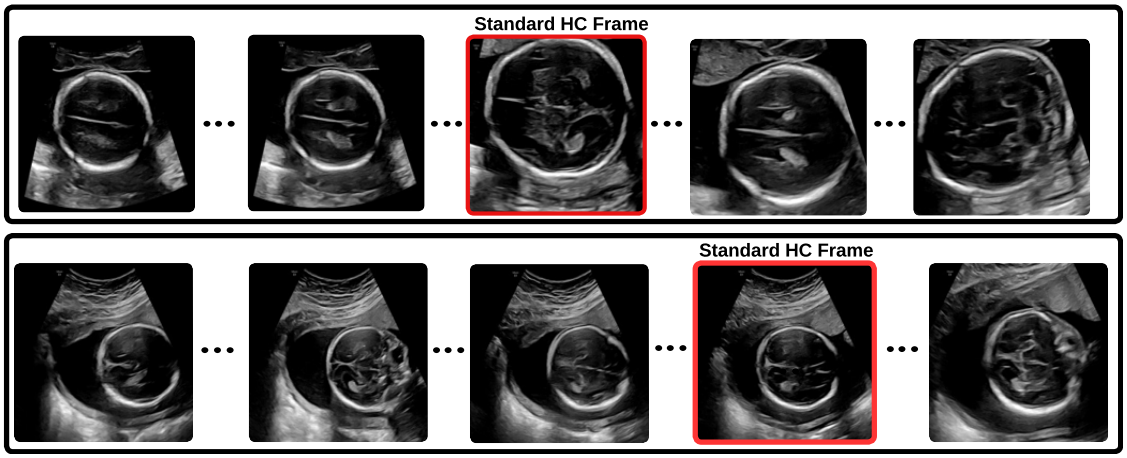


Figure 5.5: Example fetal head ultrasound frames illustrating standard and non-standard TV frames used for the head TV-frame retrieval task.

6

TIER-LOC: Visual Query-based Video Clip Localization in Fetal Ultrasound Videos with a Multi-Tier Transformer

Authors: Divyanshu Mishra, Pramit Saha, He Zhao, Netzahualcoyotl Hernandez-Cruz ,Olga Patey, Aris T. Papageorghiou, J. Alison Noble

Published in Journal: Medical Image Analysis

Background: Accurate localization of standard-view clips in fetal ultrasound videos is fundamental for reliable clinical assessment and diagnosis. However, previous visual query-based models—such as STAN-LOC—utilize single-scale feature extraction, making them less effective at capturing the subtle and localized anatomical differences that frequently arise in multi-view and fine-grained clinical scenarios. Single-scale approaches often miss critical local cues or fail to generalize across views with varying resolutions, thereby limiting their precision and robustness. In this chapter, we address these limitations by introducing TIER-LOC, a multi-tier transformer framework that learns and integrates discriminative features across multiple spatial scales. This approach enables the model to capture both coarse and fine-grained information, resulting in enhanced retrieval of clinically relevant video segments. Evaluations on two fetal ultrasound datasets and a fine-grained egocentric video benchmark demonstrate that TIER-LOC outperforms state-of-the-art models, with improvements of 7%, 4%, and 4% in mean temporal Intersection

over Union (mtIoU), respectively, highlighting its effectiveness for automated video localization in medical imaging.

Author Contribution: I was the lead technical author of the paper, responsible for formulating the problem statement, proposing the solution, designing the codebase, conducting the experiments, and preparing the original manuscript draft. Pramit Saha and He Zhao contributed to technical discussions and participated in reviewing the paper. Netzahualcoyotl Hernandez-Cruz did the data management for the project and Olga Patey helped in data collection. Aris T. Papageorghiou provided clinical supervision throughout the project. J. Alison Noble conceived the overall objectives of the study, secured funding, and supervised the entire project. All authors reviewed and approved the final version of the manuscript.

Abstract

In this paper, we introduce the Visual Query-based task of Video Clip Localization (VQ-VCL) for medical video understanding. Specifically, we aim to retrieve a video clip containing frames similar to a given exemplar frame from a given input video. To solve the task, we propose a novel visual query-based video clip localization model called TIER-LOC. TIER-LOC is designed to improve video clip retrieval, especially in fine-grained videos by extracting features from different levels, *i.e.*, coarse to fine-grained, referred to as TIERS. The aim is to utilize multi-Tier features for detecting subtle differences, and adapting to scale or resolution variations, leading to improved video-clip retrieval. TIER-LOC has three main components: 1) a Multi-Tier Spatio-Temporal Transformer to fuse spatio-temporal features extracted from multiple Tiers of video frames with features from multiple Tiers of the visual query enabling better video understanding. 2) a Multi-Tier, Dual Anchor Contrastive Loss to deal with real-world annotation noise which can be notable at event boundaries and in videos featuring highly similar objects. 3) a Temporal Uncertainty-Aware Localization Loss designed to reduce the model sensitivity to imprecise event boundary. This is achieved by relaxing hard boundary constraints thus allowing the model to learn underlying class patterns and not be influenced by individual noisy samples. To demonstrate the efficacy of TIER-LOC, we evaluate it on two ultrasound video datasets and an open-source egocentric video dataset. First, we develop a sonographer workflow assistive task model to detect standard-frame clips in fetal ultrasound heart sweeps. Second, we assess our model's performance in retrieving standard-frame clips for detecting fetal anomalies in routine ultrasound scans, using the large-scale PULSE dataset. Lastly, we test our model's performance on an open-source computer vision video dataset by creating a VQ-VCL fine-grained video dataset based on the Ego4D dataset. Our model outperforms the best-performing state-of-the-art model by 7%, 4%, and 4% on the three video datasets, respectively.

6.1 Introduction

Text query-based localization tasks, including video-temporal grounding [80, 178, 179], video moment retrieval [175, 180], and highlight detection have recently shown promising performance in the domain of natural video-understanding. In the medical domain, reports heavily rely on static images and text to convey diagnostic information, making image-text queries an apt solution. However, the advent of video technology introduces a paradigm shift in diagnostic capabilities. While static images are informative, capturing diagnostic moments in a video format provides a more comprehensive and nuanced understanding. Take, for example, a dynamic ultrasound video of a beating heart versus a single static frame – the video clip not only offers a more detailed depiction but also enables a holistic assessment of cardiac function. Moreover, in scenarios like fetal anatomy examinations, recording video clips around standard anatomy planes surpasses the limitations of singular frames. This approach allows practitioners to not only measure biometry accurately but also review the entire video sequence to ensure optimal plane selection. In instances where anomalies or concerns arise, the ability to scrutinize multiple video clips enhances diagnostic precision, emphasizing the pivotal role of dynamic visual data in the medical field. However, for the video clips, paired video-textual data is typically scarce. When available, it often comes in the form of frame-wise sparse class labels or radiology reports, where typically clinical experts provide a diagnosis for the entire video rather than offering detailed information at the clip level. Image-based queries, i.e., visual queries (VQs) offer an intuitive and direct approach to identifying objects or finding similar images. VQs not only reduce natural language barriers but also excel in expressing complex concepts or scenes that may be challenging to articulate via text. For instance, describing a medical anomaly may prove challenging with a text query, whereas presenting a model with an example frame containing the anomaly offers a more effective approach.

Literature on visual queries includes: (a) image retrieval [167–170] and (b) visual query-based 2D localization (VQ2D) [78, 79, 171, 172]. However, both lines of work output a single image or utilize datasets with more distinct classes as shown in Fig. 6.1. In the case of ultrasound video, retrieving a video clip rather than a

single frame is more challenging as observed in Fig. 6.1 because the ultrasound probe is in motion, and the (scanned) object (heart, fetus) may be in motion as well, causing various deformations, occlusions, motion blur etc. which deviate the frame appearance from the visual query, making it harder for a VQ model to localize all instances of the object.

Fetal ultrasound is crucial for monitoring prenatal development, detecting potential abnormalities, and ensuring the overall health and well-being of both the fetus and the expectant mother. In a routine pregnancy ultrasound assessment of the fetus, the sonographer scans through different fetal anatomies to evaluate fetal development and identify potential anomalies. For each anatomy, the sonographer reviews each frame meticulously and selects standard frames which are frames that contain all the anatomical landmarks in the correct anatomical orientation, size and position as defined by clinical guidelines (such as ISUOG [3, 165]). However, this process is time-consuming.

Integrating a video-clip localization model has the potential to augment the sonographer's workflow, allowing the sonographer to focus on detailed video review, analysis and anomaly detection. However, automatically detecting standard frames is challenging because of the following reasons: Firstly, the frames before and after the standard frames are often highly similar, with only small differences in anatomical landmark appearance. Most of the views have high global structural similarity with only minor local variations, thereby making the detection of anatomical temporal boundaries challenging as shown in Fig.6.1. Secondly, the localisation task is complex. Indeed, human experts can find it difficult to agree on selection of a standard or non-standard frame. This is shown in Fig. 6.2 that depicts a study where two cardiologists were asked to annotate the same 10 fetal heart videos. The kappa score [166] between the two experts was only 66% in this case, verifying the complexity of the task. This results in (naturally) noisy annotations.

Machine learning has been applied to automate a number of fetal ultrasound image analysis tasks, including automated biometry, standard-plane detection and image and video quality assessment. Automatic fetal ultrasound standard-plane detection has been well-studied for over a decade and some commercial systems have automated ultrasound standard-plane detection and quality assurance functionality.

Previously published works are predominantly image-based classification methods that classify an acquired clinical standard-plane frame using supervised [27, 181–185] or self-supervised [186] machine learning-based approaches. An interesting recent work developed an automatic method to predict the appearance of a standard plane frame in a free-hand ultrasound video clip approaching the trans-ventricular (TV) plane as a sonographer guidance aid [187]. That approach has not yet been generalized to multiple standard planes. A video-based ultrasound analysis method to detect clinical high-quality video clips containing standard planes is proposed in [188, 189]. However, that method only identifies if a clip includes standard planes as a video quality check, without identifying the specific location of standard plane frames. The current paper considers finding video clips of fetal anatomy in ultrasound sweeps, with a specific use case of standard plane partitioning of fetal echocardiography sweeps. The video characteristics that make this problem challenging relative to the work above, are the similar frame-to-frame (global) appearance, especially between standard and non-standard frames of the same anatomy and dynamic movement of the object. Further, existing standard plane detection approaches are frame-based, selecting a single frame for analysis. However, for the thorough evaluation of anomalies and assessment of dynamic anatomies such as the heart, video clips are more suitable. The heart's continuous movement and rhythmic contractions necessitate the review of a video clip to ensure accurate and comprehensive clinical assessment of all anatomical landmarks, capturing the full range of motion and functional dynamics essential for diagnosis. We have therefore chosen to automate the partitioning task using a visual query-based video clip localization task described below.

Our contributions are as follows:

1. We introduce the Visual-Query-based Video Clip Localization (VQ-VCL) task. This task requires a model to return a video clip containing a specific object when given a video and a visual query depicting that object. In the context of standard frame detection in fetal ultrasound videos, first a sonographer performs a quick sweep over the anatomies without stopping to capture specific standard planes. Then the sonographer inputs the captured video sweep along

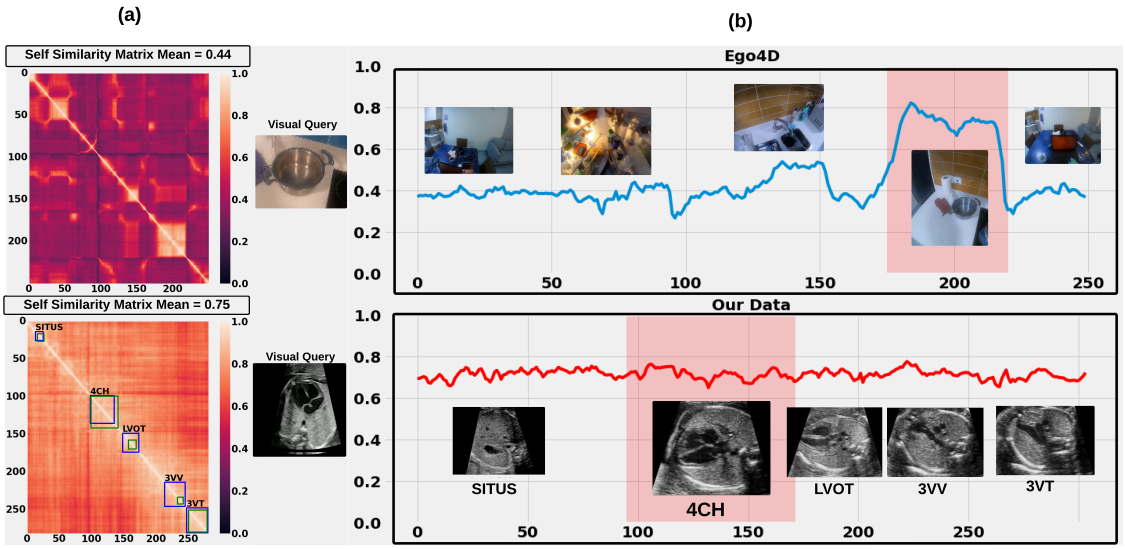
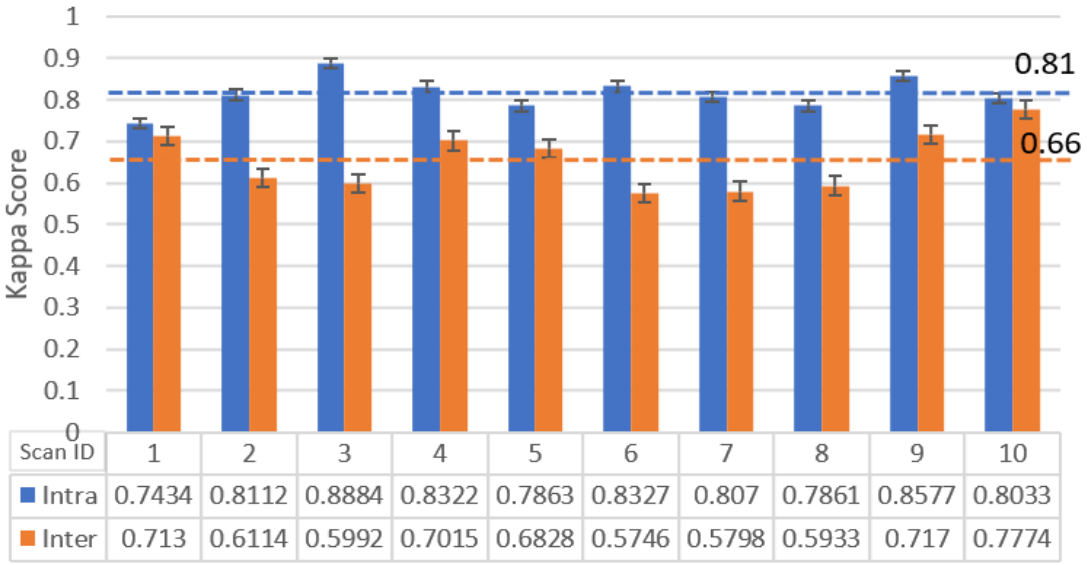


Figure 6.1: (a) Self-similarity matrix for a randomly chosen video from Ego4D (**top**, mean=0.44) [77] and our clinical video dataset (**bottom**, mean=0.75), which reveals higher task difficulty for our video clip localization task. The uncertainty in annotations of two expert cardiologists is shown in green and blue boxes, respectively. (b) Cosine similarity of the visual query with the video for both Ego4D (top) and our data (bottom). Our clinical data obtains similar scores along the video emphasizing the challenge, whereas Ego4D exhibits high scores only within the region of interest.

with a visual query of the specific anatomy standard plane that they need to analyze further into the VQ-VCL model. The model localizes a video clip from within the sweep that contains the standard frames, referred to as a standard frame clip, which the sonographer can then analyze for anomalies.

To solve this task, we propose TIER-LOC, a spatio-temporal video Transformer model designed to perform effectively in scenarios where the classes in the video are highly similar. Our approach includes a multi-tier feature extraction module that learns the spatio-temporal features in a coarse-to-fine-grained manner. The spatial information is acquired by a query-aware Transformer, and the temporal information is integrated by a learnable embedding to obtain the final spatio-temporal features.

2. To mitigate the problem of complex event boundaries and noisy labels, we propose a combined loss of Multi-Tier, Dual Anchor Contrastive Loss and Temporal Uncertainty-Aware Localization Loss.



Scan ID	1	2	3	4	5	6	7	8	9	10
Intra	0.7434	0.8112	0.8884	0.8322	0.7863	0.8327	0.807	0.7861	0.8577	0.8033
Inter	0.713	0.6114	0.5992	0.7015	0.6828	0.5746	0.5798	0.5933	0.717	0.7774

Figure 6.2: Figure showing the inter and intra-annotation agreement of annotators for the task of standard frame detection in Transversal heart sweep(TS). We can see that the Kappa score between annotators is only 66% highlighting the difficulty of the problem.

3. We demonstrate our model performance by applying it to two different real-world tasks of ultrasound standard-frame detection with limited data availability and noisy labels. We also create an open-source VQ-VCL computer vision dataset based on the Ego4D [77] dataset and evaluate our model's performance on it to allow benchmarking with respect to our approach by others in the future.

6.2 Related Work

6.2.1 Fetal Ultrasound Standard Plane Detection:

The acquisition of standard planes in 2D fetal ultrasound videos is crucial for ensuring consistent and accurate assessment of fetal growth, early detection of anomalies, and adherence to clinical guidelines. However, manually selecting these standard planes is time-consuming, prompting numerous studies to focus on automating this process [27, 181, 190–193]. These approaches follow the idea of image classification by developing a deep learning model to classify the captured standard planes into various anatomical class (e.g head, femur, abdomen etc). On the other hand, [188, 189] leverage the temporal information of ultrasound videos and develop a

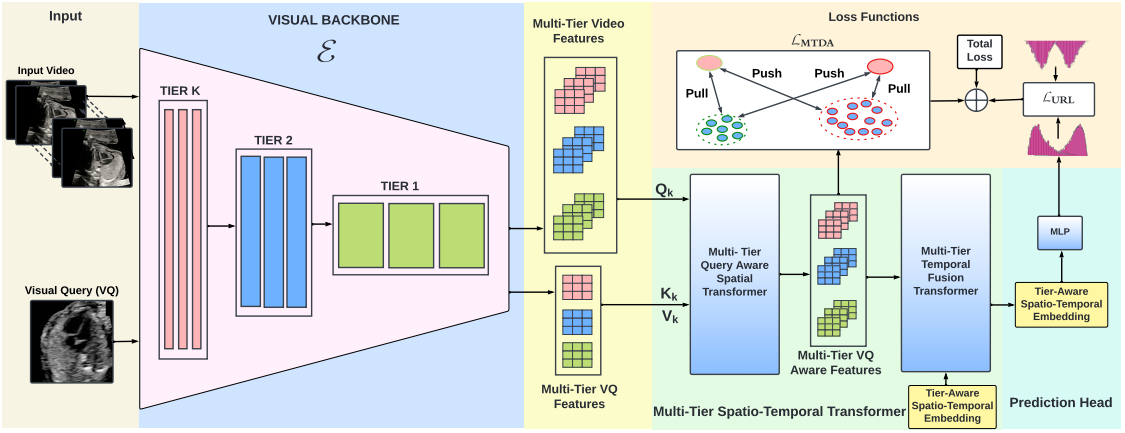


Figure 6.3: Main architecture of TIER-LOC. The input video v and visual query q are passed to the visual backbone to give multi-Tier features. These features are fused spatially using the Multi-Tier Query Aware Spatial Transformer. The Tier-specific features are passed to a) $\mathcal{L}_{MTDA}$ to learn the separation between classes, b) the Multi-Tier Temporal Fusion transformer to learn Tier-Aware Spatio-Temporal Embedding which is further passed to an MLP to make final prediction and calculate $\mathcal{L}_{URL}$ loss.

video-based anomaly detection method. This method identifies clinical high-quality video clips containing standard planes by using reconstruction error as the indicator. [187] trains a spatio-temporal generative model to synthesize a standard plane using the knowledge from the video clip. The generated image can then guide the sonographer in acquiring a good standard plane or guiding them to select a single frame similar to the generated frame. In contrast, our approach involves the sonographer capturing a sweep of the ultrasound video that includes all the relevant anatomies being analyzed, producing a standard-frame clip rather than a single frame. Using a visual query of the required anatomy, our method can automatically select the appropriate standard-frame clips from the video sweep. This reduces the manual effort required from the sonographer, enhancing efficiency and potentially enabling them to scan more patients and dedicate more time to analyzing the standard video clips.

6.2.2 Clinical Workflow Analysis:

In the context of automated fetal ultrasound clinical workflow analysis, first introduced by [194] for second-trimester scan analysis, the focus is on segmenting free-hand ultrasound videos into various anatomical regions, aiding in the analysis of the sonographer's clinical workflow. [194] introduced a (2D + T) CNN architecture,

termed Sono-2Dt-CNN, which was used for automated video classification. Building upon this approach, [195] extended the application with a dual-branch architecture called 'PULSE-v, aiming to transfer the learned knowledge to enhance the annotation of fetal anatomy in first-trimester scans.

The aforementioned work was developed for full scan workflow modeling. Furthermore, these models do not aim to locate standard frames but rather classify a given video clip to one of the anatomical classes (head, abdomen, heart etc). Our task involves detecting standard frame clips from a continuous ultrasound sweep based on a visual query that depicts the anatomy of interest. This is challenging because the smooth transitions between frames in the sweep result in high spatial similarity, making it difficult to differentiate standard frames from non-standard ones. Additionally, the process is designed to require minimal intervention from a sonographer. The sonographer only needs to perform a quick sweep over the anatomy and provide a visual query of the area they want to analyze. Our model will then identify the standard-frame video clip corresponding to the requested anatomy for further analysis.

6.2.3 Visual Query 2D Localization (VQ2D)

Introduced in Ego4D's episodic memory benchmark [77], VQ2D requires identifying the last occurrence of a queried object in an egocentric video and localizing it with a bounding box. Existing methods are broadly multi-stage or single-stage. Multi-stage approaches like Siam-RCNN [77] and its improved variant [78] treat spatial and temporal reasoning separately. In contrast, single-stage approaches—e.g., MINOTAUR [79] (multi-task) and VQLOC [172] (transformer-based)—perform spatiotemporal reasoning jointly. However, they focus on retrieving a single frame from the same video using an object crop. Our setting instead retrieves a video clip using a full-frame query drawn from a separate database, not the same video.

6.2.4 Video Temporal Grounding

Video Temporal Grounding (VTG) introduced in [196, 197] aims to localize a target moment by predicting start and end frame numbers according to a given natural language query. Earlier methods focused on generating moment proposals using

sliding windows [196, 198], proposal generation networks [199, 200] or utilized predefined anchors [201, 202]. The generated proposals were compared with the sentence queries to give the final localization. However, these methods were computationally expensive, and hence efficient proposal-free methods [80, 203–207] were introduced. Proposal-free methods encode the video and sentence features once and then model the interactions between them, resulting in efficient performance compared to proposal-based methods. VTG is similar to our task with the distinction being our task utilizes a single image as a visual query rather than a text query.

6.2.5 Multi-Scale Learning

The term Multi-Scale Learning has been used to refer to two different concepts in the computer vision literature. In the first line of work prominent in the video action recognition/classification literature, the term refers to multi-scale learning in the temporal domain. In this, the spatial features are extracted from a single spatial scale (TIER=1), the last layer of the feature extractor. The resulting feature vector is passed to their multi-scale encoder that transforms the single spatial scale feature into multi-scale temporal features by reducing the temporal dimension T and increasing channels D. This is different from our work where we employ multi-scale learning in the spatial domain and extract spatial feature vectors from multiple layers(multi-tier) of the feature extractor.

The second line of work [208–210] refers to multi-scale learning in the spatial domain. These works handle a single modality (image or video), have either a multi-scale encoder or decoder, and employ sequential fusion in the encoder/decoder, where multi-scale features are projected to common feature space and fused sequentially in a single representation from coarse to fine. In contrast, our work addresses a multi-modal scenario (image + video) with a multi-scale encoder and decoder. Further, we have an image as a query rather than text. To exploit the implicit bias of image/video modality and capture minute variations, we perform parallel fusion in our encoder. In parallel fusion, features from video and visual query at each tier are fused separately in original resolution across tiers.

6.2.6 Metric Learning

Existing metric learning methods [211–214] are built around a single positive anchor to bring similar samples closer and push dissimilar samples apart. While effective for many tasks, this single-anchor paradigm can struggle in fine-grained settings—such as fetal ultrasound—where classes share high global similarity but differ in subtle, localized landmarks. Relying solely on one anchor fails to capture these nuanced inter-class distinctions. In our work, we go beyond single-anchor approaches by introducing a dual-anchor loss, which incorporates both a positive anchor and a semantically similar negative anchor. This design aims to address the challenges of fine-grained class separation—particularly relevant for event boundary estimation in videos with high inter-frame similarity—without relying on a single source of contrast.

6.3 Methods

6.3.1 Video Clip Localization Task Formulation

The visual query-based video clip localization (VQ-VCL) task is formulated as a temporal localization task. Formally, given a video v and an exemplar frame q from a separate exemplar database $\mathcal{Q}$, the model is trained to predict the start (t_s) and end (t_e) frame number of a clip v_q where $v_q \subset v$ and contains frames semantically similar to q.

6.3.2 TIER-LOC Overall Architecture

Our proposed model, as depicted in Fig. 6.2, takes a video v and a visual query q as inputs. These inputs are passed through a shared encoder $\mathcal{E}$, resulting in K Tier video features $f_{v_K} \in R^{T\times H_k\times W_k\times C_k}$ and visual query features $f_{q_k} \in R^{H_k\times W_k\times C_k}$ where k iterates over K Tier features and T,H_k, W_k and C_k denote number of frames, height, width and channel dimensions at Tier k. The extracted features at each Tier are then spatially fused using a Multi-Tier Query-Guided Spatial Transformer to give Tier-specific query-aware features as shown in Fig. 6.2. The Tier-specific query-aware features are subsequently forwarded through feature resizing block which resizes the features at all Tiers to be equal in size for temporal fusion.

This results in K feature maps of shape $R^{T \times H_M \times W_M \times C_M}$ where M is the spatial dimension of the lowest resolution feature map. The K feature maps are forwarded to the Multi-Tier Temporal Fusion Transformer, where temporal features across Tiers are fused to learn a Tier-Aware Spatio-Temporal Embedding. Finally, the Tier-aware spatio-temporal embedding is passed through a Multi-Layer Perceptron (MLP) responsible for predicting the start and end frames of the resultant clip. To handle annotation noise and to separate highly similar classes, we introduce a Temporal Uncertainty Aware Localization loss and Multi-Tier, Dual Anchor Contrastive loss further discussed in section 6.3.4.

6.3.3 Multi-Tier Spatio-Temporal Transformer:

Multi-Tier Query Guided Spatial Transformer

The design of the encoder to fuse the video and the visual query features is crucial, especially in fine-grained video localization settings where the classes are highly similar. Previous work on visual grounding [80] and moment retrieval [174] naively concat the features from the video and query together. This can reduce the relevance of visual queries and result in features with low information about the visual query [175]. Moreover, these works are designed for text query-based video retrieval where the modality features are only extracted in a single hierarchy. However, features from videos and images can be extracted at multi-Tiers, each Tier containing coarse to fine-grained information. This variability in information can be beneficial for retrieval, especially in scenarios where the classes are highly similar with some local variations. To ensure video features at Tier k (f_{v_k}) are contextualised by visual query features (f_{q_k}) from the respective Tier, we designed a Multi-Tier Query Guided Spatial Transformer where $k = 1, 2, 3 \ldots K$. We achieve this by extracting features from K Tiers of the shared visual backbone for the video and the visual query. The extracted features for the video and the visual query for each Tier are then fused using cross-attention [159]. Formally, given the video feature f_{v_k} and visual query feature fq_k at Tier k where $k = 1, 2, 3 \ldots K$ and k=1 means the features from the last layer of the visual backbone. We project the video feature to get query Q_{v_k} whereas key K_{q_k} and value V_{q_k} are obtained from the visual query feature. Attention mechanism [159] is applied to Q_{v_k}, K_{q_k} and V_{q_k} and the output

is passed to a feed-forward network as shown in Eqn. 6.1 to give the Tier-specific query-aware video features QV_{f_k} for Tier k. The process is performed in parallel for all K Tiers to obtain Tier-specific query-aware features for each Tier.

$$QV_{f_K} = FFN\left(\text{softmax}\left(\frac{Q_{v_k}K_{q_k}^T}{\sqrt{d_k}}\right)V_{q_k}\right), \quad \text{where } K = 1, 2, 3, \ldots, k \tag{6.1}$$

Multi-Tier Temporal Fusion Transformer

To incorporate temporal information into the Tier-specific query-aware video features QV_{f_k} and to fuse these spatial-temporal features across Tiers for learning the final Tier-aware spatio-temporal embedding, we designed a multi-Tier temporal fusion transformer. Formally, given the Tier-specific query aware features for each Tier QV_{f_k} and a randomly initialized Tier-Aware Spatio-Temporal learnable embedding $E_{T_A} \in \mathbb{R}^{\mathbb{T}\times\mathbb{H}_\mathbb{M}\times\mathbb{W}_\mathbb{M}\times\mathbb{C}_\mathbb{M}}$ where k= $1, 2, 3 \ldots K$. We first add fixed 3D sine positional encoding to each Tier's feature vector and to the learnable embedding to enrich the features with positional information. Subsequently, we concatenate all the Tier features (QV_{f_K}) and learnable embedding (E_{T_A}) to give combined feature maps QV_f $\in R^{(K\times T)\times H_M\times W_M\times C_M}$. We then perform, self-attention [159] between the resulting vectors by projecting them to Q_v, K_v, and V_v to learn the Tier-Aware Spatio-Temporal Embedding, as stated in Eqn. 6.2. This helps in fusing the spatial and temporal information available across the Tiers into the Tier-Aware Spatio-Temporal Embedding which is then utilized by an MLP to make the final predictions.

$$E_{T_A} = FFN\left(\text{softmax}\left(\frac{Q_v\left(K_v^T\right)}{\sqrt{d_k}}\right)V_v\right) \tag{6.2}$$

6.3.4 Loss Functions

Multi-Tier, Dual Anchor Contrastive Loss

Existing contrastive learning methods typically rely on a single positive anchor to pull similar samples closer and push dissimilar ones away. However, this can be inadequate for fine-grained video tasks, such as fetal ultrasound analysis, where classes differ only in subtle, localized features. The reliance on a single positive anchor can result in suboptimal separation for event boundary estimation, where

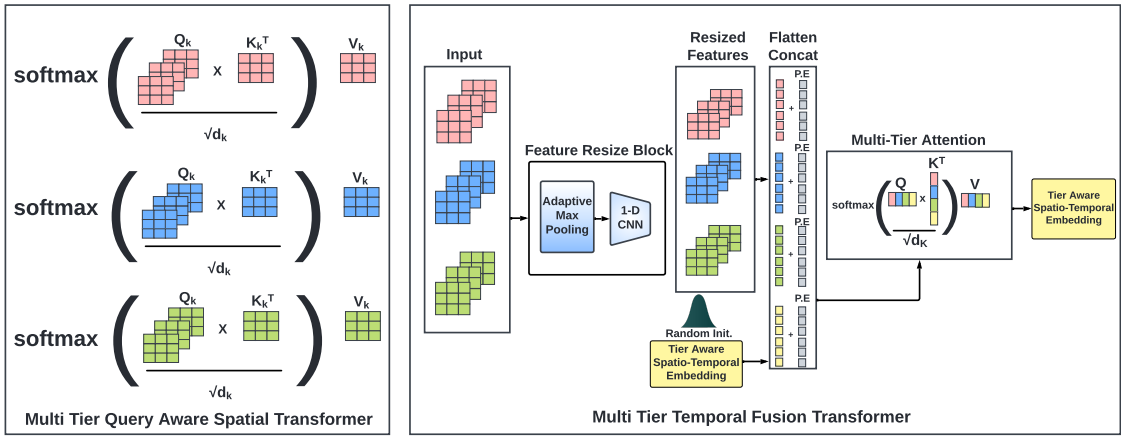


Figure 6.4: Fig (left) shows the spatial feature fusion mechanism where Tier-specific video and VQ features are spatially fused to give Tier-specific query-aware features. Figure 4 (right) shows how the Tier-specific query-aware features are first resized, flattened and enriched with positional information. The resulting features are concatenated and fused to learn the Tier-Aware Spatio-Temporal Embedding.

the challenges are further amplified by high spatial similarity among video frames (see Fig. 6.1). To overcome these limitations, we propose a Multi-Tier, Dual Anchor Contrastive Loss. This loss utilizes the positive anchor to enforce strong intra-class cohesion and the negative anchor, semantically similar but from a different class, to distinguish visually akin samples. By explicitly modeling both anchors, our method aims to achieve better separation between subtly distinct classes and capture the fine-grained inter-class boundaries essential for fetal ultrasound analysis.

The loss function has two main components 1) **Multi-Tier Positive Anchor Contrastive Loss (L_{PAC})** term which aims to bring the tier-specific visual query-aware features in the ground-truth clip together while pushing away features belonging to other classes. 2) a **Multi-Tier Negative Anchor Contrastive Loss(L_{NAC})** which utilizes a negative anchor and aims to further push the positive tier-specific query aware features away from negative.

Formally, given Tier-Specific Query Aware features f_{vq_k} for each Tier, we project the features to a shared feature space to ensure only rich-semantic features from each Tier are captured. We achieve this through a CNN projection layer $P_{\theta k}$ to give projected features f'_{vq_k}. Subsequently, we extract the video features belonging to the ground truth clip and define them as positive features ($f'^{+}_{vq_k}$) for each Tier. The video features of the frames lying outside the ground-truth clip are defined

as negative features ($f'^{-}_{vq_k}$). We randomly sample a Tier and utilize its features as anchors. A Tier's positive features serve as the positive anchor ($f^{+}_{vq_a}$) while negative features as the negative anchor ($f'^{-}_{vq_a}$) for the remaining Tiers. For the Positive Anchor Contrastive Loss, we calculate the cosine similarity between (($f'^{+}_{vq_a}$), ($f'^{+}_{vq_{k,i}}$)) and (($f'^{+}_{vq_a}$), ($f'^{-}_{vq_{k,j}}$)) as stated in Eqn. 6.3 where sim(.) denotes the cosine similarity function and i,j iterate over M_1 positive and M_2 negative samples while k iterates over $K-1$ Tiers. Finally, we optimize the loss function to pull positive features $f'^{+}_{vq_{k,i}}$ closer to the positive anchor feature $f'^{+}_{vq_a}$ while pushing all M_2 negative $f'^{-}_{vq_{k,j}}$ away as formulated in Eqn. 6.3 where τ^{+} is the positive temperature parameter.

$$\mathcal{L}_{PAC} = -\log \frac{\sum_{k=1}^{K-1}\sum_{i=1}^{M_1} \exp\left(sim(f'^{+}_{vq_a}, f'^{+}_{vq_{k,i}})/\tau^{+}\right)}{\sum_{k=1}^{K-1}\sum_{j=1}^{M_2} \exp\left(sim(f'^{+}_{vq_a}, f'^{-}_{vq_{k,j}})/\tau^{+}\right)} \tag{6.3}$$

On the other hand, for $\mathcal{L}_{NAC}$, we consider the negative features of the randomly selected Tier, as the negative anchor $f'^{-}_{vq_a}$. We calculate the cosine similarity of (($f'^{-}_{vq_a}$), ($f'^{-}_{vq_{k,i}}$)) and (($f'^{-}_{vq_a}$), ($f'^{+}_{vq_k,j}$)) where i,j iterate over M_2 negative, M_1 positive features while k iterates over $K-1$ Tiers as stated n Eqn. 6.4. Finally, we optimize the loss to pull the negative features $f'^{-}_{vq_{k,i}}$ closer to the negative anchor features $f'^{-}_{vq_a}$ while pushing all M_1 positive $f'^{+}_{vq_{k,j}}$ away as shown in Eqn. 6.4 where τ^{-} is temperature for $\mathcal{L}_{NAC}$.

$$\mathcal{L}_{NAC} = -\log \frac{\sum_{k=1}^{K-1}\sum_{i=1}^{M_2} \exp\left(sim(f'^{-}_{vq_a}, f'^{-}_{vq_{k,i}})/\tau^{-}\right)}{\sum_{k=1}^{K-1}\sum_{j=1}^{M_1} \exp\left(sim(f'^{-}_{vq_a}, f'^{+}_{vq_{k,j}})/\tau^{-}\right)} \tag{6.4}$$

The final loss $\mathcal{L}_{MTDA}$ is denoted in Eqn. 6.5 where w_p and w_n are tunable weights for $\mathcal{L}_{PAC}$ and $\mathcal{L}_{NAC}$ respectively.

$$\mathcal{L}_{MTDA} = w_p * \mathcal{L}_{PAC} + w_n * \mathcal{L}_{NAC} \tag{6.5}$$

Temporal Uncertainty Aware Localization Loss

The task of VQ-VCL becomes more challenging when there is a high similarity between the frames belonging to different classes and the event boundaries are not well defined. This leads to noisy annotations. To reduce the effect of noisy annotations, we introduce a Temporal Uncertainty Aware Localization Loss ($\mathcal{L}_{URL}$).

Instead of using binary ground truth, we generate two Gaussian distributions $T_s(x)$ and $T_e(x)$ corresponding to the true start frame (t_s) and true end frame (t_e) of the target video clip, with means $\mu_s = t_s$ and $\mu_e = t_e$ and standard deviation ($\sigma = 1$) respectively as shown in Eqn. 6.6 . Finally, we optimize the KL-divergence loss between the predicted ($P_s(x)$, $P_e(x)$) and true ($T_s(x)$, $T_e(x)$) start and end distribution and combine as shown in Eqs. 6.7 and 6.8 respectively.

$$T_s(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu_s)^2/2\sigma^2},\ T_e(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu_e)^2/2\sigma^2} \tag{6.6}$$

$$\begin{aligned} KL_s(P_s||T_s) &= \sum_x P_s(x)\log(\frac{P_s(x)}{T_s(x)}), \\ KL_e(P_e||T_e) &= \sum_x P_e(x)\log(\frac{P_e(x)}{T_e(x)}) \end{aligned} \tag{6.7}$$

$$\mathcal{L}_{URL} = KL_s + KL_e \tag{6.8}$$

Finally, we combine $\mathcal{L}_{MTDA}$ and $\mathcal{L}_{URL}$ together, to give the total loss $\mathcal{L}$ used to train our model as formulated in Eqn. 6.9.

$$\mathcal{L} = \mathcal{L}_{MTDA} + \mathcal{L}_{URL} \tag{6.9}$$

6.4 Experiments and Results

6.4.1 Dataset and Implementation

Following [215], we use ultrasound datasets from two research studies PULSE (Perception Ultrasound by Learning Sonographer Experience) [5], and CAIFE (Development of Clinical Artificial Intelligence Models in Fetal Echocardiography for the Detection of Congenital Heart Defects). Details are described next.

CAIFE dataset:

The CAIFE dataset comprises ultrasound (US) videos from participants who are over 18 years old and in their second trimester of pregnancy (20 weeks). Ethics approval was granted by the West of Scotland Research Ethics Service, UK Research Ethics Committee (Reference 18/WS/0051). An experienced fetal cardiologist collected

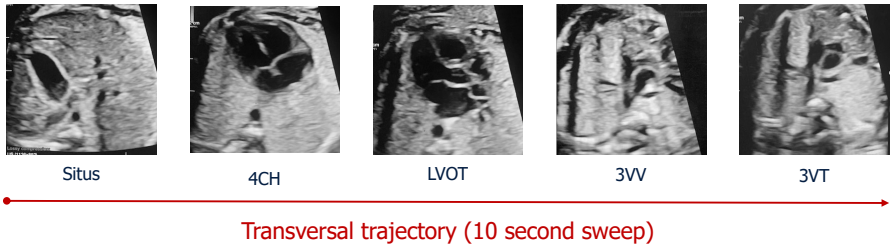


Figure 6.5: Technique for scanning fetal heart through transversal trajectory. (I) Situs view of the upper abdomen is visualized first. (II) The four-chamber view is obtained through an axial scanning plane across the fetal chest by moving and tilting the transducer in a cephalad direction. Further, cephalad movement of the transducer from the four-chamber view towards the fetal head gives the outflow-tract and great-vessel views sequentially: (III) left ventricular outflow-tract view; (IV) right ventricular outflow-tract view and the three-vessel view variants; and (V) three-vessel-and-trachea view.

the data using a standard curvilinear transducer (C2-9-D or C1-6-D) on GE Voluson US machines (models E8 or E10) from General Electric Healthcare. The videos capture a transverse trajectory with cephalad movement of the transducer, allowing visualization of five views of the fetal heart: from the cardiac situs (Situs) to the four-chamber view (4CH), through the left ventricular outflow tract (LVOT), the three-vessel view (3VV), and finally, the three-vessel trachea view (3VT) (see Figure 6.5). The data was extracted from the US machine as DICOM files to ensure high-quality graphics; videos and metadata were anonymized using DCMTK (DICOM Toolkit) and saved in high-definition resolution (1280×960 pixels). The videos are approximately 10 seconds long. We utilized 200 healthy heart sweep videos for training and 47 videos for testing. The visual query for the heart sweep data consisted of 2804 standard frames extracted from 12 held-out videos. These selected videos were manually annotated at the frame level, where individual frames received one of five labels (Situs, 4CH, LVOT, 3VV, 3VT). Unlike routine heart scans where the sonographer pauses to capture the perfect plane for each anatomical view, these sweeps continuously scan across the heart. The task involves retrieving a standard heart-view clip given a visual query of the standard heart view.

PULSE dataset:

The PULSE dataset [5] consists of free-hand ultrasound (US) videos including participants aged $>$ 18 years in their second trimester of pregnancy ($\approx$ 20 weeks).

Ethics approval was granted by the East Midlands, Leicester Central Research Ethics Committee (23/EM/0023). Data was obtained by fetal sonographers performing routine scans using a US system comprising a commercial General Electric Healthcare Voluson E8 or E10 machine equipped with standard curvilinear (C2-9-D, C1-6-D, C1-5-D) transducers. The data was recorded using the video output from the US machine, captured by a video grabbing card (DVI2PCIe) and purpose-built software to ensure real-time anonymization of the video. The saved videos include no personal details and consist of full-length scans recorded using the US machine in full high-definition resolution (1920 $\times$ 1080 pixels). We extracted video clips for 8 clinical anomaly detection anatomical planes: Transventricular and Transcerebellar views of the fetal head, Abdomen, Femur, and 4CH, LVOT, 3VV, and 3VT views of the fetal heart. We trained the TIER-LOC model on 200 videos and tested it on 30 videos. The visual query comprised 4378 standard frames extracted from the 30 test videos. The task involves retrieving a standard clip belonging to the anatomical plane represented by the visual query. Further details about dataset split and hyperparameter tuning are provided in the supplementary material.

Ego4D VQ-VCL

For the VQ-VCL task, open-source computer vision datasets are absent. Hence, we utilize the existing Ego4D [77] dataset and create a VQ-VCL fine-grained video dataset known as Ego4D VQ-VCL. We will release the dataset creation script, alongside the code, to ensure reproducibility.

We evaluated TIER-LOC on two different fetal ultrasound video datasets (CAIFE and PULSE) and one egocentric computer vision Ego4D VQ-VCL dataset. The video and visual query frames were resized to 224×224 dimensions. During training, we augmented the dataset by sampling clips with varying start and end frames, each containing 150 frames. All models were trained for 200 epochs in PyTorch version 1.8 using a Tesla V100 32 GB GPU. We utilized AdamW optimizer with StepLR learning scheduler with cosine annealing and step-size of 75. Our visual encoder was ResNet101 and both our multi-tier feature fusion transformers had 2 layers each. All hyperparameter tuning for both our model and baseline models, including temperature scaling factors in the contrastive terms (Eqs. 6.3 and 6.4)

and weighting factors in their sum (Eq. 6.5), was optimized via grid search. This optimization used a separate validation set and was conducted over 100 epochs. Specific parameter ranges for the grid search are detailed in Table 6.1. Further details regarding participant-level data splitting are available in the supplementary materials.

Hyperparameter	**Max**	**Min**
Learning Rate	1e-04	1e-06
τ^+	0.8	0.2
τ^-	0.8	0.2
w_p	0.8	0.2
w_n	0.8	0.2

Table 6.1: Grid search hyperparameter ranges used in hyperparameter optimization.

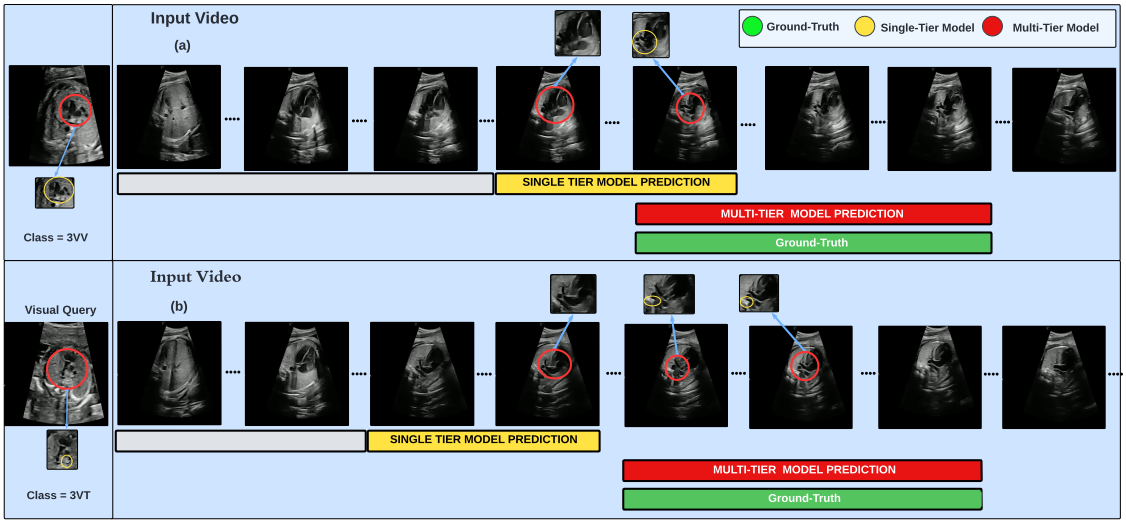


Figure 6.6: Figure showing the importance of the Multi-Tier Feature Fusion Transformer. We show that the model with a Single TIER transformer fails to detect minute local patterns (shown in patches above). For instance, in (a), the only difference between a 3VV view and the view before (LVOT) is the appearance of three-minute vessels (blobs) which are missed by a single TIER transformer while our Multi-TIER transformer can detect it and hence retrieve the correct clip.

Metrics

To quantify the performance of TIER-LOC compared to previous work on temporal video-grounding [175, 216] and our baselines [79, 172], we calculate mean temporal

intersection-over-union (mtIoU) and "R @ t" where R is recall calculated at predefined temporal IoU (tIoU) thresholds (t). For our experiments, we report recall at t = 0.1, 0.3, 0.5 and 0.7.

Table 6.2: Quantitative Comparison of TIER-LOC with Baselines on Various Datasets

Heart Sweep Data					
Method	**mtIoU(↑)**	**R@0.7(↑)**	**R@0.5(↑)**	**R@0.3(↑)**	**R@0.1(↑)**
CS Sup CNN	5.03	0.00	0.00	4.00	16.22
TubeDETR[80]	12.72	2.00	2.00	10.22	20.00
MomentDETR[174]	14.89	0.00	8.00	25.00	39.72
Resnet 3D [176]	19.79	6.00	6.00	23.22	47.17
VQLOC [172]	24.05	2.50	13.50	34.50	62.17
TIER-LOC (Ours)	**31.00**	**13.22**	**27.72**	**40.44**	**65.67**

PULSE Data [5]					
Method	**mtIoU(↑)**	**R@0.7(↑)**	**R@0.5(↑)**	**R@0.3(↑)**	**R@0.1(↑)**
CS Sup CNN	2.6	2.04	2.04	2.04	2.04
TubeDETR[80]	7.07	4.76	4.76	4.76	14.42
MomentDETR[174]	10.34	2.04	6.93	16.60	21.50
Resnet 3D [176]	17.89	14.29	17.14	22.04	**28.16**
VQLOC [172]	12.62	0.00	14.29	14.29	22.04
TIER-LOC (Ours)	**21.42**	**19.18**	**23.96**	**26.00**	26.00

Ego4D VQ-VCL Data [77]					
Method	**mtIoU(↑)**	**R@0.7(↑)**	**R@0.5(↑)**	**R@0.3(↑)**	**R@0.1(↑)**
CS Sup CNN	4.89	0.00	0.00	7.93	15.87
Resnet 3D [176]	10.72	3.57	8.33	12.70	20.24
VQLOC [172]	25.35	7.14	19.44	32.94	44.84
MomentDETR [174]	38.44	15.08	25.40	52.78	66.67
TubeDETR [80]	38.59	18.65	32.94	61.51	71.03
TIER-LOC (Ours)	**43.01**	**23.81**	**34.92**	**70.63**	**73.02**

Quantative Results

We compare our model TIER-LOC with ResNet3D CNN [176], Cosine Similarity Supervised 2D CNN, TubeDETR [80], VQLOC [172] and MomentDETR[174] as shown in Table 6.2.

1.ResNet 3D [176] We concatenate video frames and the visual query along the channel dimension and pass it to the model. We train the model with our L_{URL} loss.

2. Cosine Similarity Supervision CNN: We use ResNet 101 [173] encoder to extract features from the visual query and the video. Cosine similarity between the features of the visual query and each frame of the video is calculated and passed to cross-entropy loss to train the model.

3.TubeDETR: TubeDETR [80] is a SOTA video-grounding model. We adapt

TubeDETR to our task by replacing the text input with visual query input, the text encoder with a visual encoder and the bounding box prediction head with a temporal boundary prediction head.

4. VQLOC: VQLOC [172] is SOTA on the VQ2D task of the Ego4D dataset. We modify the prediction head to predict the start and end frame probabilities rather than bounding boxes.

5. Moment-DETR:. Moment-DETR is SOTA for Moment-Retrieval and Highlight detection tasks. We replace the prediction head to predict the start and end frame probabilities rather than saliency scores.

From Table 6.2 observe that the cosine similarity supervised baseline performs worst, achieving an mtIoU of 5.03%, R @ 0.7 = 0.00, R @ 0.5 = 0.0 . This indicates the model's inability to effectively extract, fuse, and model long-range features from both the input video and the visual query. TubeDETR, demonstrates improved performance with an mtIoU of 12.72%, R@ 0.3 = 10.22%,

This improvement may be attributed to the spatio-temporal transformer in TubeDETR, facilitating the extraction and fusion of video features in both spatial and temporal dimensions. However, the model still struggles with longer interactions, as indicated by R @ 0.7 and R @ 0.3 both being 2.00%, suggesting that the features from the visual query and the video are insufficient for modelling extended interactions. This limitation may be attributed to the direct concatenation of video and visual query features in the model. A similar pattern is observed with MomentDETR, where mtIoU is 14.89%. The model models short-range interactions well with R @ 0.3 = 25.00% and R @ 0.1 = 39.72%. However, it performs poorly in capturing longer-range interactions (R @ 0.7 = 0.00% and R @ 0.5 = 8.00%), possibly due to the concatenation of visual query and video features.

The ResNet3D baseline outperforms TubeDETR and MomentDETR, achieving a mtIoU of 19.79% and demonstrating better modelling of longer interactions with R @ 0.7 = 6.00% and R @ 0.5 = 6.00%. ResNet3D also performs well in modelling shorter interactions with R @ 0.1 = 47.17% This improvement may be attributed to the equal interaction of the visual query with each frame of the video, achieved by concatenating them together for each frame. VQLOC achieves a mtIoU of 24.05%, R @ 0.5 = 13.50, R @ 0.1 = 62.17%, indicating its ability to model both

short and long-range interactions. TIER-LOC outperforms all baselines with a mtIoU of 31.00%, nearly 7% higher than VQLOC. Its performance in modelling long-range and short-range dependencies is significantly better, with R @ 0.7 = 13.22%, R @ 0.5 = 27.72%, R @ 0.3 = 40.44, and R @ 0.1 = 65.67. TIER-LOC models the relationship between the visual query and the video in both spatial and temporal dimensions well due to Multi-Tier Spatio-Temporal Fusion Transformer. Additionally, the incorporation of boundary losses ($\mathcal{L}_{MTDA}$ and $\mathcal{L}_{URL}$) enhances its ability to detect boundaries, making it robust to noisy annotations and resulting in improved performance. A similar trend is seen in PULSE [5] data (refer Table 6.2) and Ego4D VQ-VCL (refer Table 6.2) with our model TIER-LOC outperforming the best-performing baseline by 4.00% and 4.42 % respectively.

6.4.2 Qualitative Results:

In our qualitative comparison, we evaluate a single-tier model, which uses features solely from the final layer (Tier=1), against our multi-tier model, which leverages features from multiple layers (Tiers=1, 2, 3). As illustrated in Fig. 6.6(a), the single-tier model struggles to distinguish between the LVOT and 3VV views in the video, as both views display partial four-chamber hearts. The key difference is the presence of the three vessels in the 3VV view, highlighted by the yellow bounding circle. Our multi-tier model effectively identifies and tracks this subtle variation by utilizing features from multiple tiers, resulting in significantly enhanced performance. Similar results are observed in Fig. 6.6(b), where our multi-tier model detects the subtle appearance of the trachea (highlighted in yellow bounding circle) , which helps distinguish between the 3VV and 3VT views, leading to the correct prediction of the start and end frames. In Fig. 6.7, we present additional qualitative results showing that our multi-tier model can detect subtle differences between the LVOT and 4CH views. Both views are highly similar, with four chambers visible, but the LVOT view includes the left-ventricle outflow tract (highlighted in yellow bounding circle in Fig. 6.7) , which the single-tier model misses. Our model's ability to detect this feature leads to a significant boost in localization performance.

6.4.3 Ablation Study

The section reports ablations experiments to justify the inclusion of the key components in the TIER-LOC model.

Table 6.3: Table showing the effect of multi-Tier features on TIER-LOC's performance where T is the number of frames in the input video.

Tier	**mtIoU**(↑)	**R@0.7**(↑)	**R@0.5**(↑)	**R@0.3**(↑)	**R@0.1**(↑)
1	21.24	7.00	17.22	29.72	47.22
1,2	27.74	6.72	**27.72**	38.72	55.89
1,2,3	**31.00**	**13.22**	**27.72**	40.44	65.67
1,2,3,4	27.44	4.5	11.00	**45.67**	**65.89**
1,2,3,4,5	28.78	8.5	20.72	37.94	65.39

Importance of Tiers

In this subsection, we show the importance of Tiers in our model's performance. The model utilising features only from the last layer of the visual backbone is referred to as Tier 1 while that utilising from Tier 1 and some layer before that is referred to as Tier 2 and so on. Experiments were performed for Tier = 1,2,3,4,5 where the feature sizes were $T \times 2048 \times 7 \times 7$, $T \times 1024 \times 14 \times 14$, $T \times 512 \times 28 \times 28$,$T \times 256 \times 56 \times 56$ and,$T \times 64 \times 112 \times 112$ respectively.From Table 6.3, observe that the model utilizing only Tier 1 features performs the worst with mtIoU = 21.24 %. This can be explained because Tier 1 features are insufficient to capture the fine-grained detail necessary to determine event boundaries and to distinguish between highly similar classes. When features from Tier 1 and Tier 2 are utilized observe that mtIoU increases by 6.5% showcasing the advantage of utilising features from these two Tiers. The highest performance is seen with Tier 3, which surpasses the single Tier results by almost 10% and Tier 2 by 3.26% respectively stressing the advantage of using multi-tier features to capture both low-level and high-level details to distinguish fine-grained classes. For Tier 4, 5 we see that the performance is better than Tier 1, but the balance between coarse and fine-grained is not optimal resulting in reduced performance compared to Tier 3.

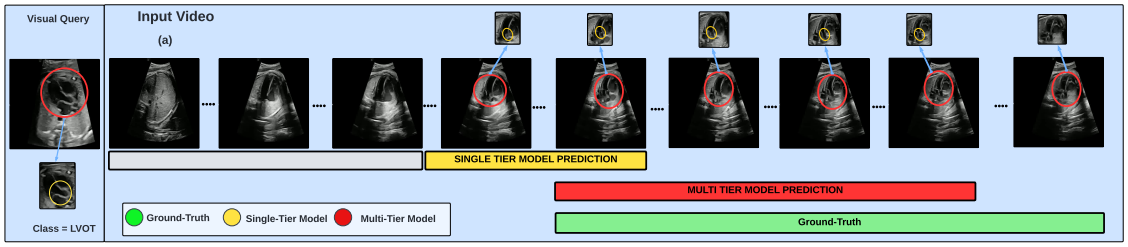


Figure 6.7: This figure compares the predictions of a single-tier model with those of our multi-tier model for LVOT visual query. Our multi-tier model excels at detecting subtle differences, resulting in precise boundary predictions. In the given example, the single-tier model struggles to differentiate between the 4CH and LVOT views in the video. Both views are highly similar, with four chambers visible. The key distinction is the appearance of an outflow tract in the left ventricle in the LVOT view as highlighted in yellow bounding circle. Our multi-tier model successfully identifies this subtle change and tracks it, leading to significantly improved performance.

Table 6.4: Ablation study showing the effect of different loss functions on TIER-LOC performance.

$\mathcal{L}_{URL}$	$\mathcal{L}_{PVAC}$	$\mathcal{L}_{NVAC}$	**mtIoU(↑)**	**R@0.7(↑)**	**R@0.5(↑)**	**R@0.3(↑)**	**R@0.1(↑)**
✗	✗	✗	16.27	5.00	9.00	19.00	40.72
✓	✗	✗	29.28	11.00	25.44	33.44	63.17
✓	✓	✗	29.17	13.00	25.22	32.22	65.67
✓	✗	✓	28.73	9.00	25.22	38.72	63.17
✓	✓	✓	**31.00**	**13.22**	**27.72**	**40.44**	**65.67**

Importance of different Loss functions

In Table 6.4, we investigate the impact of each loss function on our model's performance. Our base loss was Cross-Entropy Loss shown as the first row of Table 6.4. However, we notice, that replacing it with $\mathcal{L}_{URL}$ boosts the performance by 13% mtIoU depicting the importance of incorporating uncertainty in loss functions when the ground truth is noisy. Further, we ablate the Positive Anchor Contrastive Loss ($\mathcal{L}_{PAC}$) and Negative Anchor Contrastive Loss ($\mathcal{L}_{NAC}$) components of the Multi-Tier, Dual Anchor Contrastive loss. We observe that using $\mathcal{L}_{PAC}$, along with $\mathcal{L}_{URL}$, boosts R@0.7, R@0.1 by 2% and 2.5% respectively. Adding $\mathcal{L}_{NAC}$ along with $\mathcal{L}_{URL}$ leads to a significant increase in R @ 0.3 (5.28%) while degradation of other metrics. Combining the two anchor losses ($\mathcal{L}_{PAC} + \mathcal{L}_{NAC}$) with $\mathcal{L}_{URL}$, boosts mtIoU by 2%, R @ 0.7 by 3.22 %, R @ 0.5 by 2.28%, R @ 0.3 by 7% and R @ 0.1 by 2.5% indicating the importance of both positive and negative anchors in separating highly similar classes and reducing confusion at event boundaries.

Table 6.5: MLP Decoder vs Our Decoder

Decoder	mtIoU(↑)	R@0.7(↑)	R@0.5(↑)	R@0.3(↑)	R@0.1(↑)
MLP	22.97	10.5	16.72	31.22	42.22
Ours	**31.00**	**13.22**	**27.72**	**40.44**	**65.67**

Table 6.6: Multi-Scale Fusion

Method	mtIoU(↑)	R@0.7(↑)	R@0.5(↑)	R@0.3(↑)	R@0.1(↑)
Sequential Fusion	21.01	7.00	12.00	18.00	42.72
Parallel Fusion (Ours)	**31.00**	**13.22**	**27.72**	**40.44**	**65.67**

Importance of Multi-Tier Temporal Fusion Transformer

To investigate the importance of our Multi-Tier Temporal Fusion Transformer decoder as compared to traditional MLP-based decoders, we replace our decoder with a 2-layer MLP. As shown in Table 6.5, our decoder outperforms the MLP decoder by 8.03%, depicting the importance of optimally fusing multi-tier features across the temporal dimension as achieved by our Multi-Tier Temporal Fusion Transformer.

Sequential vs Parallel Fusion

In existing works utilising multi-scale features [208, 209] the sequential feature fusion is employed. In Sequential fusion, multi-scale features are projected to common feature space and fused sequentially in a single representation from coarse to fine. In contrast, in our work to exploit the implicit bias of image/video modality and capture minute variations, we perform parallel fusion in our encoder. In parallel fusion, features from the video and visual query at each tier are fused separately in original resolution across tiers. We compare our parallel fusion with existing sequential fusion in Table 6.6 where parallel fusion outperforms sequential fusion by 10% mtIoU.

Importance of Depth of Visual Backbone

In this section, we consider the effect of model depth on Tier features and hence model performance. For each of the three visual backbones, we extract Tier features from the same relative position to the last layer (stride) and having the same feature dimensions. In Table 6.7, observe that increasing the number of layers from 50 to 101 leads to an increase in 2.03% mtIoU. This can be explained because by increasing the number of layers, the feature hierarchies are refined to allow Tiers to extract a more abstract representation while still having relatively coarse features in shallow

Tiers (Tier 3). However, when the number of layers is increased to 152 in ResNet 152 [173] the mtIoU drops by almost 6%. This can be because increasing the number of layers might lead to the extraction of highly abstract features even in shallow Tiers (Tier 3) and thus loss of coarse features leading to a reduction in feature diversity.

Table 6.7: Table showing the effect of depth of visual backbone.

Method	**mtIoU**(↑)	**R@0.7**(↑)	**R@0.5**(↑)	**R@0.3**(↑)	**R@0.1**(↑)
ResNet50	28.97	8.5	23.22	37.94	**68.11**
ResNet101	**31.00**	**13.22**	**27.72**	**40.44**	65.67
ResNet152	25.04	6.5	20.72	37.94	54.44

Video Query Fusion

In this section, we ablate the effect of different techniques for fusing the visual query features with the video features. We experiment with the Concat Self-Attention (SA) which is popular for multi-modality fusion [80, 174, 217] and Cross-Attention feature fusion techniques. In Table 6.8, cross-attention fusion significantly outperforms Concat SA fusion by 14.39% mtIoU. This is because when we do cross-attention fusion, the Key/Value comes from the visual query while Query comes from the video. This ensures adequate contribution of visual query features resulting in fused representations that are query-aware. In Concat SA, the visual query features are directly concatenated to the video features and self-attention is performed on the whole sequence. This leads to reduced contribution of visual query features in the fused representation as the VQ feature is just an element in the sequence.

Table 6.8: Table showing the effect of different fusion techniques to fuse visual query and the video.

Method	**mtIoU**(↑)	**R@0.7**(↑)	**R@0.5**(↑)	**R@0.3**(↑)	**R@0.1**(↑)
Concat Self-Attention	16.61	4.5	11.00	24.00	35.17
Cross-Attention	**31.00**	**13.22**	**27.72**	**40.44**	**65.67**

6.5 Conclusion

This paper explores the Visual-Query-based Video Clip Localization (VQ-VCL) task and develops a video-based transformer, TIER-LOC, to model this task. Unlike related works, TIER-LOC utilizes multi-Tier features, enabling the model

to acquire a nuanced and holistic comprehension of the video and its intricate connection with the visual query. This translates into notably superior video-clip localization compared to models employing single-Tier features. Further, to tackle naturally occurring noise in real-world annotations, a temporal uncertainty-aware loss is introduced, allowing the model to focus on relevant features and reducing its over-sensitivity to noisy instances. Finally, to distinguish highly similar classes in fine-grained videos, a contrastive loss that employs multi-Tier features and guidance of multi-anchors is introduced to learn subtle class-discriminative features. TIER-LOC has been evaluated on real-world ultrasound video datasets featuring limited training data and highly similar anatomical classes. By returning a short clip rather than a single reference frame, the system captures a more comprehensive view of the anatomy—particularly for dynamic anatomies such as the fetal heart—while also reducing sonographer workload and freeing them to focus on detailed anomaly detection. This ability to retrieve clinically relevant video segments offers a promising strategy to enhance efficiency in ultrasound scanning and to streamline existing clinical workflows.

Supplementary Material

Participant Level Dataset Split

Figure 6.8 illustrates how the CAIFE fetal-heart sweep dataset was split into training, validation, and test sets. Note that the figure shows unique participants, each of whom contributes multiple short clips spanning different heart views.

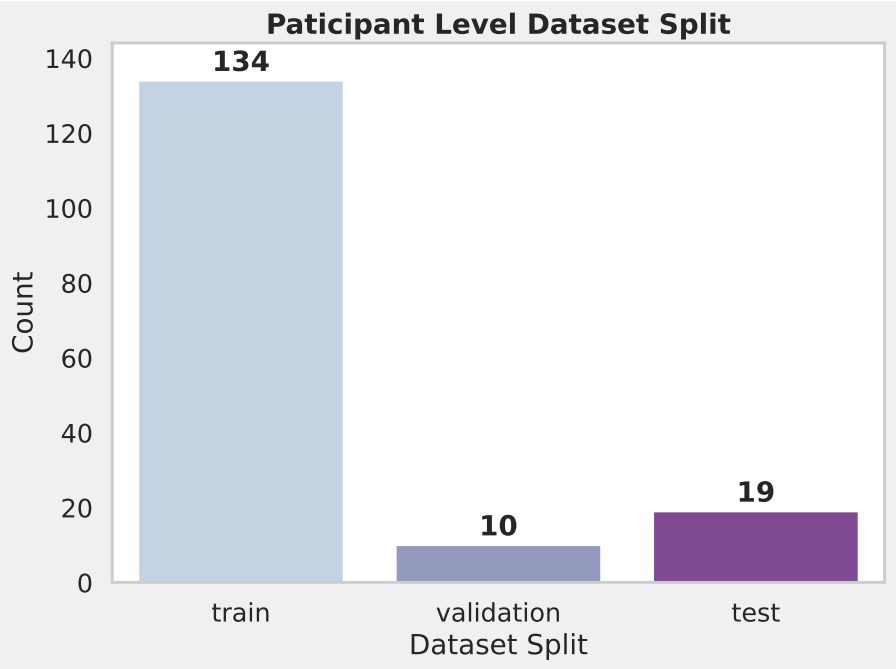


Figure 6.8: Illustration of the CAIFE fetal-heart sweep dataset split into training, validation, and test sets. Each participant contributes multiple short clips corresponding to different heart views.

7

MCAT: Visual Query-Based Localization of Standard Anatomical Clips in Fetal Ultrasound Videos Using Multi-Tier Class-Aware Token Transformer

Authors: Divyanshu Mishra, Pramit Saha, He Zhao, Netzahualcoyotl Hernandez-Cruz ,Olga Patey, Aris T. Papageorghiou, J. Alison Noble

Published in Conference: In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2025)

Background:

Existing visual query-based video clip localization (VQ-VCL) methods in fetal ultrasound rely heavily on shared or generic embeddings, making them less effective at capturing subtle, anatomy-specific differences essential for fine-grained localization in clinical practice. These limitations often result in reduced accuracy and efficiency for standard plane acquisition tasks. In this chapter, we introduce MCAT, a Multi-Tier Class-Aware Token Transformer published in AAAI 2025, which overcomes these challenges by learning class-specific tokens within a multi-tier transformer framework and representing anatomical structures at multiple scales. Evaluation on two fetal ultrasound video datasets demonstrates that MCAT achieves 10% and 13% higher mean Intersection over Union (mIoU) compared to state-of-the-art VQ-VCL baselines, and achieves a 5.35% gain on the Ego4D dataset, while reducing token

usage by 96%. These results underscore MCAT's effectiveness and efficiency for accurate standard plane localization in medical imaging.

Author Contribution: I was the lead technical author of the paper, responsible for formulating the problem statement, proposing the solution, designing the codebase, conducting the experiments, and preparing the original manuscript draft. Pramit Saha and He Zhao contributed to technical discussions and participated in reviewing the paper. Netzahualcoyotl Hernandez-Cruz did the data management for the project and Olga Patey helped in data collection. Aris T. Papageorghiou provided clinical supervision throughout the project. J. Alison Noble conceived the overall objectives of the study, secured funding, and supervised the entire project. All authors reviewed and approved the final version of the manuscript.

Abstract

Accurate standard plane acquisition in fetal ultrasound (US) videos is crucial for fetal growth assessment, anomaly detection, and adherence to clinical guidelines. However, manually selecting standard frames is time-consuming and prone to intra- and inter-sonographer variability. Existing methods primarily rely on image-based approaches that capture standard frames and then classify the input frames across different anatomies. This ignores the dynamic nature of video acquisition and its interpretation. To address these challenges, we introduce Multi-Tier Class-Aware Token Transformer (MCAT), a visual query-based video clip localization (VQ-VCL) method, to assist sonographers by enabling them to capture a quick US sweep. By then providing a visual query of the anatomy they wish to analyze, MCAT returns the video clip containing the standard frames for that anatomy, facilitating thorough screening for potential anomalies. We evaluate MCAT on two ultrasound video datasets and a natural image VQ-VCL dataset based on Ego4D. Our model outperforms state-of-the-art methods by 10% and 13% mIoU on the ultrasound datasets and by 5.35% mIoU on the Ego4D dataset, using 96% fewer tokens. MCAT's efficiency and accuracy have significant potential implications for public health, especially in low- and middle-income countries (LMICs), where it may enhance prenatal care by streamlining standard plane acquisition, simplifying US-based screening, diagnosis and allowing sonographers to examine more patients.

7.1 Introduction

Fetal ultrasound is essential for monitoring prenatal development, detecting potential abnormalities, and ensuring the health of both the fetus and the expectant mother. In routine pregnancy assessments, a sonographer scans different fetal anatomies to assess fetal development and identify anomalies. Selecting standard frames [6, 165, 218] that meet clinical guidelines (*e.g.*, ISUOG) is a time-consuming process, and a typical fetal ultrasound scan can take up to an hour. Multiple studies have attempted to streamline this process by automatically identifying standard planes in 2D fetal ultrasound using deep learning-based classification models [181, 190–193]. Other recent works have looked into leveraging temporal information for more complex tasks, such as anomaly detection in ultrasound videos [188, 189] and generative modeling of standard planes [187], although these approaches do not explicitly localize standard frames. Integrating a video-clip localization model could enhance sonographer workflow by allowing them to focus on detailed video reviews and anomaly detection. However, automatically detecting standard frames in video is challenging due to the high similarity of frames before and after the standard ones, making it difficult to determine temporal anatomical boundaries, as shown in Fig. 7.1. Additionally, even human experts may find it hard to agree on standard frame selection, as evidenced by our study showing a kappa score of only 66% between two fetal cardiologists annotating the same fetal heart videos (see Supp. Fig. 1), highlighting the complexity and inherent noise in annotations. Text query-based localization tasks, such as video-temporal grounding [80, 178, 179], video moment retrieval [175, 180], and highlight detection, have shown promising performance in natural video understanding. However, textual data often falls short of providing the dense video understanding required for some applications. In the medical domain, reports traditionally rely on static images and text to convey diagnostic information [219–222]. While image-based methods are informative, video-based analysis can offer a significant advancement in diagnostic capabilities. For instance, a dynamic ultrasound video of a beating heart provides a more detailed and holistic assessment of cardiac function compared to a single static frame [223]. Similarly, in fetal anatomy examinations, video clips allow practitioners to measure biometry

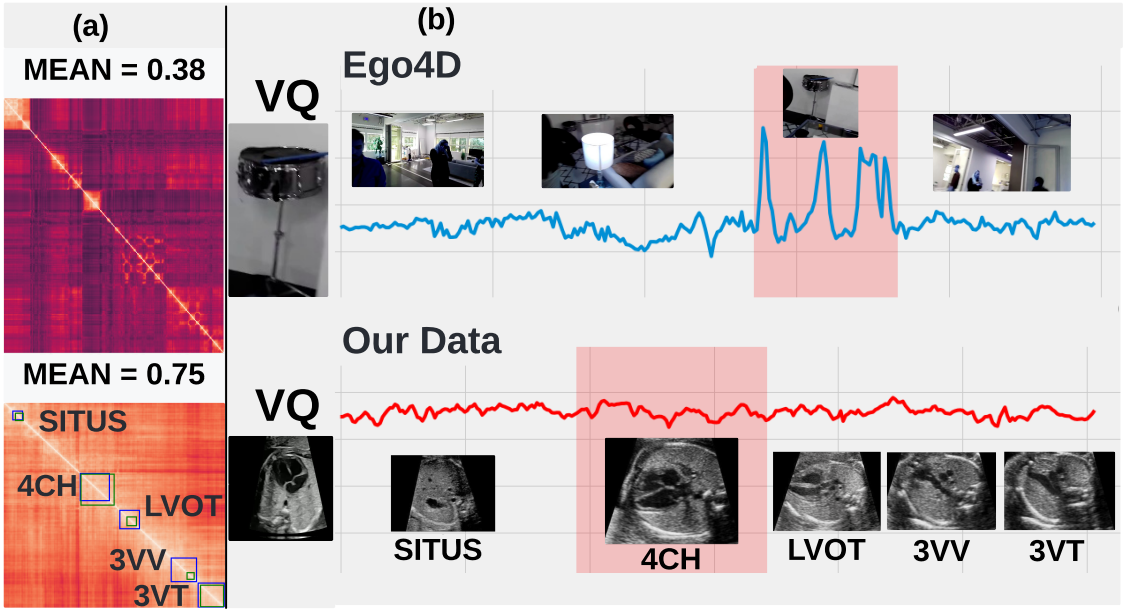


Figure 7.1: (a) Self-similarity matrix for a randomly chosen video from Ego4D (**top**, mean=0.3) [77] and our clinical video dataset (**bottom**, mean=0.75), which reveals higher task difficulty for our video clip localization task. The uncertainty in the annotations of two expert cardiologists is shown in green and blue boxes. (b) Cosine similarity of the visual query with the video for both Ego4D (top) and our data (bottom). Our clinical data obtains similar scores along the video emphasizing the challenge, whereas Ego4D exhibits high scores only within region of interest.

more accurately and review the entire sequence for optimal plane selection, thereby enhancing diagnostic precision. Despite the advantages, paired video-textual data is typically scarce in the medical field. When available, it usually includes sparse class labels or radiology reports providing a diagnosis for the entire video rather than detailed clip-level information. This is where image-based queries, or visual queries (VQs), are potentially valuable. VQs allow for intuitive and direct identification of objects or similar images, reducing language barriers and effectively expressing complex concepts that might be difficult to articulate through text. For example, describing a medical anomaly can be challenging with a text query, whereas an example frame containing the anomaly can provide a more effective query for model training. In the context of ultrasound videos, retrieving a video clip rather than a single frame is more challenging due to the motion of the ultrasound probe and the scanned object, leading to various deformations, occlusions, and motion blur, making it harder to localize all instances of the object as shown in Fig. 7.1. To

reduce the time to conduct a full scan assessment, we introduce the Visual-Query-based Video Clip Localization (VQ-VCL) task. In this approach, a sonographer performs a quick sweep to capture all relevant anatomies. With a visual query depicting the required anatomy, our method can automatically select the relevant standard-frame clips from this video sweep. This significantly reduces manual effort, enhances efficiency, and allows sonographers to scan more patients while focusing more on analyzing the standard video clips.

To tackle the challenges of the VQ-VCL task, we introduce MCAT, a Multi-Tier transformer-based model with class-specific tokens. It consists of three primary components: a Multi-Tier Class-Aware Spatio-Temporal Transformer for modeling spatial and temporal interactions and learning class-specific features through class-specific tokens, a Temporal Uncertainty Localization Loss to mitigate label noise, and a Multi-Tier, Dual Anchor Contrastive Loss for addressing complex event boundaries.

Our contributions are as follows:

1. We introduce the VQ-VCL task and propose MCAT, a spatio-temporal video Transformer model for automatic standard-plane video clip retrieval.

2. We propose a multi-tier feature extraction module to learn spatio-temporal features in a coarse-to-fine manner. A query-aware Transformer captures spatial information, while temporal information is condensed into class-specific learnable tokens. These tokens disentangle class-specific features into distinct tokens, improving video clip localization and significantly boosting model efficiency by reducing the number of tokens by 96%. This makes the approach potentially suitable for applications in resource-constrained public health including low- and middle-income country (LMIC) settings.

3. We propose a hybrid loss function comprising Multi-Tier, Dual Anchor Contrastive Loss, and Temporal Uncertainty-Aware Localization Loss to handle complex event boundaries and noisy labels.

4. We assess model performance on two real-world clinical datasets for standard-plane detection with limited data and annotations which naturally contain a high degree of noise. Additionally, we create and evaluate our model on an open-source VQ-VCL natural videos dataset based on Ego4D [77].

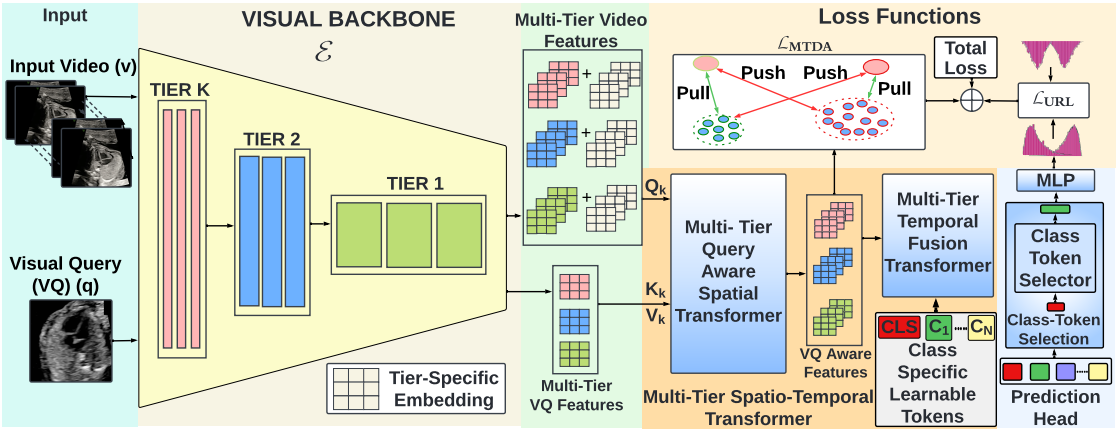


Figure 7.2: Main architecture of MCAT. The input video v and visual query q are passed to the visual backbone to give multi-Tier features. These features are fused spatially using the Multi-Tier Query Aware Spatial Transformer. The Tier-specific features are passed to a) $\mathcal{L}_{MTDA}$ to learn the separation between classes, b) the Multi-Tier Temporal Fusion transformer to learn Tier-Aware Spatio-Temporal Embedding, which is further passed to an MLP to make final prediction and calculate $\mathcal{L}_{URL}$ loss.

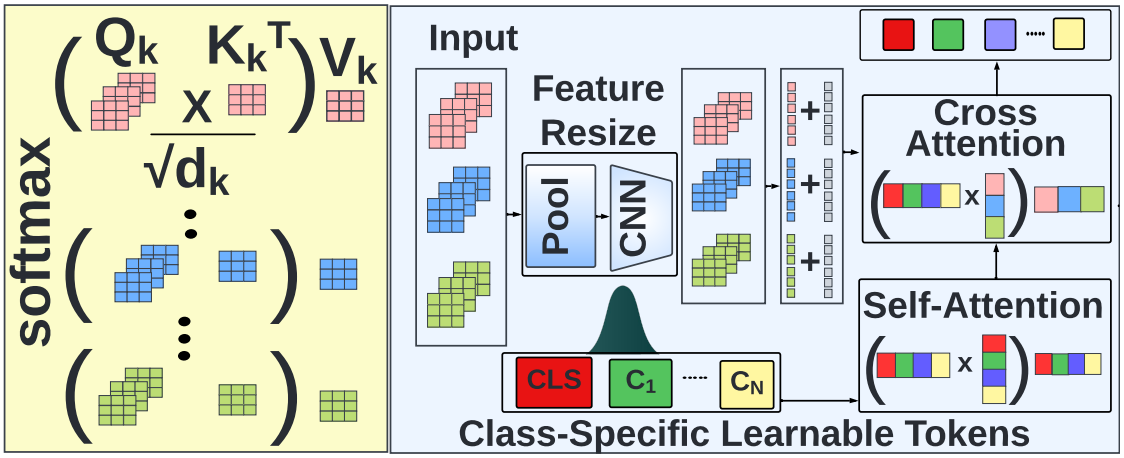


Figure 7.3: Fig (left) shows the spatial feature fusion mechanism where Tier-specific video and VQ features are spatially fused to give Tier-specific query-aware features. Figure 3 (right) shows how the Tier-specific query-aware features are first resized, flattened and enriched with positional information. The resulting features are concatenated and fused to learn the Tier-Aware Spatio-Temporal Embedding.

7.2 Methods

7.2.1 Video Clip Localization Task Formulation

The visual query-based video clip localization (VQ-VCL) task is formulated as a temporal localization task. Formally, given a video v and an exemplar frame q from a separate exemplar database $\mathcal{Q}$, the model is trained to predict the start

(t_s) and end (t_e) frame number of a clip v_q where $v_q \subset v$ and contains frames semantically similar to q.

7.2.2 MCAT Overall Architecture

Our proposed model, illustrated in Fig. 7.2, processes a video v and a visual query q as inputs. These inputs are fed through a shared encoder $\mathcal{E}$, generating K tier video features $f_{v_K} \in \mathbb{R}^{T \times H_k \times W_k \times C_k}$ and visual query features $f_{q_k} \in \mathbb{R}^{H_k \times W_k \times C_k}$, where k iterates over the K tiers and T, H_k, W_k, and C_k represent the number of frames, height, width, and channel dimensions at tier k. The features extracted at each tier from a visual backbone $\mathcal{E}$ are enriched with scale-aware learnable embeddings and spatially fused using a Multi-Tier Query-Guided Spatial Transformer, resulting in multi-tier VQ aware features as shown in Fig. 7.2. A multi-tier temporal fusion transformer is proposed to learn a series of tokens (CLS, $\{C_i | i = 1, \cdots, N\}$) from the VQ aware features where N is the number of classes. The tokens are further utilized to predict the start and end frames.

7.2.3 Multi-Tier Spatio-Temporal Transformer

Multi-Tier Query Guided Spatial Transformer

The design of the encoder to fuse the video and visual query features is crucial, especially in fine-grained video localization settings where the classes are highly similar. Previous work on visual grounding [80] and moment retrieval [174] naively concatenates the features from the video and query together. This approach can diminish the relevance of visual queries and result in features with low information about the visual query [175]. Moreover, these works are designed for text query-based video retrieval where modality features are only extracted in a single hierarchy. In contrast, features from videos and images can be extracted at multiple tiers, each tier containing coarse to fine-grained information. This variability in information can be beneficial for retrieval, especially in scenarios where the classes are highly similar with some local variations. To ensure video features at Tier k (f_{v_k}) are contextualized by visual query features (f_{q_k}) from the respective tier, we designed a Multi-Tier Query Guided Spatial Transformer where $k = 1, 2, 3, \ldots, K$. We achieve this by extracting features from K tiers of the shared visual backbone for the video and the visual query. Tier-specific learnable embeddings (emb_K) are added to each tier video feature to ensure optimal learning and fusion of tier-specific information from the video and visual query. The resulting video features and visual query features for each tier are then fused using cross-attention [159] to learn these tier-specific embeddings (emb_K). Formally, given the video feature f_{v_k} and visual query feature f_{q_k} at tier k where $k = 1, 2, 3, \ldots, K$ and $k = 1$ means the features from the last layer of the visual backbone. We first add to each tier video

feature the tier-specific learnable embedding (emb_K) such that $f_{v_k} = f_{v_k} + emb_k$. We project the video feature to get query (Q_{v_k}), whereas key (K_{q_k}) and value (V_{q_k}) are obtained from the visual query feature. The attention mechanism [159] is applied to Q_{v_k}, K_{q_k}, and V_{q_k}, and the output is passed to a feed-forward network as shown in Eq. 7.1 to produce the tier-specific query-aware video features (QV_{f_k}) for tier k. This process is performed in parallel for all K tiers to obtain tier-specific query-aware features QV_{f_K} for each tier.

$$QV_{f_K} = FFN\left(\text{softmax}\left(\frac{Q_{v_k}K_{q_k}^T}{\sqrt{d_k}}\right)V_{q_k}\right) \tag{7.1}$$

Multi-Tier Temporal Fusion Transformer

To incorporate temporal information into the Tier-specific query-aware video features QV_{f_k} and to fuse these spatio-temporal features across Tiers for learning the class-specific Tier-aware spatio-temporal tokens (CLS, $\{C_i|i = 1, \cdots, N\}$) where N is number of classes, we designed a multi-Tier temporal fusion transformer. Formally, given the Tier-specific query-aware features for each Tier QV_{f_k}, and randomly initialized class-selection token CLS, Class-Specific Tier-Aware Spatio-Temporal learnable tokens $C_N \in \mathbf{R}^{(\mathbf{N})\times\mathbf{H_M}\times\mathbf{W_M}\times\mathbf{C_M}}$ where k= $1, 2, 3 \ldots K$. We first perform self-attention between the $E_T = CLS + C_N$ tokens as in Eq. 7.2.

$$E_T = FFN\left(\text{softmax}\left(\frac{Q_v\left(K_v^T\right)}{\sqrt{d_k}}\right)V_v\right) \tag{7.2}$$

Cross-attention is performed between the resulting vector and the Tier-specific query-aware video features QV_{f_k} as formulated in Eq.7.3 and shown in Fig. 7.3

$$E_T = FFN\left(\text{softmax}\left(\frac{Q_{E_T}\left(K_{QV_f}^T\right)}{\sqrt{d_k}}\right)V_{QV_f}^T\right) \tag{7.3}$$

This helps fuse the spatial and temporal information available across the Tiers into the class-specific Tier-Aware Spatio-Temporal tokens. The spatio-temporal information-rich E_T tokens are fed to the token selection block that helps select the token corresponding to the VQ and updates only the selected token with the current spatio-temporal class-specific information.

Class-Specific Token Selection and Learning

The block is designed to select the class-specific token (C_S) corresponding to the visual query and to enable class-specific token learning. During training, the class-selection token (CLS) obtained after spatio-temporal fusion is passed through a multi-layer perceptron (MLP) to predict the class to which the visual query (VQ) belongs. This is formulated as an N-class classification problem, and cross-entropy loss is used to train the MLP. Since the class of the VQ is known during training, we use this information to select the class-specific token (C_S) and only update the token for the specific VQ class. During inference, the prediction from the trained MLP is used to select C_S and predict the start and end frames of the ground-truth video clip.

7.2.4 Loss Functions

Multi-Tier, Dual Anchor Contrastive Loss

In settings with high spatial similarity between the video frames, as seen in Fig. 7.1, estimating the correct event boundary is challenging. Moreover, in such a case, object appearance can significantly vary from that of the visual query as the objects of interest and the data acquisition device are both in motion. To mitigate the above challenges and learn subtle differences between the classes, we propose a Multi-Tier, Dual Anchor Contrastive Loss, where the anchors and samples are selected from different tiers. The loss function has two main components: 1. **Multi-Tier Positive Anchor Contrastive Loss (L_{PAC})**, which aims to bring the tier-specific visual query-aware features in the ground-truth clip together while pushing away features belonging to other classes. 2. **Multi-Tier Negative Anchor Contrastive Loss (L_{NAC})**, which utilizes a negative anchor to further push the positive tier-specific query-aware features away from the negative ones. Formally, given Tier-Specific Query Aware features f_{vq_k} for each tier, we project the features to a shared feature space to ensure that only rich-semantic features from each tier are captured. This is achieved through a CNN projection layer $P_{\theta k}$, resulting in projected features f'_{vq_k}. Subsequently, we extract the video features belonging to the ground-truth clip and define them as positive features ($f'^{+}_{vq_k}$) for each tier. The video features of the frames lying outside the ground-truth clip are defined as negative features

($f'^{-}_{vq_k}$). We randomly sample a tier and utilize its features as anchors. A tier's positive features serve as the positive ($f^{+}_{vq_a}$), while the negative features serve as the negative anchor ($f'^{-}_{vq_a}$) for the remaining tiers. For the Positive Anchor Contrastive Loss, we calculate the cosine similarity between ($f'^{+}_{vq_a}$, $f'^{+}_{vq_k,i}$) and ($f'^{+}_{vq_a}$, $f'^{-}_{vq_k,j}$) as stated in Eq. 7.4, where sim(.) denotes the cosine similarity function and i, j iterate over M_1 positive and M_2 negative samples, while k iterates over $K-1$ tiers.

$$\mathcal{L}_{PAC} = -\log \frac{\sum_{k=1}^{K-1}\sum_{i=1}^{M_1} \exp\left(sim(f'^{+}_{vq_a}, f'^{+}_{vq_k,i})/\tau^{+}\right)}{\sum_{k=1}^{K-1}\sum_{j=1}^{M_2} \exp\left(sim(f'^{+}_{vq_a}, f'^{-}_{vq_k,j})/\tau^{+}\right)} \tag{7.4}$$

Finally, we optimize the loss function to pull positive features $f'^{+}_{vq_k,i}$ closer to the positive anchor feature $f'^{+}_{vq_a}$ while pushing all M_2 negative features $f'^{-}_{vq_k,j}$ away, as formulated in Eq. 7.4, where τ^{+} is the positive temperature.

On the other hand, for $\mathcal{L}_{NAC}$, we consider the negative features of the randomly selected tier as the negative anchor ($f'^{-}_{vq_a}$). We calculate the cosine similarity between ($f'^{-}_{vq_a}$, $f'^{-}_{vq_k,i}$) and ($f'^{-}_{vq_a}$, $f'^{+}_{vq_k,j}$), where i and j iterate over M_2 negative and M_1 positive features, respectively, while k iterates over $K-1$ tiers, as stated in Eq. 7.5.

$$\mathcal{L}_{NAC} = -\log \frac{\sum_{k=1}^{K-1}\sum_{i=1}^{M_2} \exp\left(sim(f'^{-}_{vq_a}, f'^{-}_{vq_k,i})/\tau^{-}\right)}{\sum_{k=1}^{K-1}\sum_{j=1}^{M_1} \exp\left(sim(f'^{-}_{vq_a}, f'^{+}_{vq_k,j})/\tau^{-}\right)} \tag{7.5}$$

Finally, we optimize the loss to pull the negative features ($f'^{-}_{vq_k,i}$) closer to the negative anchor features ($f'^{-}_{vq_a}$) while pushing all M_1 positive features ($f'^{+}_{vq_k,j}$) away, as shown in Eq. 7.5, where τ^{-} is the temperature parameter for $\mathcal{L}_{NAC}$. The final loss $\mathcal{L}_{MTDA}$ is given in Eq. 7.6 where w_p and w_n are tunable weights for $\mathcal{L}_{PAC}$ and $\mathcal{L}_{NAC}$ respectively.

$$\mathcal{L}_{MTDA} = w_p * \mathcal{L}_{PAC} + w_n * \mathcal{L}_{NAC} \tag{7.6}$$

Temporal Uncertainty Aware Localization Loss

The VQ-VCL task becomes more challenging when there is a high similarity between the frames belonging to different classes and the event boundaries are not well defined. This leads to noisy manual annotations. To reduce the effect of noisy annotations, we introduce a Temporal Uncertainty Aware Localization Loss ($\mathcal{L}_{URL}$). Instead of using binary ground truth, we generate two Gaussian distributions $T_s(x)$

and $T_e(x)$ corresponding to the true start frame (t_s) and true end frame (t_e) of the target video clip, with means $\mu_s = t_s$ and $\mu_e = t_e$ and standard deviation ($\sigma = 1$) respectively as shown in Eq. 7.7.

$$T_s(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu_s)^2/2\sigma^2},\ T_e(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu_e)^2/2\sigma^2} \tag{7.7}$$

Finally, we optimize the KL-divergence loss between the predicted ($P_s(x)$, $P_e(x)$) and true ($T_s(x)$, $T_e(x)$) start and end distribution and combine as shown in Eqs. 7.8 and 7.9 respectively.

$$\begin{aligned} KL_s(P_s||T_s) &= \sum_x P_s(x) \log(\frac{P_s(x)}{T_s(x)}), \\ KL_e(P_e||T_e) &= \sum_x P_e(x) \log(\frac{P_e(x)}{T_e(x)}) \end{aligned} \tag{7.8}$$

$$\mathcal{L}_{URL} = KL_s + KL_e \tag{7.9}$$

Finally, we combine Eqs 7.6 and 7.9 to give the total loss $\mathcal{L}$ used to train our model as formulated in Eq. 7.10.

$$\mathcal{L} = \mathcal{L}_{MTDA} + \mathcal{L}_{URL} \tag{7.10}$$

7.3 Experiments and Results

Dataset and Implementation

We evaluated MCAT on two distinct fetal ultrasound video datasets following [215] and one egocentric computer vision dataset, Ego4D VQ-VCL, which we created based on the Ego4D dataset [77]. The first dataset consists of fetal heart video sweeps from CAIFE (Development of Clinical Artificial Intelligence Models in Fetal Echocardiogra- phy for the Detection of Congenital Heart Defects). It includes 10-second transversal heart sweeps over the fetal heart (see Supp. Fig 3), scanning from the cardiac situs (Situs) to the four-chamber view (4CH), through the left ventricular outflow tract (LVOT), the three-vessel view (3VV), and finally, the three-vessel trachea view (3VT) of the fetal heart. Unlike routine heart scans, where the sonographer pauses to capture the perfect plane for each anatomical view, these sweeps continuously scan across the heart. The VQ-VCL task retrieves

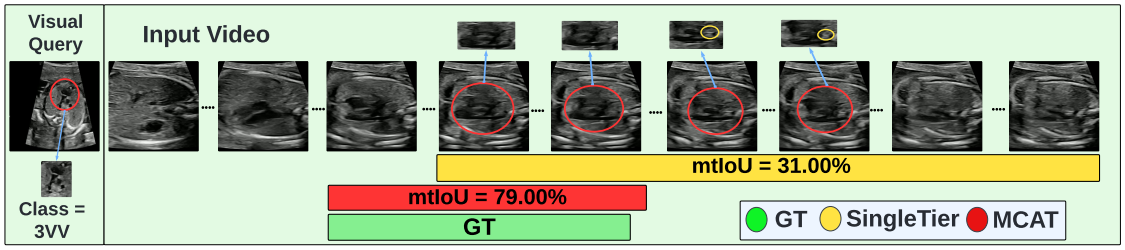


Figure 7.4: This figure compares the predictions of a single-tier model with our multi-tier model for an LVOT visual query.

a standard heart-view clip given a visual query of the standard heart view. We used 200 healthy heart sweep videos for training and 47 videos for testing. The visual query for the heart sweep data consisted of 2804 standard frames extracted from 12 held-out videos. Further details about our unique heart sweep data are in ultrasound dataset details section of Supplementary. Our second dataset is derived from the PULSE [5] fetal ultrasound anomaly scan video dataset. We extracted clips for 8 fetal anatomical planes utilized for clinical anomaly detection, including Transventricular and Transcerebellar Views of the fetal head, Abdomen, Femur, and the 4CH, LVOT, 3VV, and 3VT views of the fetal heart. We trained the MCAT model on 200 videos and tested it on 30 videos. The visual query comprised 4378 standard frames extracted from 30 videos. As we introduce the VQ-VCL task, we acknowledge the lack of open-source datasets for model reproducibility. Therefore, we utilized the existing Ego4D [77] dataset to create the Ego4D VQ-VCL dataset. We plan to release the dataset creation script along with our code to ensure the reproducibility of our work. Video and visual query frames were resized to dimensions of 224×224. During training, we augmented the dataset by sampling clips with varying start and end frames, each containing 150 frames. All models were trained for 200 epochs in PyTorch version 1.8 using a Tesla V100 32 GB GPU. We employed AdamW optimizer with a StepLR learning scheduler, utilizing cosine annealing with a step-size of 75. Our visual encoder was ResNet101, and both our multi-tier feature fusion transformers consisted of 2 layers each.

Metrics

To evaluate the performance of MCAT, we follow previous works on temporal video grounding [175, 216] and our baselines [79, 172]. Hence, we compute the mean

temporal intersection-over-union (mtIoU) and "R @ t", where R represents recall measured at predefined temporal IoU (tIoU) thresholds (t). For our experiments, we report recall at thresholds $t = 0.1, 0.3, 0.5$ and 0.7.

Heart Sweep Data					
Method	**mtIoU**	**R@0.7**	**R@0.5**	**R@0.3**	**R@0.1**
CS Sup CNN	5.03	0.00	0.00	4.00	16.22
TubeDETR	12.72	2.00	2.00	10.22	20.00
MomentDETR	14.89	0.00	8.00	25.00	39.72
Resnet 3D	19.79	6.00	6.00	23.22	47.17
VQLOC	24.05	2.50	13.50	34.50	62.17
MCAT (Ours)	**34.1**	**11.00**	**30.17**	**56.17**	**66.17**
PULSE [5] Data					
Method	**mtIoU**	**R@0.7**	**R@0.5**	**R@0.3**	**R@0.1**
CS Sup CNN	2.6	2.04	2.04	2.04	2.04
TubeDETR	7.07	4.76	4.76	4.76	14.42
MomentDETR	10.34	2.04	6.93	16.60	21.50
Resnet 3D	17.89	14.29	17.14	22.04	28.16
VQLOC	12.62	0.00	14.29	14.29	22.04
MCAT (Ours)	**30.63**	**26.80**	**31.70**	**34.56**	**39.32**
Ego4D [77] VQ-VCL Dataset					
Method	**mtIoU**	**R@0.7**	**R@0.5**	**R@0.3**	**R@0.1**
CS Sup CNN	4.89	0.00	0.00	7.93	15.87
Resnet 3D	10.72	3.57	8.33	12.70	20.24
VQLOC	25.35	7.14	19.44	32.94	44.84
MomentDETR	38.44	15.08	25.40	52.78	66.67
TubeDETR	38.59	18.65	32.94	61.51	71.03
MCAT (Ours)	**43.94**	**32.54**	**39.68**	**64.68**	**71.83**

Table 7.1: Quantitative comparison of MCAT

Quantitative Results

We compare, MCAT with ResNet3D CNN [176], Cosine Similarity Supervised 2D CNN, TubeDETR [80], VQLOC [172], and MomentDETR [174], as in Table 7.1. Further details about baselines are in Supp.

From Table 7.1, we observe that the cosine similarity supervised baseline performs worst, achieving an mtIoU of 5.03%, R @ 0.7 = 0.00% , R @ 0.5 = 0.0% . This indicates the model's inability to effectively extract, fuse, and model long-range

features from both the input video and the visual query. TubeDETR demonstrates improved performance with an mtIoU of 12.72%, R @ 0.3 = 10.22%. This improvement may be attributed to the spatio-temporal transformer in TubeDETR, facilitating the extraction and fusion of video features in both spatial and temporal dimensions. However, the model still struggles with longer interactions, as indicated by R @ 0.7 and R @ 0.3 both being 2.00%, suggesting that the features from the visual query and the video are insufficient for modeling extended interactions. This limitation may be due to the direct concatenation of video and visual query features in the model. A similar pattern is observed with MomentDETR, where mtIoU is 14.89%. The model handles short-range interactions well with R @ 0.3 = 25.00% and R @ 0.1 = 39.72%, but it performs poorly in capturing longer-range interactions (R @ 0.7 = 0.00% and R @ 0.5 = 8.00%), possibly due to the concatenation of visual query and video features. The ResNet3D baseline outperforms TubeDETR and MomentDETR, achieving an mtIoU of 19.79% and demonstrating better modeling of longer interactions with R @ 0.7 = 6.00% and R @ 0.5 = 6.00%. ResNet3D also performs well in modeling shorter interactions with R @ 0.1 = 47.17%. This improvement may be attributed to the equal interaction of the visual query with each frame of the video, achieved by concatenating them together for each frame. VQLOC achieves an mtIoU of 24.05%, R @ 0.5 = 13.50%, R @ 0.1 = 62.17%, indicating its ability to model both short and long-range interactions. MCAT outperforms all baselines with a mtIoU of 34.10%, 10.05% higher than VQLOC. Its performance in modeling long-range and short-range dependencies is significantly better, with R @ 0.7 = 11.00%, R @ 0.5 = 30.17%, R @ 0.3 = 56.17%, and R @ 0.1 = 66.17%. MCAT effectively models the relationship between the visual query and the video in both spatial and temporal dimensions due to the Multi-Tier Spatio-Temporal Fusion Transformer and disentanglement of class-specific spatio-temporal features through class-specific tokens. Additionally, the incorporation of boundary losses ($\mathcal{L}_{MTDA}$ and $\mathcal{L}_{URL}$) enhances its ability to detect boundaries, making it robust to noisy annotations and resulting in improved performance. A similar trend is seen in the PULSE [5] data (refer to Table 7.1) and Ego4D VQ-VCL (refer to Table 7.1), with our model MCAT outperforming the best-performing baseline by 12.74% and 5.35%, respectively.

Qualitative Results

In the qualitative comparison, we analyze the single-tier model, which only utilizes features from the last layer (Tier=1), against our multi-tier model (Tiers=1, 2, 3). As shown in Fig. 7.4, the single-tier model struggles to differentiate between the 3VV and 3VT views in the video, as both views display three vessels. The critical distinction is the appearance of the trachea in the 3VT view, as highlighted by the yellow circle. Our multi-tier model successfully identifies this subtle change by leveraging features from multiple tiers, leading to significantly improved performance. Additional qualitative results are provided in the supplementary.

7.3.1 Ablation Study

This section reports ablation experiments to justify the inclusion of the key components in the MCAT model.

Importance of Tiers

First we show the importance of Tiers to model performance. The model utilizing features only from the last layer of the visual backbone is referred to as Tier 1, while that utilising from Tier 1 and some layer before that is referred to as Tier 2, and so on. Experiments were performed for Tier = 1,2,3 where the feature sizes were $T \times 2048 \times 7 \times 7$, $T \times 1024 \times 14 \times 14$ and $T \times 512 \times 28 \times 28$ respectively. Table 7.2 shows that the model utilising only Tier 1 features performs the worst with mtIoU = 28.23 %. This can be explained because Tier 1 features are insufficient to capture the fine-grained detail necessary to determine event boundaries and to distinguish between highly similar classes. When features from Tier 1 and Tier 2 are utilized, mtIoU increases by 3.1%. The highest performance is seen with Tier 3, which surpasses the single Tier results by almost 6% and Tier 2 by 2.77% respectively stressing the advantage of using multi-Tier features to capture both low- and high-level details to distinguish fine-grained classes.

Tier	mtIoU	R@0.7	R@0.5	R@0.3	R@0.1
1	28.23	10.5	19.72	38.72	**66.61**
1, 2	31.33	**13.0**	28.22	42.44	**66.61**
1, 2, 3	**34.10**	11.00	**30.17**	**56.17**	**66.17**

Table 7.2: Showing importance of multi-Tier features.

Importance of different Loss functions

In Table 7.3, we investigate the impact of each loss function on model performance. Our baseline loss is Cross-Entropy Loss, as shown in the first row of Table 7.3. Replacing it with $\mathcal{L}_{URL}$ improved performance by 14% mtIoU, highlighting the significance of incorporating uncertainty in loss functions when the ground truth annotation contains a high degree of noise. Additionally, we ablate the Multi-Tier, Dual Anchor Contrastive loss ($\mathcal{L}_{MTDA}$). Including this loss, which consists of our dual-anchor losses ($\mathcal{L}_{PAC}$ and $\mathcal{L}_{NAC}$), alongside $\mathcal{L}_{URL}$, further enhances performance. Specifically, mtIoU increases by 4.46%, R @ 0.5 by 3.95%, R @ 0.3 by 11.23%, and R @ 0.1 by 9.78%. These improvements indicate the importance of both positive and negative anchors to distinguish highly similar classes and to reduce confusion at event boundaries.

$\mathcal{L}_{URL}$	$\mathcal{L}_{MTDA}$	mtIoU	R@0.7	R@0.5	R@0.3	R@0.1
✗	✗	15.93	2.00	12.00	22.22	42.44
✓	✗	29.64	11.50	26.22	44.94	56.39
✓	✓	**34.10**	**11.00**	**30.17**	**56.17**	**66.17**

Table 7.3: Analysis of contribution of different loss functions.

Sequential vs Parallel Fusion

In existing works utilizing multi-scale features [208, 209], sequential feature fusion is employed. In sequential fusion, multi-scale features are projected into a common feature space and fused sequentially from coarse to fine. In contrast, our work employs parallel fusion in the encoder to exploit the implicit bias of image/video modalities and capture minute variations. In parallel fusion, features from the video and visual query at each tier are fused separately, maintaining the original resolution across tiers. As shown in Table 7.4, parallel fusion outperforms sequential fusion,

improving R@0.3 by 11.45%, R@0.5 by 6.67%, and mtIoU by 2.33%, demonstrating its superiority in capturing short-term and long-term interactions between the video and visual query.

Method	mtIoU	R@0.7	R@0.5	R@0.3	R@0.1
Sequential Fusion	31.77	**11.00**	23.50	44.72	**68.39**
Parallel Fusion (Ours)	**34.10**	**11.00**	**30.17**	**56.17**	66.17

Table 7.4: Effect of sequential and parallel fusion.

Video Query Fusion

We examined the impact of Concat Self-Attention (SA) method, which is popular for multi-modality fusion [80, 174, 217], with Cross-Attention feature fusion for fusing visual query features with video features. As shown in Table 7.5, cross-attention fusion significantly outperforms Concat SA fusion by 20.63% mtIoU. This improvement is because, in cross-attention fusion, the Key/Value is derived from the visual query while the Query comes from the video. This ensures a substantial contribution from the visual query features, resulting in query-aware fused representations. In contrast, Concat SA involves directly concatenating the visual query features with the video features and performing self-attention on the entire sequence. This approach reduces the contribution of visual query features in the fused representation, as the VQ feature becomes just one element within the sequence.

Method	mtIoU	R@0.7	R@0.5	R@0.3	R@0.1
Concat Self-Attention	14.47	6.00	8.00	20.00	**34.22**
Parallel Fusion (Ours)	**34.10**	**11.00**	**30.17**	**56.17**	**66.17**

Table 7.5: Comparing methods for video-visual query fusion.

Class-Specific Tokens vs Generic Embedding

In existing works such as [80, 172], the temporal transformer learns a generic embedding that is shared across classes and corresponds to the number of frames in the video. While this approach may be effective for coarse-grained videos, it results in sub-optimal performance for fine-grained videos. As shown in Table

7.6, our class-specific embedding outperforms the generic embedding by 3.1% with 96% fewer tokens.

Method	mtIoU	R@0.7	R@0.5	R@0.3	R@0.1
Generic Embedding	31.00	**13.22**	**27.72**	40.44	65.67
Class-Specific Embedding (Ours)	**34.10**	11.00	**30.17**	**56.17**	**66.17**

Table 7.6: Generic Embedding vs Class-Specific Token

Tier-Specific Embedding

We demonstrate the importance of using scale-specific embedding in the Query-Guided Spatial Transformer. As shown in Table 7.7, Tier-Specific Embedding improves model performance by 3.52%, highlighting its crucial role in capturing tier-specific features essential for fine-grained video retrieval.

Method	mtIoU	R@0.7	R@0.5	R@0.3	R@0.1
W/O Tier-Specific Embedding	30.58	**15.50**	25.72	42.67	62.61
W/ Tier-Specific Embedding	**34.10**	11.00	**30.17**	**56.17**	**66.17**

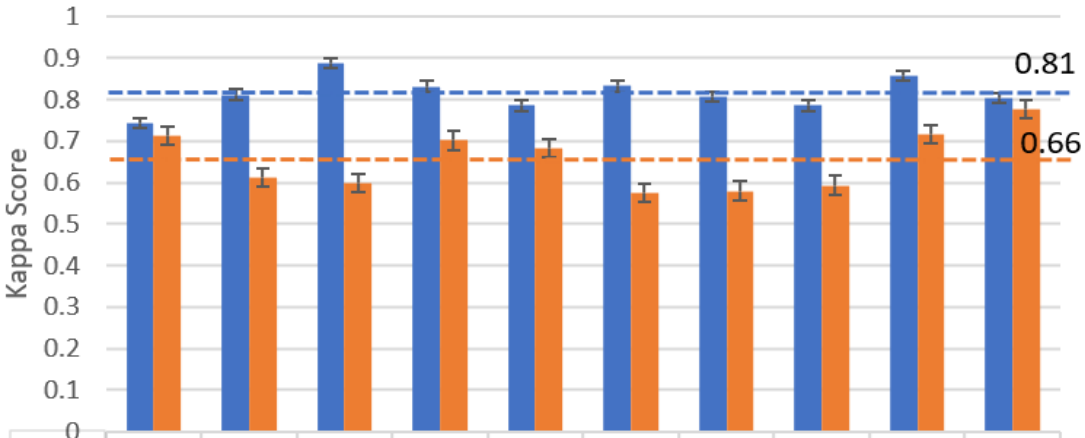
Table 7.7: Importance of Tier-Specific Embedding

7.4 Conclusion

This paper introduces an visual-query based solution for detecting standard anatomy video clips in fetal US videos. Our model MCAT, is a video-based transformer that leverages multi-tier features and class-specific token learning to understand the video with the visual query. This significantly improves video-clip localization compared to models that use single-tier features with 96% less tokens.This enables the model to retrieve the relevant video clip based on a visual query in just 2.69 seconds while using only 4.62 GB of memory during inference, allowing it to run effectively on affordable GPUs, even in resource-limited settings. Additionally, we introduce a temporal uncertainty-aware loss to handle the inherent noise in real-world annotations. Furthermore, to differentiate highly similar classes in fine-grained videos, we propose a contrastive loss that utilizes multi-tier features and multi-anchor guidance to learn subtle class-discriminative features. We apply MCAT to real-world standard plane video-clip detection task with limited data and

fine-grained classes and validate its effectiveness through comparisons with SOTA baselines, demonstrating significant improvements in localization accuracy and resource efficiency. These traits make our model beneficial for prenatal care in LMICs, where access to advanced diagnostic tools and skilled health professionals is limited.

Supplementary Material



Scan ID	1	2	3	4	5	6	7	8	9	10
Intra	0.7434	0.8112	0.8884	0.8322	0.7863	0.8327	0.807	0.7861	0.8577	0.8033
Inter	0.713	0.6114	0.5992	0.7015	0.6828	0.5746	0.5798	0.5933	0.717	0.7774

Figure 7.5: Figure showing the inter and intra-annotation agreement of annotators for the task of standard frame detection in Transversal heart sweep(TS). We can see that the Kappa score between annotators is only 66%, highlighting the difficulty of the problem

7.5 Inter and Intra-Annotator Analysis

Two experienced cardiologists were asked to annotate 10 fetal heart transverse (TV) sweep videos in this analysis. Each annotator was required to annotate the same set of 10 videos frame-by-frame, utilizing five cardiac labels. For intra-annotator analysis, the annotators had to annotate the same videos twice, with a two-week interval between annotations. Moreover, inter-annotator agreement was evaluated between the two cardiologists. Both intra and inter-annotator agreements were quantified on a frame-by-frame basis using the Cohen Kappa score [166], and the mean Kappa score for each video is reported in Fig. 7.1. As shown in Fig. 7.1, the mean inter-annotator agreement across videos was only 66%, highlighting the task's difficulty. This results in naturally occurring noise in the video annotations, especially at event boundaries, making the task harder.

7.6 Baseline Details:

1. ResNet 3D [176]: We concatenate video frames and the visual query along the channel dimension and pass it to the model. We train the model with our $\mathcal{L}_{URL}$ loss.

2. Cosine Similarity Supervision CNN: We use a ResNet 101 [173] encoder to extract features from the visual query and the video. Cosine similarity between the features of the visual query and each frame of the video is calculated and passed to cross-entropy loss to train the model.

3. TubeDETR: TubeDETR [80] is a SOTA video-grounding model. We adapt TubeDETR to our task by replacing the text input with visual query input, the text encoder with a visual encoder, and the bounding box prediction head with a temporal boundary prediction head.

4. VQLOC: VQLOC [172] is SOTA on the VQ2D task of the Ego4D dataset. We modify the prediction head to predict the start and end frame probabilities rather than bounding boxes.

5. MomentDETR: MomentDETR [174] is SOTA for Moment-Retrieval and Highlight detection tasks. We replace the prediction head to predict the start and end frame probabilities rather than saliency scores.

7.7 Qualitative Results

This section compares the importance of the Multi-Tier Feature Fusion Transformer. As shown in Fig. 7.6, the model with a Single Tier transformer fails to detect minute local patterns, highlighted using the red bounding circles. For instance, in Fig. 7.6(a), the model cannot detect the subtle appearance of four chambers and confuses it with SITUS. In contrast, our model detects subtle differences across the two views. Specifically, it successfully identifies the four chambers and predicts the start frames. Additionally, the multi-tier feature enables our model to differentiate between the highly related 4CH and LVOT views. Although both views contain four chambers, they differ in the presence of a very small left-outflow tract in the LVOT view. Our multi-tier feature detects this distinction and accurately predicts the end of the 4CH clip just before the left-outflow tract appears.

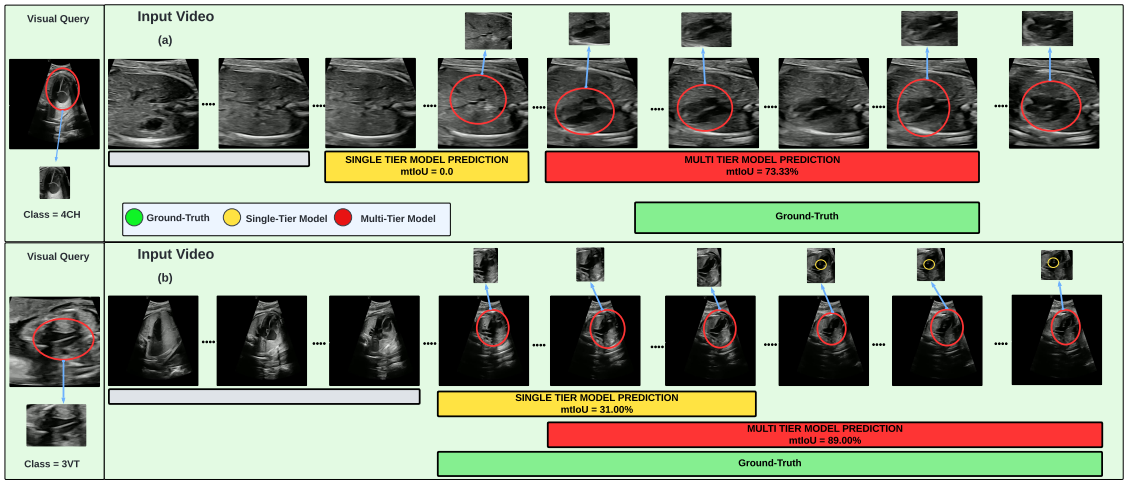


Figure 7.6: Qualitative comparison between single-Tier and Multi-Tier (Ours) models. Multi-Tier model can understand both coarse and fine-grained features, leading to improved video-clip localization.

In Fig. 7.6(b), the single-tier model successfully detects the starting frames of the 3VT view but fails to identify the outline of the pulmonary artery (highlighted in the yellow circle) and misses the remaining frames of the 3VT view. In contrast, with its ability to analyze both coarse and fine-grained features, our multi-tier model captures minute details and detects all frames of the 3VT view.

7.8 Ultrasound dataset details

7.8.1 Heart Sweep Dataset

The heart sweep dataset consists of ultrasound (US) videos collected from participants over 18 years old who are in their second trimester of pregnancy (approximately 20 weeks). Ethical approval for the study was obtained from the Health Research Authority, Care Research Wales and the Research Ethics Committee (Ref: 23/EM/0023; IRAS Project ID: 317510) [128]. Data collection was performed by an experienced fetal cardiologist using a standard curvilinear transducer (C2-9-D or C1-6-D) on GE Voluson US machines, specifically models E8 or E10 from General Electric Healthcare.

These videos document a transverse sweep with cephalad movement of the transducer, capturing five key views of the fetal heart introduced by our project CAIFE [128]. The views range from the cardiac situs (Situs) to the four-chamber

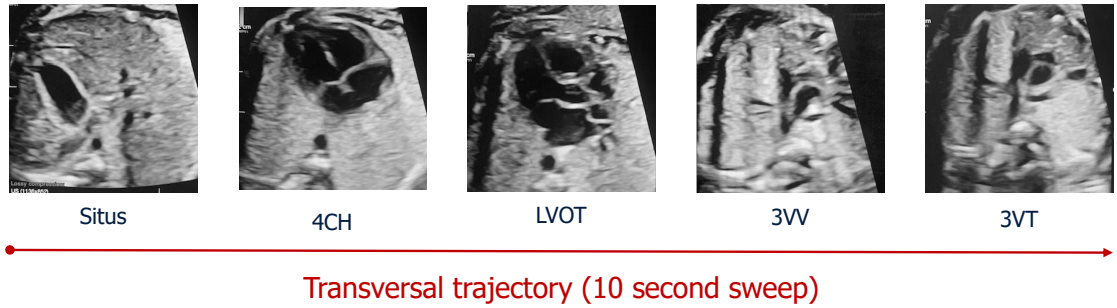


Figure 7.7: Figure depicting our 10 seconds traversal heart sweep.

view (4CH) and continue through the left ventricular outflow tract (LVOT), the three-vessel view (3VV), and the three-vessel trachea view (3VT) (refer to Figure 7.7). The data was extracted from the US machine as DICOM files to maintain high resolution. The videos and associated metadata were anonymized using the DCMTK (DICOM Toolkit) and stored in high-definition resolution (1280 × 960 pixels), with each video lasting approximately 10 seconds. This work used 200 videos of healthy heart sweeps for training, while 47 videos were reserved for testing. The visual query for the heart sweep data included 2804 standard frames taken from 12 separate videos. These selected videos were manually annotated at the frame level, with each frame being assigned one of the five labels: Situs, 4CH, LVOT, 3VV, or 3VT.

7.9 Training Details

In Table 7.8, we present the training details. All models were trained using the PyTorch framework version 1.8 on Tesla V100 32 GB GPUs. We trained all models and baselines for 200 epochs with a Step-LR learning rate scheduler with a step size of 75 and a batch size of 1. We uniformly sampled 150 frames from each video and utilized the Adam optimizer with weight decay to optimize all models. The Multi-Tier Query Guided Transformer comprised 2 layers for each tier, while the Multi-Tier Temporal Fusion Transformer consisted of 2 transformer encoder layers. For the Multi-Tier, Dual Anchor Contrastive Loss ($\mathcal{L}_{MTDA}$), we found that setting $\tau^{+} = 0.4$ and $\tau^{-} = 0.2$ resulted in the best performance.

Table 7.8: Training Details

Component	**Value**
Framework	Pytorch
Pytorch Version	1.8
Optimizer	AdamW
Epochs	200
Number of Frames	150
Learning Rate	1e−05
LR Scheduler	Step
LR Scheduler Step Size	75
τ^+	0.4
τ^-	0.2
w1	1
w2	1
Layers in Query Aware Spatial Transformer (Per-Tier)	1
Layers in Temporal Fusion Transformer Layers	2
Visual Encoder ($\mathcal{E}$) Architecture	ResNet101
GPU	Tesla V100 32 GB

8

Self-supervised Normality Learning and Divergence Vector-guided Model Merging for Zero-shot Congenital Heart Disease Detection in Fetal Ultrasound Videos

Authors: Pramit Saha* , Divyanshu Mishra* , Netzahualcoyotl Hernandez-Cruz , Olga Patey , Aris Papageorghiou , Yuki M. Asano , and J. Alison Noble (* equal contribution)

Published in Conference: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2025)

Background: Collaborative deep learning for congenital heart disease (CHD) detection in fetal ultrasound is hampered by limited annotated data, strict privacy requirements, and the logistical barriers of pooling information across multiple hospital sites—challenges that are especially acute for rare diseases. In this chapter, we present Sparse Tube Ultrasound Distillation (STUD), a privacy-preserving zero-shot detection framework published in MICCAI 2025, which enables decentralized anomaly detection without the need for raw data sharing. STUD leverages self-supervised learning on normal cases at each institution and then merges models using a divergence-guided strategy (DivMerge) that preserves both privacy and generalization. When evaluated on real-world fetal ultrasound data from five hospitals, STUD's merged model delivers 23.77% higher accuracy and a 30.13%

boost in F1-score over individual site-specific models, underscoring its potential for scalable, secure clinical deployment.

Author Contribution: I was joint first author of the paper, responsible for self-supervised video understanding, video normality modeling, and congenital heart disease (CHD) detection, including formulating the problem statement, proposing the solution, designing the codebase, conducting experiments, and preparing the original manuscript draft. Pramit Saha was responsible for the model merging components of the paper and preparing the manuscript draft. Yuki Asano contributed to technical discussions relating to self-supervised learning. Netzahualcoyotl Hernandez-Cruz managed the project data, and Olga Patey assisted with data collection. Aris T. Papageorghiou provided clinical supervision throughout the project. J. Alison Noble conceived the study objectives, secured funding, and supervised the entire project. All authors reviewed and approved the final version of the manuscript.

Abstract

Congenital Heart Disease (CHD) is one of the leading causes of fetal mortality, yet the scarcity of labeled CHD data and strict privacy regulations surrounding fetal ultrasound (US) imaging present significant challenges for the development of deep learning-based models for CHD detection. Centralised collection of large real-world datasets for rare conditions, such as CHD, from large populations requires significant co-ordination and resource. In addition, data governance rules increasingly prevent data sharing between sites. To address these challenges, we introduce, for the first time, a novel privacy-preserving, zero-shot CHD detection framework that formulates CHD detection as a normality modeling problem integrated with model merging. In our framework dubbed Sparse Tube Ultrasound Distillation (STUD), each hospital site first trains a sparse video tube-based self-supervised video anomaly detection (VAD) model on normal fetal heart US clips with self-distillation loss. This enables site-specific models to independently learn the distribution of healthy cases. To aggregate knowledge across the decentralized models while maintaining privacy, we propose a Divergence Vector-Guided Model Merging approach, DivMerge, that combines site-specific models into a single VAD model without data exchange. Our approach preserves domain-agnostic rich spatio-temporal representations, ensuring generalization to unseen CHD cases. We evaluated our approach on real-world fetal US data collected from 5 hospital sites. Our merged model outperformed site-specific models by 23.77% and 30.13% in accuracy and F1-score respectively on external test sets.

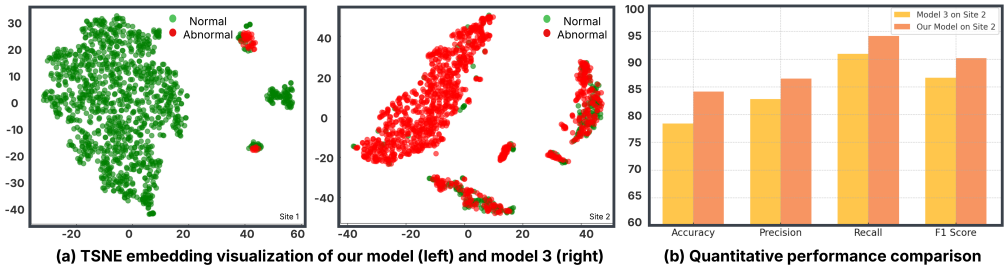


Figure 8.1: (a) t-SNE visualization (left) shows that our proposed method (after merging models trained on 3 sites) achieves nearly distinct clustering, suggesting well-separated feature representations. On the other hand, Model 3 (trained on Site 3) is observed to achieve low separability (right). (b) The quantitative comparison of both models evaluated on Site 2 further illustrates the benefit of our proposed model merging technique.

8.1 Introduction and Background

Congenital heart disease (CHD) accounts for approximately 28% of all congenital anomalies worldwide [224]. CHD encompasses a diverse range of heart conditions, varying in frequency and severity, and can be diagnosed early by fetal ultrasound (US) scanning. As a widely used non-invasive screening tool, fetal US is favored for its rapid data acquisition, affordability, portability, and ability to perform assessments without ionizing radiation. Early detection of CHD from fetal US is crucial to ensure long-term health outcomes [225–227]. However, diagnosing CHD remains challenging and time-intensive due to the subtle nature of certain heart defects and variable fetal US video quality. Further, fetal heart assessment [228] presents significant challenges due to several factors, including fetal movement, rapid heart rate, small size, and limited accessibility. This suggests a clinical need for approaches to automated CHD detection, and an opportunity for deep learning-based analysis. However, applying supervised deep learning methods is often impractical, as they are not designed for highly imbalanced data scenarios such as our clinical setting. Many forms of CHD are extremely rare, resulting in highly imbalanced datasets. Conversely, a large volume of fetal ultrasound videos from healthy fetuses is routinely collected during standard screening procedures. In this work, we exploit the availability of healthy population video to train a novel anomaly detection framework to identify CHD cases during inference.

Anomaly detection models in the literature, such as [229–232], offer promising solutions for normality modeling. For effective performance, such models need to

be trained on diverse, centralized data sourced from multiple hospital sites, thereby capturing a broad spectrum of fetal cardiac variations in appearance, geometry and disease. However, privacy regulations prohibit cross-hospital data sharing [233], creating a significant bottleneck in development of centralized models. To this end, instead of combining data from different sites, we propose to effectively merge models trained at individual sites. Our approach allows the aggregation of knowledge learned from individual hospital datasets without the need for data sharing and while avoiding interference and conflicts due to domain shifts. This ensures that essential task information is preserved, leading to a merged model that leverages the strengths of each local model while maintaining privacy compliance. The primary contributions of this work are as follows:

1. To the best of our knowledge, this is the first work to introduce **video normality learning for CHD detection in fetal US videos**. We train a self-supervised video anatomy detection network on healthy fetal US clips using self-distillation loss that incorporates a student-teacher model **(section 8.2.1)**. Our novel **Sparse Tube Ultrasound Distillation (STUD)** model learns the spatio-temporal representation related to healthy fetal hearts and is leveraged to detect previously unseen CHD anomalies during test time. Our model is light-weight as we sparsely sample 3D space-time tubes of varying sizes from the US video to create learnable tokens, which are then processed by a vision transformer. This enables us to develop strong and computationally efficient video models.

2. This is also the first work to investigate **model merging for multi-site US analysis as well as for normality modeling**. We propose a two-step model merging procedure dubbed DiVMerge to enhance robustness with respect to model noise and model drifts while preserving normality information **(section 8.2.2)**. We first compute the geometric median of local models, acting as a denoising mechanism and then compute the divergence vectors as the difference between individual models and the geometric median. The parameters with small divergence vector component are retained, while others are replaced by the geometric median. In addition, the overall magnitude of the divergence vector for each site model is used to dynamically weight the updated local models before merging.

3. Our method enables **zero-shot CHD detection** coupling of the normality model and k-Nearest Neighbours (KNN) algorithm, thereby eliminating the need for additional fine-tuning. Trained on healthy data from 3 hospitals in a privacy-preserving manner, our normality model demonstrates the ability to detect anomalies [126] (such as Hypoplastic Left Heart Syndrome (HLHS), Coarctation of the Aorta (COA), Right Aortic Arch (RAA), Left Superior Vena Cava (LSVC), Ventricular Septal Defects (VSD), and Cardiomegaly (CM)). In particular, DiVMerge outperforms the centralized model and all individual site-specific models on 2 external hospital datasets with distinct domain shifts.

8.2 Methodology

8.2.1 Site-specific Self-supervised Video Anomaly Detection

Ultrasound videos are inherently fine-grained and require dense temporal sampling to capture subtle changes essential for accurate understanding and detection. However, conventional tokenization methods using 2D patching [234] or fixed 3D kernels [107] generate an excessive number of tokens, making dense sampling computationally expensive and reducing the number of frames that can be processed on limited computing. To address this, our Sparse Tube Ultrasound Distillation (STUD) network employs a sparse tube sampling [235] that drastically reduces token redundancy while preserving spatio-temporal detail. For self-supervised video normality modeling, we integrate a video-focused self-distillation loss inspired by DINO [236], which trains a teacher-student network to learn consistent feature representations across diverse augmented views.

Sparse Tube Construction and Feature Extraction We adopt a sparse sampling strategy to address the limitations of dense tokenization. A standard 2D convolution with a 16×16 kernel is applied on frames sampled with a large temporal stride (e.g., every 16th frame). The total number of tokens generated from a video clip of dimensions $T \times H \times W$ is defined by: $N_{\text{tokens}} = \frac{T}{s_T} \times \frac{H}{s_H} \times \frac{W}{s_W}$ where s_T, s_H, and s_W denote the temporal and spatial strides, respectively. We adapt two primary types of tubes: **(a) Image Tubes** of shape $1 \times 16 \times 16 \times d$ (where d is hidden dimension), which tokenize individual frames and **(b) Video**

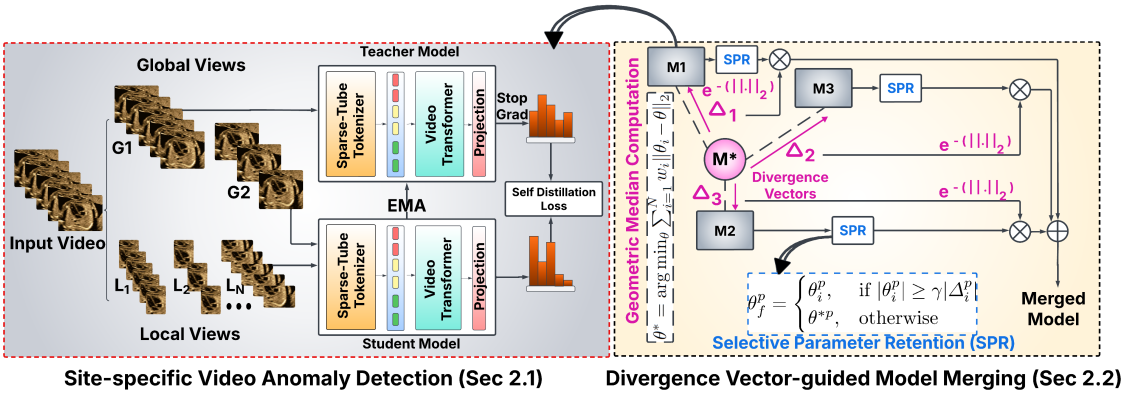


Figure 8.2: Overview of the proposed technique. Left figure shows self-supervised video anomaly network training at each site leveraging sparse-tube tokenizer and teacher-student model via self-distillation loss. This leads to the development of models M_1, M_2, and M_3 at three sites. The right figure shows the geometric median computation followed by estimation of divergence vectors for each site. The divergence vectors are then employed for selective parameter retention to reduce model drift and for adaptively weighting different models for final model merging.

Tubes of shape $8 \times 8 \times 8 \times d$, which capture the spatio-temporal context over multiple frames. Both types of tubes use a stride of $16 \times 16 \times 16$. To further capture diverse motion patterns in ultrasound videos, we incorporate variations such as temporally elongated tubes ($16 \times 4 \times 4$) for long-duration actions and spatially focused tubes ($2 \times 16 \times 16$) for fine spatial detail (see Fig. 8.2). A space-to-depth transformation is applied reduce the channel dimension of the feature map (by a factor of 2), effectively enlarging the receptive field without increasing the number of parameters. Additionally, we learn a single 3D kernel ($8 \times 8 \times 8$) reshaped by trilinear interpolation to adapt to various tube configurations ($4 \times 16 \times 16$ or $32 \times 4 \times 4$). These enhancements ensure that our sparse sampling method captures all the necessary details while reducing computational demands.

Self-Supervised Learning via Self-Distillation in the Video Domain We generate multiple spatio-temporal augmentations (each called a 'view') to capture global context and fine-grained local details. Specifically, we create two global views encompassing a large portion of the video's temporal and spatial dimensions and eight local views focusing on smaller, more detailed regions (see Fig. 8.2). In our self-distillation framework, the teacher network processes only the global views to produce target feature representations, while the student network processes both global and local views. The self-distillation loss encourages the student

representations to align with the teacher by minimizing the discrepancy between their outputs. To ensure stable learning, the teacher model parameters are updated by an exponential moving average (EMA) of the student model.

8.2.2 Divergence Vector-guided Model Merging (DiVMerge)

We propose a two-step model merging procedure designed to improve robustness to model drifts and interference while preserving essential normality information.

Step 1: Geometric Median Computation: The first step involves computing the geometric median of the locally trained models from individual sites (see Fig. 8.2). Given a set of N locally trained models $\mathcal{M} = \{M_1, M_2, \ldots, M_N\}$, each represented by its parameter set $\theta_i \in \mathbb{R}^d$, the geometric median is the point θ^* that minimizes the sum of Euclidean distances to all individual models: $\theta^* = \arg\min_\theta \sum_{i=1}^{N} w_i \|\theta_i - \theta\|_2$ where θ_i represents the parameter vector of the i-th local model, w_i is the weighting factor for each model (uniform in our case). This method is particularly beneficial in our distributed settings, where model updates may vary significantly across clients due to the natural heterogeneity in data distributions. The geometric median filters out inconsistencies, outliers, and extreme deviations in model updates that may arise due to small or biased datasets, noisy labels, or domain shifts in individual hospitals.

Step 2: Divergence Vector-Based Adaptation: We define the divergence vector for each site model as the difference between the local model trained on that site and the geometric median model. For each model M_i, the divergence vector Δ_i is computed as: $\Delta_i = \theta_i - \theta^*$ where θ_i represents the parameter vector of the locally trained model at site i, θ^* is the geometric median computed in Step 1, and Δ_i represents the site-specific deviation from the geometric median. This vector captures how much each model deviates from the robust median representation and serves two key purposes: **(A) Dynamic Model Weighting:** It adaptively assigns importance to the model contributions of each site in the final weighted averaging step. A model with a small divergence vector is likely to be more stable and reliable, whereas a model with a large divergence vector may reflect domain-specific biases or noise. We use the magnitude of the divergence vector to assign dynamic weights α_i to each site model, allowing models to contribute proportionally based on their

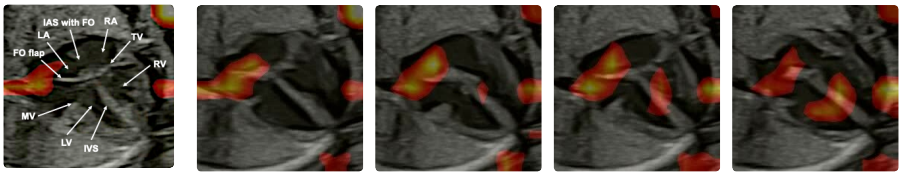


Figure 8.3: Feature map visualization overlaid on sequential US frames, highlighting the model's capability to focus on key anatomical fetal heart structures for CHD

distance from the geometric median where $\alpha_i = \exp(-\lambda\|\Delta_i\|_2)$. λ is a scaling factor that controls the influence of the magnitude of the divergence vector, $\|.\|_2$ is the L2-norm. The final normalized weight for model i is: $\tilde{\alpha}_i = \frac{\alpha_i}{\sum_{j=1}^{N}\alpha_j}$. The final merged model θ_f is computed as: $\theta_f = \sum_{i=1}^{N}\tilde{\alpha}_i\theta_i$. This adaptive weighting reduces the effect of noisy local models.

(B) Selective Parameter Retention: The divergence vector is also utilized to localize normality information at the parameter level. Each parameter θ^p is retained only if its weight magnitude exceeds a $\gamma-$rescaled magnitude of the divergence vector, as it likely represents a confident and stable feature. Otherwise, it is replaced by the corresponding geometric median parameter to prevent excessive deviation from the global consensus. $\theta_f^p = \begin{cases} \theta_i^p, & \text{if } |\theta_i^p| \geq \gamma|\Delta_i^p| \\ \theta^{*p}, & \text{otherwise.} \end{cases}$

8.3 Experiments and Results

Dataset and Experimental Settings The five datasets used in this work are fetal heart ultrasound video sweeps collected as part of an international collaboration involving five hospital sites hereafter referred to as Site 1 through Site 5. Data were acquired using 10 different ultrasound machines by sonographers and fetal cardiologists. Healthy fetal US videos from Sites 1–3 (8,878, 16,074, and 1,573, respectively) were used to train the model. The evaluation set comprised a mix of normal and abnormal cases (667, 2,088, and 667, in total respectively), with the abnormal cases comprising instances of COA and HLHS. Videos from Sites 4 and 5 (29 and 18 samples) were reserved for zero-shot testing and comprised normal cases along with four other anomalies, *viz.*, RAA, LSVC, VSD, and CM. All videos (mean length: 125 frames) were pre-processed with an automatic cropping model to extract the heart region.

Training and Implementation Details During training, we randomly sampled a clip of 64 frames per video with a sampling rate of 3, while during evaluation we uniformly sampled N clips from each video to extract features. A KNN classifier was then applied to these clip features, classifying a video as an anomaly if any clip was flagged abnormal. For self-distillation, we used 2 global crops (size 224) and 8 local crops (size 96), applying spatial transforms: color jittering, solarization, Gaussian blur and varying temporal sampling rates for both crop types. All models were trained for 200 epochs on a RTX6000 GPU (VRAM 25GB) with a batch size of 12 and a cosine learning rate schedule with a 5e-04 initial learning rate. Scaling factor λ was fixed at 0.005 via grid search.

8.3.1 Performance analysis of site-specific Video Anomaly Detection

Table 8.1 shows the comparison of our model (trained and tested individually at each site) with two baseline methods *viz.*, TimeFormer [234] (w/ supervised pre-training) and VideoMAE [107] (w/ self-supervised pre-training) in detecting normal and abnormal fetal heart clips. Our model provides a favorable trade-off between computational efficiency and predictive performance. TimeFormer achieves slightly higher accuracy for some sites at the cost of 10x more tokens. VideoMAE underperforms compared to our method and TimeFormer for all internal sites. The highest overall F1-score of our model shows its best generalization capability and most stable performance across different sites while reducing computational costs. Figure 8.3 shows a visualization of the attention maps from the final layer. This demonstrates effective localization of our model on sequential US video frames around the tricuspid valve, foramen ovale flap, and inter-ventricular septum which are key anatomical structures for CHD detection.

Table 8.1: Performance of site-specific models for CHD detection in sites 1-3

Model	# tokens	Site 1		Site 2		Site 3		Average	
		Prec	**F1**	**Prec**	**F1**	**Prec**	**F1**	**Prec**	**F1**
TimeFormer	12522 (100%)	40.00	57.14	81.13	88.61	**89.37**	**93.85**	70.17	79.87
VideoMAE	6272 (50%)	43.48	46.51	69.95	78.62	88.13	93.09	67.19	72.74
Ours	**1176 (9.39%)**	**48.72**	**64.41**	**87.13**	**90.97**	83.84	90.14	**73.66**	**82.20**

8.3.2 Performance analysis of Model Merging

Evaluation on sites 1-3

Table 8.2 shows the performance (precision and F-1 score) of the model built using DiVMerge on sites 1, 2, and 3, compared to the centralized model (*i.e.*, trained on data combined from all sites) and individual models (*i.e.* models trained locally on their own sites). We also compare with 5 SOTA model merging methods, *viz.*, Model Soup (2022)[237], Task vector (2023) [238], Ties Merging (2023) [239], DARE (2024) [240], Model Stock (2024) [241]. Note that all SOTA methods other than Model Soup need a base model for task vector computation whereas our merging strategy does not require one. This enhances the usability and potential scope of application of our model, including for scenarios where a base model is unavailable. We observe that divergence-guided merging of models (ours) is almost as good as or better than the model built from centralized data. In addition, the overall performance of our model is higher than that of individual local models. While performance slightly drops with respect to individual models for Site 2 (which has the highest amount of data), our model outperforms individual models trained on Sites 1 and 3, which have less data. The improvement in precision is 3.83% for Site 3 and 4.06% for Site 1. This reveals a benefit of model merging for sites with limited data availability. In addition, this shows that our merged model eliminates the need to store site-specific models. While Model Soup and other SOTA merging techniques have a slight performance drop due to model drifts induced by heterogeneous sites, our model shows overall stable performance. This demonstrates that the use of the divergence vector can effectively mitigate the inter-site model conflicts.

Evaluation on sites 4 and 5

Sites 4 and 5 are different geographical locations and cover different patient demographics to those used to train the merged model. Evaluation of our merge model on sites 4 and 5 data is shown in Tab. 8.3 and Fig. 8.4. These show that our model generalises well to these new data scenarios where there are domain shifts due to different ultrasound scanners and data acquisition procedures at these sites. The centralized model and Model 1 fails to detect most abnormal cases resulting in an overall F1 score of 23.55 and 20.0 respectively. By contrast,

Table 8.2: Performance comparison of model merging for sites 1-3. (B.M = Base Model)

Model	B.M. needed?	Site 1		Site 2		Site 3		Average	
		Prec	F1	Prec	F1	Prec	F1	Prec	F1
Centralized	-	50.00	65.52	87.30	91.20	89.45	93.98	75.58	83.57
Individual	-	48.72	64.41	87.13	90.97	83.84	90.14	73.23	81.84
Model Soup [237]	No	41.30	57.58	**86.75**	89.79	86.34	92.45	71.74	79.94
Task Vector [238]	Yes	35.19	51.35	76.86	88.53	88.52	93.54	66.85	77.00
Ties Merging [239]	Yes	32.60	44.44	79.45	85.98	89.42	93.50	67.16	74.64
DARE [240]	Yes	32.60	44.40	83.00	86.80	88.40	93.60	68.00	74.93
Model Stock [241]	Yes	36.70	52.20	84.27	88.11	**89.68**	**94.18**	70.21	78.16
Ours ($\gamma = 0.01$)	No	**52.78**	**67.86**	86.50	**90.18**	87.67	92.31	75.65	83.45
Ours ($\gamma = 0.1$)	No	**52.78**	**67.86**	86.21	89.64	87.67	**94.18**	**76.22**	**83.89**

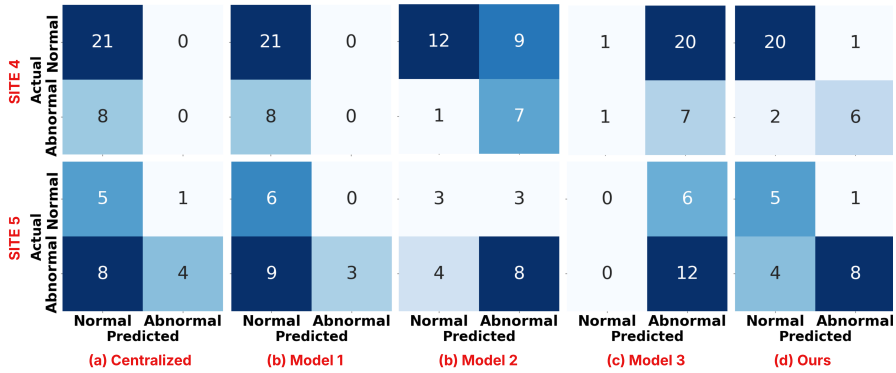


Figure 8.4: Confusion matrices illustrating the performance of various models on external sites 4 and 5. The results indicate that the Centralized Model and Model 1 struggle to detect most abnormal cases, while Model 3 frequently misclassifies normal cases as abnormal due to domain shift. In contrast, our model achieves the best performance, accurately distinguishing most normal and abnormal cases even with domain gap.

Model 3 considers almost all abnormal cases as healthy hearts. Model 2 performs better but misidentifies 15 out of 20 healthy samples as CHD. Our model achieves the best performance, improving on the centralized model by 20% and 54.6% in accuracy and F1 score respectively.

Table 8.3: Performance comparison of merged model w.r.t. baselines on Sites 4 and 5

Model	Site 4				Site 5				Average			
	Accuracy	Prec	Recall	F1	Accuracy	Prec	Recall	F1	Accuracy	Prec	Recall	F1
Centralized	72.0	0.0	0.0	0.0	50.0	80.0	33.3	47.1	61.0	40.0	16.7	23.5
Model 1	72.4	0.0	0.0	0.0	50.0	100.0	25.0	40.0	61.2	50.0	12.5	20.0
Model 2	65.5	43.8	87.5	58.3	61.1	72.7	66.7	69.6	63.3	58.3	77.1	63.9
Model 3	27.6	25.9	87.5	40.0	66.7	66.7	100.0	80.0	47.2	46.3	93.8	60.0
Ours	**89.7**	**85.7**	**75.0**	**80.0**	**72.2**	**88.9**	**66.7**	**76.2**	**81.0**	**87.3**	**70.9**	**78.1**

8.4 Conclusion

In this work, we have introduced a novel privacy-preserving, zero-shot CHD detection framework that leverages self-supervised video normality learning (STUD) and a novel divergence vector-guided model merging (DiVMerge) technique to overcome the challenges of scarce labeled data and cross-hospital privacy constraints in fetal US. Evaluation with real-world fetal US datasets from five hospitals demonstrates that our method achieves superior performance than SOTA methods in detecting CHD anomalies. Compared with a model built with centralized data, our merged model achieves a 20% improvement in accuracy and a 54.6% increase in F1-score on unseen test data sets, highlighting its strong zero-shot generalization capabilities even under domain shifts caused by variations in ultrasound scanners, acquisition procedures, and patient demographics.

Our results highlight that model merging can serve as a viable alternative to centralized learning in privacy-sensitive clinical settings. Our experiments show that model merging has the potential to improve performance particularly for sites with limited data through the knowledge acquired from models trained on other sites via merging. Another notable observation is that our merged model significantly outperforms model trained with centralized data by mitigating domain shifts, thanks to its adaptive weighting of site-specific models and the resolution of inter-site model conflicts via the proposed divergence vector.

9

Self-supervised Learning of Echocardiographic Video Representations via Online Cluster Distillation

Authors: Divyanshu Mishra, Mohammadreza Salehi, Pramit Saha, Olga Patey, Aris T. Papageorghiou, Yuki M. Asano, J. Alison Noble

Published in Conference: The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025)

Background: Self-supervised learning (SSL) has delivered transformative advances in natural image and video representation learning by reducing reliance on manual labeling. However, translating these successes to cardiac ultrasound video remains a substantial challenge, due to the subtlety of cardiac anatomical features, the complex temporal dynamics of heart motion, and frequent low signal-to-noise ratio in ultrasound acquisitions. Conventional SSL approaches—such as contrastive, masked modeling, or clustering-based methods—often falter in echocardiography because they struggle to distinguish between highly similar samples, and their augmentations can inadvertently obscure or distort critical clinical patterns. Furthermore, the lack of large, domain-specific pre-trained models limits their ability to capture fine-grained anatomy and meaningful temporal structure crucial to cardiac assessment. In this chapter, we address these limitations by introducing DISCOVR, a dual-branch self-supervised framework that specifically integrates semantic knowledge

from static images with temporal modeling of video dynamics, aiming to produce richer, anatomically meaningful representations for challenging cardiac ultrasound data. Extensive evaluation across six echocardiography datasets demonstrates that DISCOVR outperforms domain-specific video anomaly detection models and state-of-the-art SSL baselines on zero-shot classification, linear probing, and segmentation transfer tasks, establishing DISCOVR as an effective approach for self-supervised learning in complex medical video domains.

Author Contribution: I was the lead technical author of the paper, responsible for formulating the problem statement, proposing the solution, designing the codebase, conducting the experiments, and preparing the original manuscript draft. Mohammadreza Salehi contributed in overall paper discussion. Pramit Saha contributed to paper review and dataset curation for open-source datasets. Olga Patey helped in data collection and annotation. Yuki Asano participated in technical discussions of the paper. Aris T. Papageorghiou provided clinical supervision throughout the project. J. Alison Noble conceived the overall objectives of the study, secured funding, and supervised the entire project. All authors reviewed and approved the final version of the manuscript.

Abstract

Self-supervised learning (SSL) has achieved major advances in natural images and video understanding, but challenges remain in domains like echocardiography (heart ultrasound) due to subtle anatomical structures, complex temporal dynamics, and the current lack of domain-specific pre-trained models. Existing SSL approaches such as contrastive, masked modeling, and clustering-based methods struggle with high intersample similarity, sensitivity to low PSNR inputs common in ultrasound, or aggressive augmentations that distort clinically relevant features. We present DISCOVR (Distilled Image Supervision for Cross Modal Video Representation), a self-supervised dual branch framework for cardiac ultrasound video representation learning. DISCOVR combines a clustering-based video encoder that models temporal dynamics with an online image encoder that extracts fine-grained spatial semantics. These branches are connected through a semantic cluster distillation loss that transfers anatomical knowledge from the evolving image encoder to the video encoder, enabling temporally coherent representations enriched with fine-grained semantic understanding. Evaluated on six echocardiography datasets spanning fetal, pediatric, and adult populations, DISCOVR outperforms both specialized video anomaly detection methods and state-of-the-art video-SSL baselines in zero-shot and linear probing setups, and achieves superior segmentation transfer.
Code available at: https://github.com/mdivyanshu97/DISCOVR

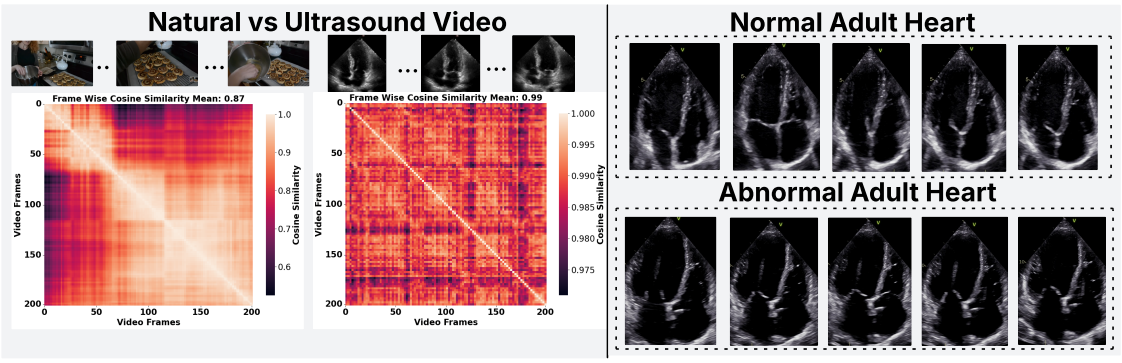


Figure 9.1: Figure (left) compares two fine-grained videos: a natural scene of a person baking (left) and an adult fetal heart ultrasound (right). The frame-level cosine similarity matrix, computed using a pretrained VideoMAE model, shows that ultrasound frames are highly similar (mean=0.99), with only minor local variations. This highlights the difficulty in distinguishing individual frames in such medical videos. Figure (right) compares normal and abnormal adult echocardiograms that appear nearly identical. However, on close inspection, it is revealed that the abnormal heart shows severe biventricular systolic dysfunction and a dilated, globular left ventricle, underscoring the subtlety of cardiac defects and the need for fine-grained structural analysis.

9.1 Introduction

Modeling dynamic content in video data presents significant challenges due to complex spatio-temporal relationships, high redundancy between frames, and the need to capture both short- and long-range temporal dependencies [10, 242]. Echocardiography (heart or cardiac ultrasound) exemplifies these video understanding challenges [7, 8]. With high frame rates (30–80 fps) [9], complex anatomical motion, and variability in image appearance caused by speckle, shadowing artifacts, and ultrasound probe variability [243], automated echocardiography analysis requires sophisticated temporal modeling approaches. The information density in these videos is high, where features critical for diagnosis may appear as subtle variations in wall motion, valve function, or blood flow patterns that manifest only when viewed dynamically across multiple frames. Moreover, the appearance of the heart can change drastically across different cardiac views, patient populations, and imaging equipment. Developing robust video SSL models for comprehensive video understanding in echocardiography faces additional obstacles due to data limitations. Expert annotations are costly, labor-intensive, and if based on real-world hospital data often incomplete, capturing only specific aspects of the rich information contained in these videos. This scarcity of labeled data motivates

SSL approaches that can leverage abundant unlabeled echocardiograms for model development [6, 10, 244].

Several SSL frameworks have been proposed for learning meaningful video representations, each with particular limitations in the echocardiography context. Masked video modeling methods [107, 245, 246] tend to focus on reconstructing low-level image features like textures or edges, limiting their ability to capture high-level semantic information critical for clinical interpretation. This is especially problematic for ultrasound, which inherently exhibits a low signal-to-noise ratio (SNR), making approaches that rely on low-level pixel representations ineffective. Contrastive learning methods [247, 248] struggle due to high inter-sample similarity and limited effective augmentations, making it difficult to construct informative positive and negative pairs, often leading to representation collapse. Clustering-based SSL methods have demonstrated strong semantic learning through self-distillation but rely heavily on aggressive augmentations that risk disrupting essential anatomical details required for fine-grained understanding.

To address these limitations, we propose DISCOVR (*Distilled Image Supervision for Cross-Modal Video Representation*), a dual branch SSL framework tailored for echocardiography that jointly captures temporal dynamics and fine-grained semantic structure. The video encoder is trained to model temporal features using a clustering-based objective applied to masked video tokens, while an online image encoder separately learns spatially rich and anatomically meaningful representations from masked image views. To bridge the gap between spatial and temporal learning, we introduce a semantic cluster distillation loss that transfers knowledge from the evolving image encoder to the video encoder through semantic cluster alignment. This enables the video encoder to embed fine-grained semantic detail into its temporally coherent representations, without relying on pretrained models or heavy augmentations.

We extensively evaluate DISCOVR on six echocardiography datasets that span fetal, pediatric, and adult populations, covering anomaly detection, classification (linear probing and zero-shot transfer), and segmentation tasks. DISCOVR consistently outperforms prior self-supervised and anomaly detection methods. It achieves an average F1 improvement of 3.4% for anomaly detection, a 2.4% gain in linear

probing, and a 1.5% increase in balanced accuracy under zero shot evaluation. For segmentation, DISCOVR delivers a 3.1% relative improvement in Dice score (from 81.9 to 84.4), despite using a simple segmentation head compared to more complex baseline architectures. These results demonstrate that integrating spatial semantics with temporal dynamics through cross-modal distillation yields robust and generalizable cardiac ultrasound video representations.
Overall, our contibutions are as follows:

- We develop an SSL method that jointly models temporal dynamics and spatial semantics by integrating video self-distillation with an evolving semantic image encoder, without labels, pretrained models, or augmentations.
- We introduce a novel online semantic distillation loss that continually transfers anatomical knowledge from the evolving image encoder to the video encoder, enriching its temporal representations with fine-grained spatial semantics to better capture clinically relevant spatio-temporal patterns in echocardiography.
- DISCOVR is, to our knowledge, the most comprehensive self-supervised video representation model for echocardiography to date. Trained solely on normal videos, it models healthy heart dynamics and detects pathology as deviations, eliminating the need for labeled abnormal cases. Evaluated across six datasets spanning fetal, pediatric, and adult cohorts, DISCOVR demonstrates strong generalization in zero shot classification, linear probing, anomaly detection, and segmentation, making it a versatile backbone for ultrasound analysis.

9.2 Related Work

Self-supervised learning (SSL) aims to learn feature extractors directly from raw data by solving an intrinsic task using supervision signals derived from the data itself, eliminating the need for manual labels. Early image-based SSL relied on handcrafted pretext tasks such as solving jigsaw puzzles [249], predicting rotations [250], or colorizing grayscale inputs [251]. Recent methods have shifted towards instance discrimination via contrastive learning [247, 252, 253]. To

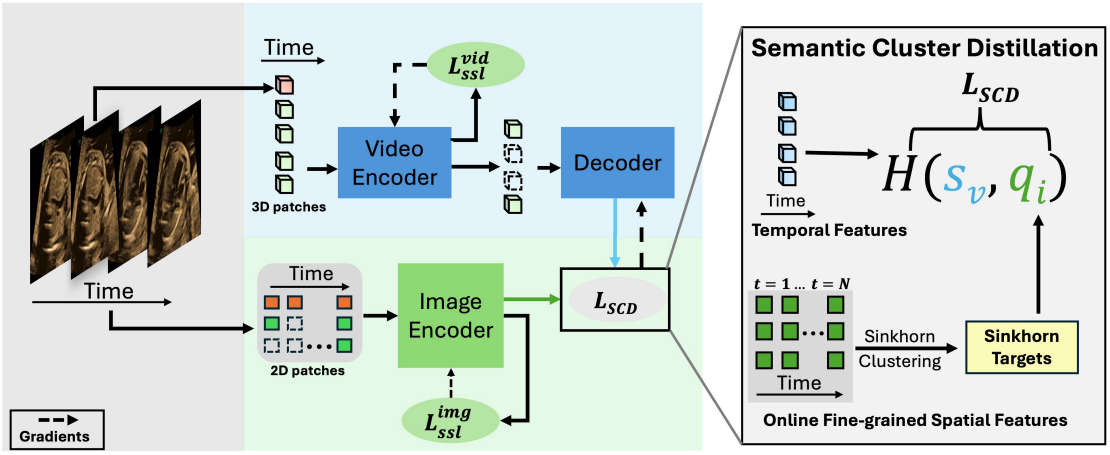


Figure 9.2: Overview of the DISCOVR framework. An input video is tokenized into 3D patches for the video branch and per-frame 2D patches for the image branch. Both encoders perform masked self-distillation. Masked video tokens are reconstructed by the video decoder, and dense semantic features are extracted from the image encoder. The $\mathcal{L}_{\text{SCD}}$ loss then aligns these outputs, distilling fine-grained spatial semantics into the video representation to produce rich spatio-temporal features.

understand how these ideas extend to video and medical domains, we review the most relevant self-supervised methods in both areas, highlighting shared limitations and how DISCOVR addresses them.

Video Self-Supervised Learning. Extending SSL to video introduces additional temporal complexity, inspiring tasks such as frame order prediction [254, 255], spatio-temporal jigsaws [256], and playback pace prediction [103, 257]. Recently, masked video modeling has become the dominant approach: VideoMAE [107] reconstructs raw pixels from masked tubelets using a ViT backbone. MGMAE [110] predicts optical flow to enhance temporal modeling, and motion-aware masking [246] highlights dynamic regions. SIGMA [258] replaces pixel-level targets with Sinkhorn-regularized cluster assignments, encouraging learning of semantic features. Yet, these approaches often rely on frozen teachers, handcrafted objectives, or sensitive clustering parameters. DISCOVR addresses these issues by introducing video self-distillation with evolving semantic guidance from an image encoder, aligning fine-grained spatial and temporal features to produce coherent, high-level video representations, without external supervision, handcrafted tasks, or modality-specific assumptions.

Self-Supervised Pretraining for Medical Videos. Given the limited availability of annotated data, several works have adapted video SSL techniques

to medical domains. Jiao et al. [259] explored frame order and transformation prediction for fetal ultrasound. EchoFlow[260] generated synthetic echocardiograms via adversarial VAEs and latent flow. Although effective in context, these methods inherit key limitations from natural video SSL, including reliance on frozen teachers, hand-crafted objectives, and sensitive clustering parameters. In addition, they adopt design choices tailored to natural images, such as short clip lengths and the lack of mechanisms for capturing fine-grained spatial cues, both of which are inadequate for clinical video analysis, where longer temporal context and detailed spatial reasoning are critical. In contrast, DISCOVR uses long (64-frame) clips and introduces dynamic semantic guidance from an evolving image encoder, enabling the video backbone to learn rich, fine-grained spatio-temporal representations without reliance on pretrained models or handcrafted supervision.

9.3 Methodology

The modelling of echocardiography video-based tasks poses unique challenges, as models must simultaneously detect fine-grained anatomical details, such as subtle septal defects, and accurately track how these features evolve throughout the cardiac cycle to reliably identify anomalies. We propose a unified self-supervised framework addressing these aspects without relying on labeled data or external pretrained models. Our method integrates three complementary techniques: (1) video self-distillation to capture global cardiac motion, (2) online spatial guidance to learn fine-grained structural information, and (3) semantic cluster distillation (SCD) loss to transfer fine-grained semantic knowledge from the evolving image encoder to the video model.

9.3.1 Video Self-Distillation

To capture how cardiac structures evolve throughout the cardiac cycle, it is essential to learn spatio-temporal representations from echocardiography videos. We propose a video-level self-distillation framework based on a student-teacher architecture with Vision Transformer (ViT)-based encoders (Fig. 9.2) that models temporal dynamics and improves understanding of global heart motion. Given a video

input v, we partition it into non-overlapping 3D space-time patches (tube tokens), and prepend a learnable class (CLS) token, resulting in a sequence $x_0, x_1, \ldots, x_N$, where x_0 is the CLS token.

The teacher encoder E_{θ_t} processes the complete, unmasked video to produce a global representation, whereas the student encoder E_{θ_s} processes multiple masked variants $v_{\mathcal{M}1}, \ldots, v_{\mathcal{M}_M}$, each applying distinct random space-time masks to enforce inference of missing content.

Both encoders output a global video representation via the CLS token:

$$z_t = E_{\theta_t}(v)[0], \quad z_s^{(m)} = E_{\theta_s}(v_{\mathcal{M}_m})[0]. \tag{9.1}$$

The teacher parameters are updated using an exponential moving average (EMA) of the student parameters:

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s, \quad \lambda \in [0, 1). \tag{9.2}$$

These CLS embeddings are subsequently mapped through linear projection heads characterized by learnable weight matrices W_t (teacher) and W_s (student). The resulting embeddings are transformed into probability distributions via temperature-scaled softmax operations:

$$P_t = \text{softmax}\left(\frac{W_t z_t}{\tau_t}\right), \quad P_s^{(m)} = \text{softmax}\left(\frac{W_s z_s^{(m)}}{\tau_s}\right), \tag{9.3}$$

where τ_t and τ_s are temperature parameters for the teacher and student, respectively.

We align these probability distributions using the cross-entropy loss:

$$\mathcal{L}_{\text{ssl}}^{vid} = \frac{1}{M}\sum_{m=1}^{M} H(P_t, P_s^{(m)}), \tag{9.4}$$

where H denotes cross-entropy. This approach encourages the student to match the teacher's global representation of cardiac motion, despite observing only incomplete views of the video. Through video-level self-distillation, the student learns to recover the evolving dynamics of anatomical landmarks, capturing coherent motion patterns and structural features relevant to global heart function throughout the cardiac cycle.

9.3.2 Fine-Grained Online Spatial Guidance

Although video self-distillation promotes temporal consistency and global abstraction, it tends to overlook fine-grained spatial features, particularly those critical to clinical interpretation in echocardiography. Echocardiography imaging captures the dynamics and appearance of anatomically complex structures, where capturing subtle spatial details, such as mitral valve leaflet motion, septal wall thickness, or endocardial border definition, is crucial. To address this, we introduce a two-part strategy for enriching spatial detail and semantic structure in video representations:
a). Masked Image Self-Distillation. An online image encoder is trained to learn spatially rich features from partially masked images, enabling the extraction of fine-grained semantic concepts.
b). Semantic Cluster Distillation (SCD). A cross-modal clustering objective aligns reconstructed video tokens with spatial image features, encouraging the video model to organize its representation space around semantically meaningful structures.

Masked Image Self-Distillation

To learn fine-grained semantic features, we train an image encoder $\mathcal{I}_\theta$ in parallel with the video encoder. Each video v is decomposed into individual frames $\{x_t\}$, which are processed independently. For each frame x, the teacher image encoder $\mathcal{I}_{\theta_t}$ receives the full-resolution image, while the student encoder $\mathcal{I}_{\theta_s}$ is given N randomly masked variants $\{x_{\mathcal{M}_i}\}_{i=1}^{N}$. Each output is projected using distinct learnable heads W_t (teacher) and W_s (student), followed by softmax normalization:

$$P_s^{(i)} = \text{softmax}\left(\frac{W_s \mathcal{I}_{\theta_s}(x_{\mathcal{M}_i})}{\tau_s}\right), \quad P_t = \text{softmax}\left(\frac{W_t \mathcal{I}_{\theta_t}(x)}{\tau_t}\right), \tag{9.5}$$

where τ_s and τ_t are temperature parameters. The loss function encourages the student to match the teacher's predictions across all masked views:

$$\mathcal{L}_{\text{ssl}}^{img} = \frac{1}{N}\sum_{i=1}^{N} H(P_t, P_s^{(i)}), \tag{9.6}$$

with $H(\cdot,\cdot)$ denoting the cross-entropy. This training objective promotes the emergence of spatially grounded representations that encode fine-grained clinical concepts such as fetal heart valves, ventricular anatomy, and septal delineation that may be underrepresented in purely temporal learning.

Semantic Cluster Distillation (SCD)

While Masked Image Self-Distillation enables the image encoder to learn spatially grounded representations that capture fine-grained clinical concepts, it does not transfer this knowledge to the video encoder. As a result, the spatial and temporal representations remain disjoint. To bridge this gap, we introduce *Semantic Cluster Distillation (SCD)*, a cross-modal objective that distills semantic structure from the image encoder, guiding the video encoder to incorporate fine-grained spatial detail into its token representations.

Given a masked video input, the student video encoder E_{θ_s} processes the visible tokens to produce latent representations, which are then passed to a decoder ψ that reconstructs token-level features $\hat{\mathbf{z}}_v \in \mathbb{R}^{B\times N\times D}$, where B is the batch size, N is the number of masked tokens, and D is the feature dimension. In parallel, the corresponding video frames are processed by the image encoder $\mathcal{I}_{\theta_t}$, producing spatial features $\hat{\mathbf{z}}_i \in \mathbb{R}^{B\times N\times D}$. These image features are detached from the gradient flow and serve as semantic targets. Both sets of features are projected onto a shared set of learnable prototypes $P \in \mathbb{R}^{K\times D}$, resulting in similarity scores:

$$\mathbf{s}_v = \frac{\hat{\mathbf{z}}_v P^\top}{\tau}, \quad \mathbf{s}_i = \frac{\hat{\mathbf{z}}_i P^\top}{\tau}, \tag{9.7}$$

where τ is a temperature scaling parameter and K is the number of prototypes. The resulting scores are transformed into Sinkhorn soft cluster targets using the Sinkhorn-Knopp algorithm:

$$\mathbf{q}_v = \text{Sinkhorn}(\mathbf{s}_v), \quad \mathbf{q}_i = \text{Sinkhorn}(\mathbf{s}_i). \tag{9.8}$$

The SCD loss symmetrically aligns the two modalities by minimizing the cross-entropy between their soft cluster assignments:

$$\mathcal{L}_{\text{SCD}} = \text{CE}(\mathbf{s}_v, \text{stopgrad}(\mathbf{q}_i)) + \text{CE}(\mathbf{s}_i, \text{stopgrad}(\mathbf{q}_v)), \tag{9.9}$$

where gradients are propagated only through the video model and the prototype matrix P, while the image encoder is updated solely via its own self-distillation loss. This guides the video encoder to anchor its token representations to the spatially grounded clusters discovered by the image encoder, thereby distilling fine-grained anatomical detail into its temporal feature space. Semantic Cluster Distillation thus embeds spatial semantics within temporal features, yielding spatio-temporal representations that capture anatomically relevant detail in echocardiography videos.

Table 9.1: Comparison of video anomaly detection methods on three echocardiography datasets. Our method consistently outperforms SOTA approaches, demonstrating improved effectiveness in identifying cardiac abnormalities across diverse patient populations.

Dataset	Model	Balanced Acc.	F1	AUC
EchoNet-Dynamic	MNAD	52.25	52.08	53.15
	MemAE	49.22	46.33	49.69
	C2FPL	57.36	57.35	59.00
	Ours	**63.20**	**61.45**	**67.06**
RVENET	MNAD	52.34	52.18	54.05
	MemAE	47.65	32.10	44.68
	C2FPL	47.88	47.86	46.30
	Ours	**56.23**	**53.88**	**57.42**
Echo Pediatric-LVH	MNAD	47.86	47.85	47.31
	MemAE	47.28	47.28	47.23
	C2FPL	51.39	51.31	50.68
	Ours	**55.63**	**54.63**	**57.23**

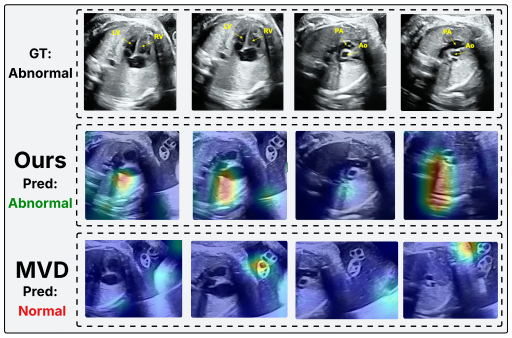


Figure 9.3: Zero-Shot classification comparison: (**Top**) The sweep from four-chamber to three-vessel view reveals smaller left-sided structures (LV and Ao) versus right-sided (RV and PA), consistent with coarctation of the aorta. (**Middle**) DISCOVR correctly identifies the abnormality, focusing on the ventricles in the four-chamber view and the Ao and PA in the vessel view. (**Bottom**)A backbone pretrained with MVD, in contrast, misclassifies the video as normal.

9.4 Experiments and Results

Datasets. We use five ultrasound video datasets across fetal, pediatric, and adult populations. Two private fetal heart datasets, FetalEcho1 and FetalEcho2, were each collected from different hospital partners in the UK, comprising 10-second transverse, cephalad sweeps capturing five standard cardiac views (Situs, 4CH, LVOT, 3VV, 3VT). FetalEcho1 includes 8273/414/317 and FetalEcho2 includes 4154/320/305 videos for training/validation/testing. For adult and pediatric echocardiography, we use 3 public datasets: EchoNet Dynamic (apical 4CH adult; 7378/1326/1326) [129], EchoPediatric LVH (parasternal long-axis pediatric; 7837/1592/1592) [130], and RVENet (right ventricular pediatric/adult; 2516/487/573) [131]. Videos for adult and pediatric populations are labeled as *normal* or *abnormal* based on ejection fraction (EF), with *abnormal* defined as EF $<$ 45% or EF $>$ 75% [261]. Fetal videos are labeled as *normal* or *abnormal* based on expert evaluation by two fetal cardiologists(+10 years of experience). For the downstream segmentation task, we utilize the CAMUS [262] dataset.

Evaluation. All baseline models use official implementations, with videos sampled in 64-frame clips at a stride of 3. We adopt space-time tube embeddings from VideoMAE [107], treating each $2 \times 16 \times 16$ cube as a token with 90% masking ratio. All models use a ViT base backbone with consistent configurations. We evaluate representations using **zero-shot classification** and **linear probing**. Zero-shot evaluation uses a weighted kNN classifier [236, 252] on frozen features, with k selected based on validation balanced accuracy. Linear probing trains a linear classifier for 30 epochs on a frozen backbone using a labeled validation set. During inference, each test video is divided into 64-frame clips and classified independently; a video is labeled abnormal if any clip is predicted abnormal. For segmentation evaluation, we add a linear layer followed by Conv2D upsampling blocks to generate pixel-level masks while keeping the backbone frozen.

Baselines. We compare DISCOVR with SOTA video SSL methods SIGMA [258], MGMAE [110], MVD [263], VideoMAE [107], and RAD-DINO [264], covering masked modeling, clustering, and dense feature learning. For anomaly detection, we include SOTA methods MNAD [265], MemAE [266], and C2FPL [267], which rely solely on spatial-temporal learning without external modules like object detectors, pose estimators, or optical flow, often tailored to natural images.

9.4.1 Comparison with Video Anomaly Detection Methods

Table 9.1 compares the anomaly detection performance of DISCOVR with several state-of-the-art approaches. DISCOVR achieves the highest F1 score for all datasets (61.45% for EchoNet Dynamic, 53.88% for RVENET, and 54.63% for EchoPediatric LVH) as well as the highest balanced accuracy (63.20%, 56.23%, and 55.63%, respectively), substantially outperforming C2FPL, MemAE, and MNAD for all reported metrics. C2FPL relies on a multi-stage pseudo-labeling process to enhance anomaly discrimination, while both MemAE and MNAD incorporate sophisticated memory mechanisms and feature regularization in their inference pipelines. These methods employ targeted, anomaly-specific inference strategies and complex architectures.

In contrast, DISCOVR builds on a simple self-supervised learning framework that jointly learns spatial and temporal features, utilizing only a straightforward zero shot

Table 9.2: Linear probing classification results on five echocardiography datasets spanning fetal, adult, and pediatric populations. Our method achieves SOTA results, outperforming prior video SSL baselines and generalizing effectively across diverse clinical cohorts.

Dataset	Model	Acc	Bal. Acc.	F1
Fetal-Echo 1	VideoMAE	60.19	60.01	59.82
	MGMAE	59.55	59.40	59.30
	SIGMA	63.11	62.93	62.78
	Ours	**65.70**	**65.52**	**65.39**
Fetal-Echo 2	VideoMAE	56.39	53.12	51.60
	MGMAE	60.98	60.49	60.43
	SIGMA	56.07	56.06	55.81
	Ours	**65.25**	**63.53**	**63.59**
Echonet-Dynamic	VideoMAE	71.04	70.86	70.85
	MGMAE	61.84	61.81	61.81
	SIGMA	75.57	75.48	75.50
	Ours	**77.68**	**77.61**	**77.63**
Echo Pediatric-LVH	VideoMAE	60.87	60.94	60.71
	MGMAE	54.71	51.70	49.46
	SIGMA	58.42	57.27	57.24
	Ours	**62.81**	**61.64**	**61.66**
RVENET	VideoMAE	60.03	60.31	59.70
	MGMAE	59.16	59.15	59.15
	SIGMA	59.51	59.25	58.98
	Ours	**62.65**	**62.68**	**62.65**

Table 9.3: Zero-shot evaluation across five echocardiography datasets covering fetal, adult, and pediatric populations. Our method consistently outperforms existing video SSL baselines, demonstrating robust generalization across diverse clinical populations.

Dataset	Population	Model	Acc	Bal. Acc.	F1
Fetal-Echo 1	Fetal	RAD-DINO	55.34	55.35	55.34
		VideoMAE	60.52	60.81	60.00
		SIGMA	54.37	54.91	51.90
		MGMAE	60.84	61.03	60.64
		MVD	59.87	60.20	59.15
		Ours	**62.46**	**62.79**	**61.79**
Fetal-Echo 2	Fetal	RAD-DINO	54.10	51.46	50.62
		VideoMAE	50.49	48.01	47.21
		SIGMA	55.41	51.90	49.92
		MGMAE	59.34	56.71	56.09
		MVD	59.34	55.45	53.14
		Ours	**59.67**	**57.18**	**56.69**
Echonet-Dynamic	Adult	RAD-DINO	59.43	59.63	59.34
		VideoMAE	57.16	57.91	55.07
		SIGMA	53.47	54.46	49.04
		MGMAE	51.21	52.23	46.13
		MVD	60.11	60.94	57.56
		Ours	**62.59**	**63.20**	**61.45**
Echo Pediatric-LVH	Pediatric	RAD-DINO	53.14	52.27	52.26
		VideoMAE	51.57	53.98	50.47
		SIGMA	47.55	49.56	46.80
		MGMAE	46.61	48.91	45.45
		MVD	49.56	51.91	48.46
		Ours	**54.65**	**55.63**	**54.63**
RVENET	Adult, Pediatric	RAD-DINO	55.67	55.65	**55.65**
		VideoMAE	54.97	55.64	52.24
		SIGMA	52.36	53.18	47.64
		MGMAE	53.23	54.08	48.17
		MVD	54.62	55.12	53.17
		Ours	**55.67**	**56.23**	53.88

kNN classifier at inference. DISCOVR not only achieves state-of-the-art scores, including the highest AUCs of 67.06 on EchoNet Dynamic, 57.42 on RVENET, and 57.23 on EchoPediatric LVH, but also demonstrates that richer spatio-temporal representations learned via simple SSL can offer more effective and efficient anomaly detection than more sophisticaed anomaly detection techniques without reliance on specialized or resource intensive modules.

Linear Probing. Table 9.2 shows that DISCOVR achieves the highest balanced accuracy and F1 score in linear probing for anomaly detection across all echocardiography datasets. For example, on Echonet Dynamic, DISCOVR attains an F1 of 77.63 compared to 75.50 for SIGMA, and on FetalEcho 2, achieves 63.59 versus 60.43 for MGMAE. These improvements are consistent across fetal, pediatric, and adult cohorts. While VideoMAE and MGMAE rely on high masking ratios and pixel-level reconstruction, their representations often miss subtle anatomical landmarks and temporally distributed abnormalities, reflecting a lack of deeper semantic abstraction. Clustering-based approaches such as SIGMA can capture

some temporal variation but lack explicit semantic guidance, limiting their ability to identify clinically relevant landmarks. In contrast, DISCOVR leverages semantic supervision from the image encoder through online distillation, combined with temporal modeling in the video branch. This enables DISCOVR to capture fine-grained spatial features and their evolution over time, resulting in representations that are both robust and clinically meaningful for anomaly detection in cardiac ultrasound.

Zero-Shot. Table 9.3 shows that DISCOVR achieves the highest balanced accuracy and F1 score for zero shot classification across all echocardiography datasets. For example, on Echonet Dynamic, DISCOVR reaches an F1 of 61.45 compared to 57.56 for the best baseline, and on FetalEcho 1, achieves 61.79 versus 60.64 for MGMAE. These improvements are consistent across fetal, pediatric, and adult cardiac cohorts. This stronger performance reflects DISCOVR's ability to integrate semantic features captured by the image encoder with temporal dynamics modeled by the video branch, explicitly aligned through the SCD loss during self-supervised training. Pixel reconstruction models such as VideoMAE and MGMAE focus primarily on low-level appearance and texture, and clustering approaches like SIGMA, while using temporal clips, lack explicit semantic guidance. Image-based baselines like RAD-DINO do not leverage temporal information, while methods such as MVD that rely on external pretrained teachers may be less adaptable to the clinical and domain-specific challenges of ultrasound video. DISCOVR's capabilities are further highlighted in the qualitative example of Fig. 9.3, where it detects subtle cardiac structures and correctly classifies a challenging fetal video as abnormal, while MVD fails to capture these cues and predicts a normal outcome. This underscores how DISCOVR's features are sufficiently fine-grained to enable accurate zero shot anomaly detection, even without task-specific tuning.

9.4.2 Segmentation Evaluation

We evaluate the effectiveness of DISCOVR representations for downstream cardiac segmentation using the CAMUS dataset [262]. As shown in Fig 9.4, DISCOVR achieves the highest Dice score (0.844), outperforming specialized segmentation architectures such as UNet and DeepLabV3 (0.816 and 0.819, respectively, both

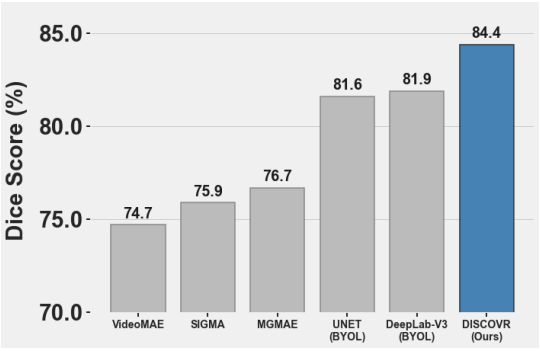


Figure 9.4: Barplot comparing the segmentation performance across different models. Our proposed DISCOVR approach achieves the highest Dice score of 0.844, outperforming both specialized segmentation architectures (DeepLab-V3, UNET) and other self-supervised methods.

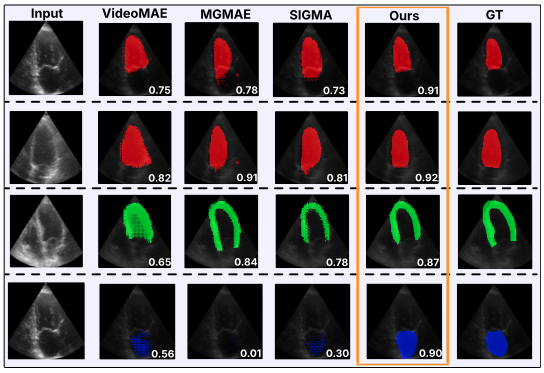


Figure 9.5: Segmentation comparison on the CAMUS dataset for left ventricular endocardium (LV Endo), left ventricular epicardium (LV Epi), and left atrium (LA). Our method produces accurate and consistent masks, achieving higher Dice scores (bottom right) than baseline methods.

with BYOL pretraining).When compared using a simple linear+upsampling head on a frozen backbone, DISCOVR also surpasses other SSL-based video models, including VideoMAE (0.747), MGMAE (0.767), and SIGMA (0.759). Fig. 9.5 highlights these advantages: DISCOVR produces consistently accurate and well-aligned segmentation masks for LV Endo, LV Epi, and especially the left atrium. For the challenging left atrium segmentation (blue mask), MGMAE misses the structure entirely (Dice = 0.01), while SIGMA and VideoMAE also perform poorly (Dice = 0.30 and 0.56). DISCOVR, in comparison, achieves 0.90, demonstrating superior ability to segment subtle structures and delineate boundaries due to its fine-grained feature learning.

9.4.3 LVEF Prediction

We evaluate the effectiveness of DISCOVR representations for downstream cardiac function estimation using the EchoNet-Dynamic ejection fraction dataset [129]. As shown in Table 9.4, DISCOVR achieves the lowest Mean Absolute Error (MAE) of 7.79 under the standard linear probing setup, outperforming other self-supervised baselines such as VideoMAE (8.02) and MGMAE (8.88). When fine-tuning only the last three encoder blocks, DISCOVR further reduces the MAE to 6.32, demonstrating the strength of its learned representations even with limited adaptation. In comparison, the fully supervised EchoNet-Dynamic model [129] is

trained end to end with all parameters updated. Under an ejection fraction–only setup without segmentation labels, DISCOVR surpasses these fully supervised baselines, including MC3 with an MAE of 6.59, the base EchoNet-Dynamic model with 7.35, and R3D with 7.63. The full EchoNet-Dynamic architecture achieves an MAE of 4.05 using a large multi-task design with 71.1 million parameters co-trained on 20,060 manual segmentation tracings. These results show that DISCOVR, through self-supervised pretraining and partial fine-tuning, learns powerful cardiac representations that rival or exceed fully supervised models trained end to end.

Table 9.4: LVEF prediction results on the EchoNet-Dynamic dataset. Our self-supervised method is compared against other SSL methods and fully-supervised baselines from [129].

Model	MAE ↓	RMSE ↓	EF Labels	Seg. Labels
Self-Supervised (Linear Probing)				
VideoMAE	8.02	11.16	✓	
MGMAE	8.88	12.47	✓	
DISCOVR (Ours)	**7.79**	**10.89**	✓	
Self-Supervised (Fine-tuning)				
DISCOVR (finetune last 3 blocks)	**6.32**	**8.62**	✓	
Fully-Supervised Baselines [1] trained only with EF Data				
MC3 (All frames)	6.59	9.39	✓	
EchoNet-Dynamic (EF, All frames)	7.35	9.53	✓	
R3D (All frames)	7.63	9.75	✓	
DISCOVR (finetune last 3 blocks,64 frames)	**6.32**	**8.62**	✓	
EchoNet-Dynamic (Full model)	**4.05**	**5.30**	✓	✓

9.5 Ablation Study

Table 9.5: Effect of loss terms.

$\mathcal{L}_{\text{ssl}}^{vid}$	$\mathcal{L}_{SCD}$	Bal. Acc.	Precision	F1
✓	✗	52.27	53.16	48.23
✓	✓	**63.20**	**65.35**	**61.45**

Table 9.6: Backbone size.

Backbone	Bal. Acc.	Precision	F1
ViT-Small	59.44	61.03	57.52
ViT-Base	**63.20**	**65.35**	**61.45**

Table 9.7: Masking ratio.

Mask (%)	Bal. Acc.	Precision	F1
50	55.60	56.90	52.98
75	56.25	57.58	53.85
90	**63.20**	**65.35**	**61.45**

Table 9.8: Number of frames.

Frames	Bal. Acc.	Precision	F1
16	57.89	59.45	55.68
32	59.54	61.36	57.45
64	**63.20**	**65.35**	**61.45**

In this section, we ablate the key components of the training objective in our model, **DISCOVR**. All experiments are conducted on the **Echonet Dynamic** dataset and evaluated using the **k-nearest neighbor (kNN) protocol**. This setup allows us to assess the discriminative quality of the learned representations in a fully frozen setting without additional fine-tuning.

Effect of Loss Components. We evaluate the effect of two core loss components used in DISCOVR: (i) the video self-distillation component ($\mathcal{L}_{\text{ssl}}^{\text{vid}}$), and (ii) the semantic cluster distillation component with online image guidance ($\mathcal{L}_{SCD}$). Table 9.5 reports the performance of these losses individually and in combination in zero-shot settings. Using only $\mathcal{L}_{\text{ssl}}^{\text{vid}}$ yields modest performance (F1 = 48.23%), as it primarily captures global temporal structure via CLS tokens but lacks guidance for fine-grained semantics. Introducing $\mathcal{L}_{SCD}$ leads to a substantial improvement (F1 = 61.45%, Balanced Accuracy = 63.20%), as the evolving image-based semantic clusters enrich the temporal features learned by the video model and encourage focus on more fine-grained, spatially grounded information. For more detailed ablation, refer to supplementary section B.1.4.

Effect of Backbone Size. We investigate how transformer backbone size impacts DISCOVR's representation quality. We evaluate ViT-Small and ViT-Base variants, each paired with matching DINO image encoders, on the Echonet Dynamic dataset using kNN evaluation (Table 9.6). ViT-Base achieves superior performance (F1=61.45%, balanced accuracy=63.20%) compared to ViT-Small (F1=57.52%,

balanced accuracy=59.44%). The smaller model's reasonable performance indicates DISCOVR learns meaningful representations even with limited capacity.

Effect of Number of Frames. In this ablation, we evaluate how the number of frames sampled from each video clip affects the representational quality learned by our model. We experiment with three temporal lengths: 16, 32, and 64 frames. All other training settings are kept constant, and the results are reported in Table 9.8. We observe a clear upward trend in performance with increasing frame count. Using 16 frames results in an F1 score of 55.68%, which improves to 57.45% with 32 frames. The best performance is achieved with 64 frames, yielding an F1 score of 61.45% and balanced accuracy of 63.20%. These results support the intuition that ultrasound, being a temporally dense and dynamic modality, benefits from longer clips. More frames provide richer temporal context, enabling the model to capture fine-grained spatial and temporal motion patterns across the cardiac cycle.

Effect of Masking Ratio. Table 9.7 shows a steady improvement in performance as the masking ratio increases, with F1-score rising from 52.98% (50%) to 61.45% (90%). Higher masking forces both the video encoder and the semantic image guidance branch to infer more from sparse visual cues, encouraging the model to focus on the most salient and non-redundant features. This promotes the learning of richer representations that better capture subtle and fine-grained spatio-temporal patterns, resulting in improved anomaly detection performance.

Computational Cost and Scalability. We report the computational cost, of our method in Table 9.9. The table shows both training and inference statistics, showing GPU memory usage and F1-score for each method on EchoNet-Dynamic, for a batch size of 1, 16 frames, and a spatial size of 112×112. During training, DISCOVR uses slightly more GPU memory (10.5GB) compared to prior methods (between 9.0 and 9.5GB) but achieves a notable +6.38% improvement in F1-score over the closest competitor. At inference, all methods, including DISCOVR, use identical ViViT-like encoders, resulting in nearly the same GPU memory footprint and FLOPS. This demonstrates that our method's performance improvements come with minimal extra training cost and no penalty for inference efficiency.

Table 9.9: Training and inference GPU memory, FLOPS, and F1-score on EchoNet-Dynamic, batch size = 1, 16 frames, 112 × 112 resolution.

Model	Train Mem (GB)	F1-score	Infer Mem (GB)	Infer. FLOPS
MGMAE [110]	9.0	46.13	1.153	101.85
VideoMAE [107]	9.0	55.07	1.153	101.85
SIGMA [258]	9.2	49.04	1.153	101.85
Video-distillation	9.5	48.23	1.153	101.85
DISCOVR (Ours)	10.5	**61.45**	1.153	101.85

9.6 Conclusion

We introduce DISCOVR, a self-supervised model for learning video representations in echocardiography across diverse patient populations. Our approach combines masked video modeling, temporal self-distillation, and online spatial supervision, unified by a Semantic Cluster Distillation (SCD) objective that aligns video and image features through cross-modal clustering, without relying on labeled anomalies or pretrained models. Extensively evaluated on six echocardiography datasets spanning fetal, pediatric, and adult populations, DISCOVR consistently outperforms previous self-supervised and anomaly detection methods for multiple tasks, including anomaly detection, classification (zero-shot and linear probing), and segmentation. DISCOVR's task-agnostic design and its applicability to diverse patient groups establish it as a strong foundation for screening cardiac conditions and developing assistive tools for echocardiography.

Acknowledgments

We acknowledge financial support from InnoHK-funded Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE), UKRI grant EP/X040186/1, UK EPSRC grant EP/T028572/1 (VisualAI), UK EPSRC Doctoral Training Partnership award and, UKRI AIRR Early Access Project No. ANON-BYYG-VX4C-Z.

Supplementary Material

9.7 Dataset Distribution

This section presents the dataset distributions for our five echocardiography video datasets: FetalEcho1 (Fig.9.6), FetalEcho2 (Fig.9.7), EchoNet-Dynamic (Fig.9.8), EchoNet-Pediatric (Fig.9.9), and RVENET (Fig.9.10). For each dataset, the bar chart displays the number of unique samples in the training, validation, and test sets. The accompanying pie charts illustrate the class distributions (Normal vs. Abnormal) within the validation and test sets.

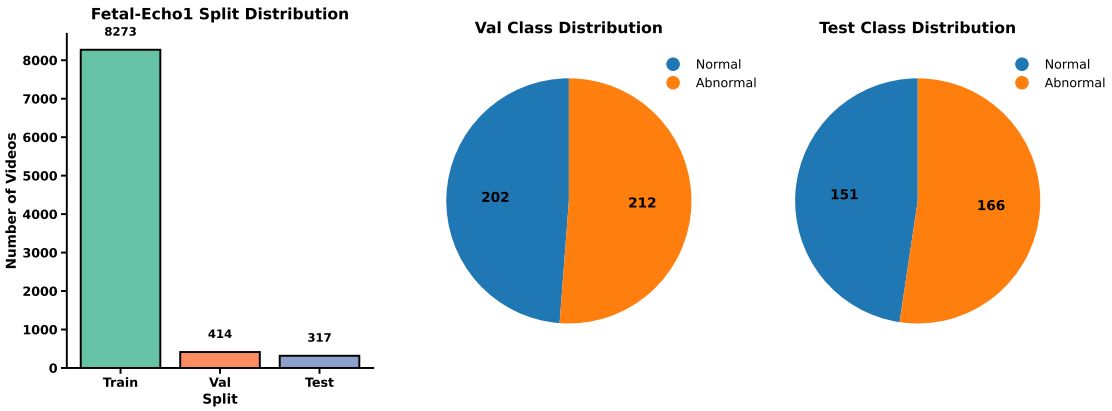


Figure 9.6: Dataset Distribution for Fetal-Echo1 dataset.

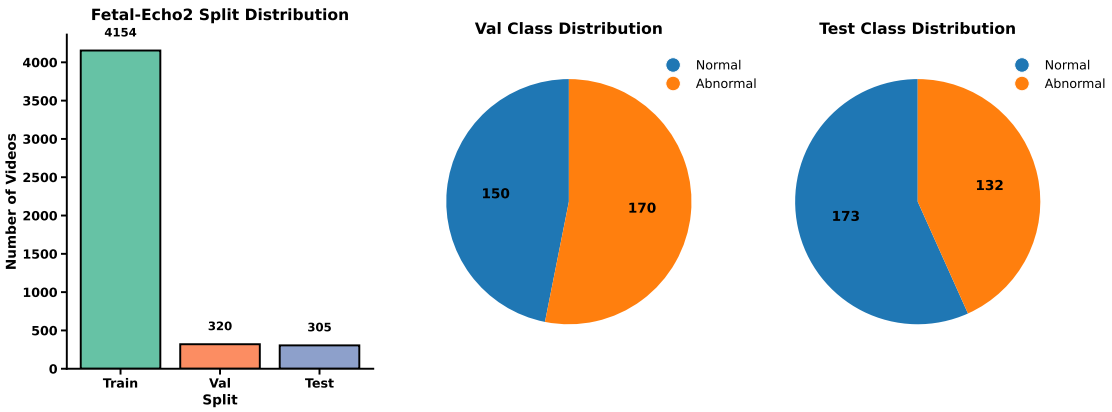


Figure 9.7: Dataset Distribution for Fetal-Echo2 dataset.

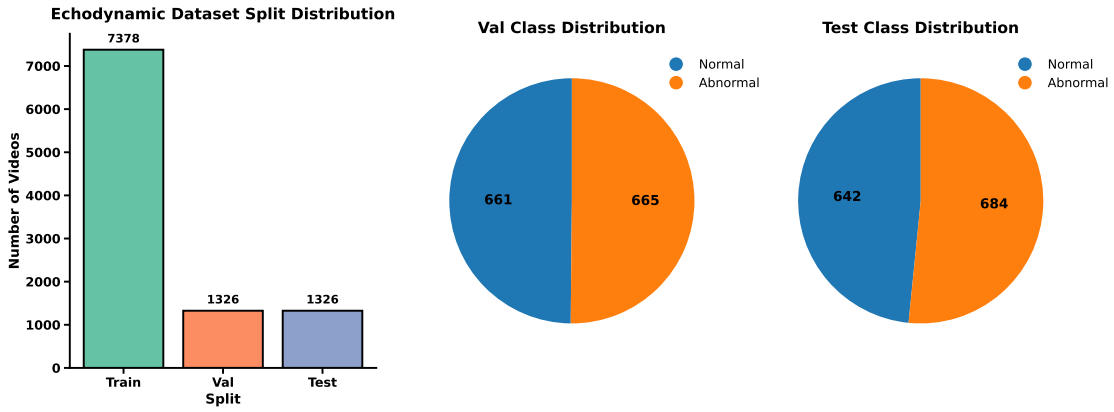


Figure 9.8: Dataset Distribution for Echo-Dynamic dataset.

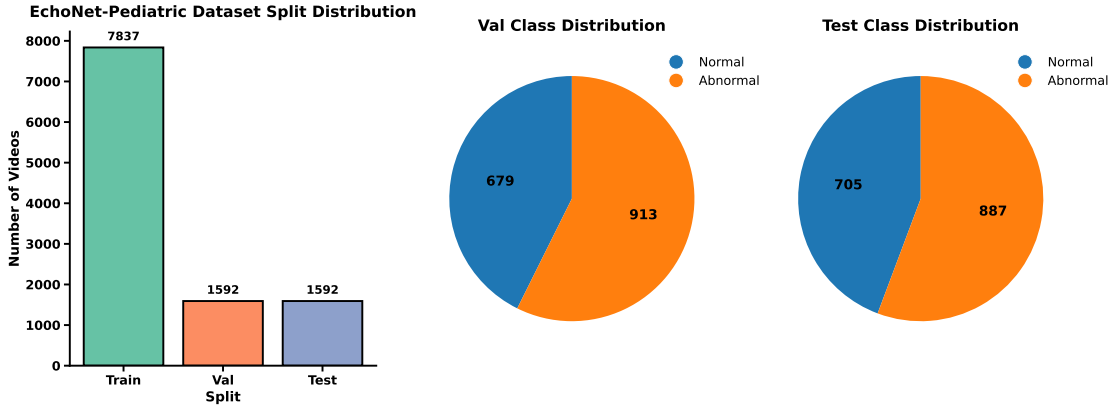


Figure 9.9: Dataset Distribution for Echo-Pediatric dataset.

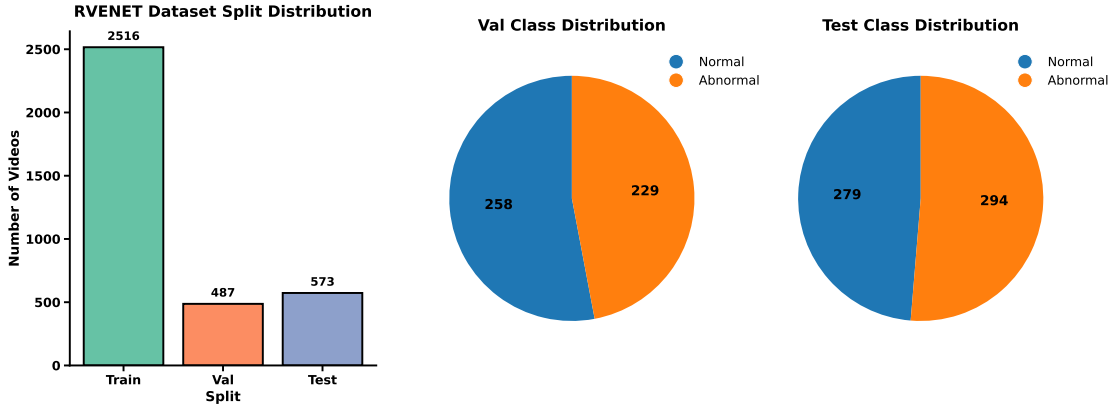


Figure 9.10: Dataset distribution for RVENET dataset.

9.8 Additional Results:

9.8.1 Full Finetuning

Evaluation Setup

We follow the same evaluation procedure as described in the experiments section in the main paper, but fine-tune the entire backbone along with the linear layer. All other evaluation settings remain unchanged. Results are reported on the Echonet Dynamic dataset to assess end-to-end supervised performance.

Evaluation Result:

Under full fine-tuning, as shown in Table 9.10, all models experience a drop in performance compared to their linear probing results, reflecting overfitting due to the limited labeled validation data. Despite this, DISCOVR achieves the highest F1 score of 70.44%, outperforming MGMAE (65.99%), SIGMA (61.46%), and VideoMAE (57.31%). DISCOVR's structured representation learning, through temporal distillation and cross-modal clustering, appears to provide more robust and adaptable features, enabling it to generalize better even when fully fine-tuned on a small dataset.

Table 9.10: Table showing the full-finetuning result of DISCOVR compared to other baselines on the Echo-Dynamic Dataset

Model (%)	Accuracy	Balanced Acc.	Precision	Recall	F1-Score
VideoMAE	57.62	57.94	58.27	57.94	57.31
SIGMA	61.69	62.00	62.41	62.00	61.46
MGMAE	65.99	66.08	66.10	66.08	65.99
Ours	**70.51**	**70.42**	**70.50**	**70.42**	**70.44**

Generalisation to Other Modalities.

To test whether DISCOVR generalizes beyond echocardiography, we evaluated its transfer performance on two distinct medical image benchmarks: the Breast Ultrasound Images dataset [268] (cancer detection across 600 patients) and DermMNIST [269] (skin lesion classification). Both breast ultrasound and echocardiography require the detection of small, irregular regions of altered tissue, such as hypoechoic

tumors in the breast or localized wall motion abnormalities in the heart, making the ability to identify subtle structural changes in one domain directly applicable to the other. Similarly, DermMNIST demands fine-grained visual discrimination between morphologically similar skin lesions. For both benchmarks, we froze the DISCOVR encoder and trained a linear classifier, comparing performance directly across methods.

As shown in Table 9.11, our method demonstrates strong generalization across both tasks. On the Breast ultrasound dataset, DISCOVR improves balanced accuracy by 2.01% over VideoMAE, 19.83% over SIGMA, and 12.01% over MGMAE. For DermMNIST, DISCOVR achieves an accuracy of 71.68%, outperforming VideoMAE by 2.85%, SIGMA by 3.00%, and MGMAE by 3.85%. These results demonstrate strong generalization to multiple medical image analysis tasks beyond echocardiography. Further, to assess generalization to natural video data, we pretrained and evaluated all models on the Kinetics 400 action recognition benchmark, using a zero-shot protocol where KNN classification with $K = 20$ was applied to features using 64 frames from the frozen video backbone. As shown in Table 9.12, DISCOVR achieves the highest Top-1 accuracy at 22.3%, outperforming MVD by 3.6%, MME by 3.2%, and VideoMAE by 1.6%, while also requiring the fewest pretraining epochs. These results highlight that DISCOVR not only excels at medical video tasks but also learns generalizable representations efficiently for large-scale natural video datasets.

Table 9.11: Linear Probing results on the **Breast ultrasound** dataset and **DermMNIST**. For Breast ultrasound, we report Balanced Accuracy (Bal. Acc.) and F1; for DermMNIST, we report overall Accuracy (Acc.).

Method	**Breast ultrasound**		**DermMNIST**
	Balanced Accuracy	**F1 Score**	**Accuracy**
VideoMAE	61.75	64.45	68.83
SIGMA	43.93	42.21	68.68
MGMAE	51.75	52.34	67.83
DISCOVR (Ours)	**63.76**	**65.44**	**71.68**

Table 9.12: Zero-Shot KNN classification performance on Kinetics-400.

Model	Epochs	Top-1 Accuracy (%)
MVD [263]	1600	18.7
MME [270]	800	19.1
VideoMAE [107]	800	20.7
DISCOVR (Ours)	**400**	**22.30**

Loss function Ablation detailed.

To rigorously evaluate each component, we have added baselines using only masked image or only video self-distillation. Indeed, we find that the settings perform suboptimally, as shown in Table 1, confirming that spatial or temporal cues alone are insufficient for strong representation learning. In contrast, combining both with the SCD loss, which explicitly distills fine-grained semantic structure from the image branch into the video backbone, achieves the best results. This supports our intuition that SCD is crucial for aligning spatial semantics with temporal dynamics, enabling more robust and clinically meaningful video representations.

Table 9.13: Effect of the different loss terms on classification performance (Balanced Accuracy, Precision, and F1).

$\mathcal{L}_{\text{ssl}}^{vid}$	$\mathcal{L}_{\text{ssl}}^{img}$	$\mathcal{L}_{\text{SCD}}$	Bal. Acc.	Precision	F1
✓	✗	✗	52.27	53.16	48.23
✗	✓	✗	53.66	55.22	49.43
✓	✓	✓	**63.20**	**65.35**	**61.45**

9.9 Implementation Details

All models are implemented in PyTorch 2.6 and trained on RTX 8000 GPUs (48 GB) with a batch size of 8 using the AdamW optimizer. Videos are processed as 64-frame clips sampled at a stride of 3 and resized to 112×112.

For both video and image self-distillation, we use a student-teacher setup where the teacher processes the full input and the student observes $N = 4$ randomly masked views. The teacher network is updated via an exponential moving average (EMA) of the student with momentum $\lambda = 0.996$. A fixed temperature $\tau_s = 0.1$ is

used for the student, while the teacher temperature τ_t is linearly warmed from 0.04 to 0.07 over the first 30 epochs. Semantic Cluster Distillation (SCD) uses $K = 3000$ learnable prototypes, with similarity scores computed via temperature-scaled dot products ($\tau = 0.1$) and cluster assignments generated using the Sinkhorn-Knopp algorithm (10 iterations, $\epsilon = 0.05$). Models are trained for 400 epochs with a learning rate of 1.5×10^{-4}, weight decay of 0.05, and 40 warmup epochs.

9.10 Broader Impact and Limitations

In this work, we introduce **DISCOVR**, a novel self-supervised model for echocardiography video understanding across fetal, pediatric, and adult populations. Trained without labeled abnormal cases, DISCOVR learns rich spatiotemporal representations and enables zero-shot inference. One key application of DISCOVR is in the *early screening of heart diseases*, where it can assist clinicians by flagging potential anomalies in echocardiography videos. This has significant clinical relevance, as congenital heart defects affect approximately 1 in 100 newborns, with up to 50% missed during prenatal screening [4, 224, 271], and cardiovascular diseases remain the leading global cause of death [272]. By reducing reliance on large, labeled datasets, DISCOVR offers a scalable and accessible solution, particularly for deployment in low-resource settings.

While DISCOVR shows strong potential, its current scope is focused specifically on echocardiography, and it has not yet been evaluated on other imaging modalities. The model was trained and tested on five datasets collected from distinct clinical sites, each with its own imaging protocols, devices, and patient cohorts. As a result, the demographic and geographic diversity of the data may be limited. Further validation is needed to assess the model's generalizability across broader clinical settings, populations, and imaging systems.

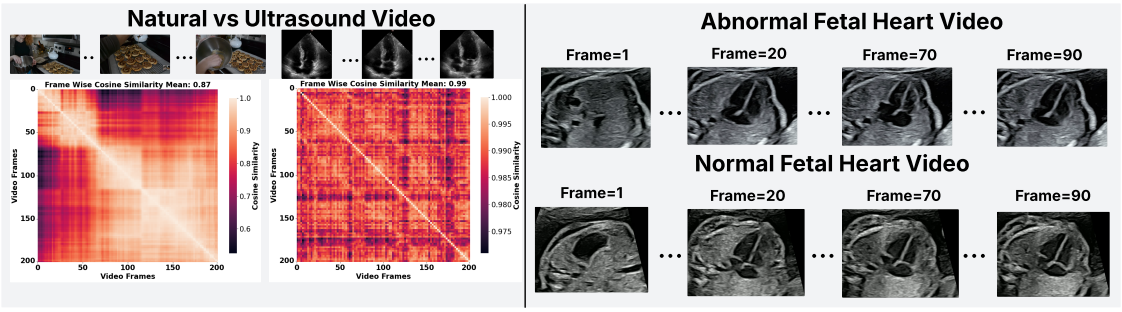


Figure 9.11: Figure (left) compares two fine-grained videos: a natural scene of a person baking (left) and an adult heart ultrasound (right). The frame-level cosine similarity matrix, computed using a pretrained VideoMAE model, shows that ultrasound frames are highly similar (mean=0.99), with only minor local variations. This highlights the difficulty in distinguishing individual frames in such medical videos. Figure (right) compares normal and abnormal fetal echocardiograms, which appear almost identical despite one being abnormal. This illustrates the inherent difficulty of distinguishing subtle cardiac abnormalities in fetal imaging.

10 Conclusion and Future Work

In this final chapter, we summarize the main contributions of the thesis, discuss the limitations of the proposed approaches, and highlight future research directions. The work presented in this thesis advances the automation of congenital heart disease (CHD) screening by shifting from static image analysis to video understanding of fetal ultrasound. Traditional image-based approaches capture only limited anatomical information from brief scans, while the fetal heart is inherently dynamic, with morphology and motion that evolve throughout the cardiac cycle. By modelling the spatio-temporal structure of ultrasound videos, this research leverages temporal coherence, cardiac motion cues, and anatomical context to improve the accuracy and efficiency of CHD detection and aims to support sonographers in screening more effectively.

10.1 Discussions and Contributions

Chapter 4 – Unsupervised Out-of-Distribution Detection Using Diffusion Models

The first stage of this research focused on identifying heart views in free-hand ultrasound videos, with the aim of reducing the manual effort required to review large numbers of frames. Dual-Conditioned Diffusion Models (DCDM) were introduced for unsupervised out-of-distribution (OOD) detection, formulating heart frame

identification as a reconstruction problem. By conditioning the generative process on both in-distribution class information and latent image features, DCDM reconstructs anatomically consistent heart frames while rejecting non-cardiac content. The proposed approach achieved improvements of 12% in accuracy, 22% in precision, and 8% in F1-score compared with existing reconstruction-based baselines. By isolating heart-specific video content from free-hand ultrasound scans, this work enables subsequent modelling and analysis to focus on cardiac structure and dynamics.

Chapter 5 – STAN-LOC: Visual Query-Based Video-Clip Localisation

Building on the ability to identify cardiac frames, the next stage addressed the challenge of retrieving diagnostically relevant standard cardiac views from continuous heart sweeps. In clinical practice, locating these standard planes is time-consuming and highly dependent on operator expertise. In response, the task of visual query-based video-clip localisation (VQ-VCL) was formulated. The proposed STAN-LOC model uses a query-aware spatio-temporal fusion transformer to combine information from a visual query image and an ultrasound video, together with a multi-anchor contrastive loss to handle annotation noise and a query selection module for robust inference. STAN-LOC achieved a 22% improvement in mean temporal Intersection over Union (mtIoU), showing that visual query-based retrieval can reliably localize standard cardiac views in heart sweep videos.

Chapter 6 – TIER-LOC: Multi-Tier Transformer for Multi-View Localisation

To extend localisation beyond a single view, TIER-LOC was developed to retrieve multiple standard cardiac views from each sweep. The model extracts and integrates features across several spatial and temporal tiers, capturing both coarse and fine-grained motion patterns. A dual-anchor contrastive loss improves view separation, while a temporal uncertainty-aware loss reduces the effect of noisy boundaries. TIER-LOC achieved consistent performance gains of 7–8% mtIoU across datasets, demonstrating robust localisation across different cardiac views and acquisition conditions. This extension supports multi-view analysis within routine heart sweeps, strengthening the clinical applicability of visual query-based video-clip localisation.

Chapter 7 – MCAT: Multi-Tier Class-Aware Token Transformer

Recognising that shared embeddings limit the discrimination of subtle anatomical differences, the Multi-Tier Class-Aware Token Transformer (MCAT) was introduced to learn view-specific tokens that represent distinct anatomical patterns. Each token selectively activates for the queried cardiac view, improving localisation accuracy while reducing computational cost. MCAT achieved 10–13% higher mean temporal Intersection over Union on fetal ultrasound datasets and a 5.35% improvement on the Ego4D benchmark, while reducing token usage by 96%. In combination with STAN-LOC and TIER-LOC, this work extends visual query-based video-clip localisation from single-view retrieval to efficient, class-aware multi-view localisation, aligning the modelling approach with the practical requirements of cardiac screening workflows.

Chapter 8 – STUD and DiVMerge: Self-Supervised Normality Learning

While the earlier chapters focus on identifying and localising standard cardiac views, many hospitals lack sufficient annotated abnormal data for supervised disease detection. In this setting, Sparse Tube Ultrasound Distillation (STUD) was proposed as a self-supervised, zero-shot anomaly detection framework that models normal heart function using sparsely sampled spatio-temporal tubes. By operating on sparse tube representations rather than densely tokenised video sequences, STUD substantially reduces the number of processed tokens, enabling efficient modelling of long ultrasound videos. To combine knowledge across hospitals without data sharing, a divergence-guided model merging approach (DiVMerge) was developed to aggregate site-specific models using geometric medians and adaptive weighting. Across five clinical sites, the merged model achieved 23.77% higher accuracy and 30.13% higher F1-score than local baselines, demonstrating that decentralized, privacy-preserving, and computationally efficient training can support robust CHD detection across institutions.

Chapter 9 – DISCOVR: Online Cluster Distillation for Echocardiographic Video Representation

Finally, DISCOVR (Distilled Image Supervision for Cross-Modal Video Representation) was proposed to learn transferable video representations applicable across different cardiac imaging modalities. DISCOVR integrates a clustering-based video encoder with an online image encoder through semantic cluster distillation, allowing anatomical information learned from static images to guide temporal video representations. The model achieved a 3.4% improvement in F1-score for anomaly detection, a 2.4% gain in linear probing accuracy, and a 3.1% increase in Dice score for segmentation across six fetal, paediatric, and adult echocardiography datasets.

By producing task-agnostic representations that capture both spatial semantics and temporal dynamics, DISCOVR supports a range of downstream cardiac ultrasound analysis tasks without task-specific supervision.

Summary

Collectively, the studies presented in this thesis establish a coherent approach to video-based analysis of fetal cardiac ultrasound, shaped by both the characteristics of the data and the evolving research questions. The work begins with free-hand ultrasound videos, which are widely used in routine screening and provide broad anatomical coverage across the fetus. Exploring this setting enabled the development of unsupervised methods for identifying heart-related frames within large volumes of free-hand ultrasound video and established an initial foundation for automated analysis of fetal heart data. As the research progressed, it became clear that many clinically relevant questions require not only identifying the heart, but also analysing its spatio-temporal motion in a more focused and structured manner. This motivated a transition to continuous CAIFE heart sweeps, which provide richer spatial and temporal coverage of the fetal heart and enable more detailed modelling of cardiac anatomy and dynamics. Within this setting, the work progressed from unsupervised identification of heart frames to query-based localisation of standard cardiac view video clips, with the aim of guiding sonographers and simplifying clinical workflow. As the modelling shifted from single-view tasks to learning representations

across multiple standard views, the approach evolved toward multi-view and class-aware localisation. Building on these developments, later chapters shifted focus toward assisting disease detection through self-supervised normality modelling and, ultimately, toward learning more general video representations of cardiac structure and motion. This progression reflects a deliberate change in emphasis over the course of the PhD: from specialized, task-driven systems constrained by explicit human definitions to more representational, task-agnostic frameworks that aim to capture the underlying structure of the cardiac ultrasound domain. The work is supported by a multi-institutional dataset comprising scans from five clinical sites, acquired using multiple ultrasound systems and operators across a diverse patient population, enabling systematic evaluation of robustness and generalisation. By treating ultrasound as a temporal signal rather than a collection of static images, the proposed methods leverage the anatomical and dynamic richness of cardiac video to reduce manual effort and support a wide range of downstream tasks. While full clinical validation remains ongoing, this thesis provides a principled foundation for future assistive systems designed to support more consistent and effective congenital heart disease screening across diverse clinical settings.

Clinical Implications

From a clinical perspective, the methods developed in this thesis can be viewed as enabling two assistive roles within fetal ultrasound practice. First, visual query–based video clip localisation models, particularly MCAT as the strongest-performing VQ-VCL approach, can support sonographers during review by allowing them to specify a cardiac view of interest using a visual query and automatically retrieve the corresponding video clip from a heart sweep. This enables focused inspection of the relevant anatomy, facilitates cross-checking against clinical guidelines, and supports downstream measurements, while leaving all interpretation and decision-making with the sonographer.

Second, the video foundation model DISCOVR, used for the downstream task of video anomaly detection, addresses a different clinical need that arises in routine practice, particularly in the UK, where most anomaly scans are performed by general obstetric sonographers rather than specialist fetal cardiologists. In such settings,

a screening model could assist by highlighting scans with non-standard cardiac motion or structure for further review or referral, helping to prioritise cases that may benefit from specialist input. While the evaluation in this thesis is retrospective and draws on data from five clinical sites, future work would require prospective validation on data from these sites, followed by extension to additional hospitals to assess generalisability before any clinical adoption.

10.2 Limitations

Although the methods proposed in this thesis advance the automation of fetal cardiac screening, several limitations remain at both methodological and practical levels.

- **Annotation noise and ground truth variability:** Manual annotations of cardiac view boundaries and standard planes contain subjectivity and inter-observer variability. Although the proposed contrastive and uncertainty-aware loss functions reduce the impact of such noise, inherent ambiguity in annotations limits the achievable upper bound of performance.
- **Computational complexity:** Multi-tier transformers and self-supervised models such as DCDM, TIER-LOC, and DISCOVR require substantial computational resources for training and inference. Future work should explore parameter-efficient fine-tuning or model compression to enable real-time deployment on clinical workstations.
- **Limited abnormal data:** CHD cases are relatively rare - this is why CAIFE is collating data from multiple sites. The lack of comprehensive and well-annotated abnormal cases restricts evaluation of anomaly detection models across diverse CHD subtypes. Although STUD and DISCOVR mitigate this limitation through self-supervision, further clinical validation is required to confirm diagnostic reliability.
- **Workflow integration:** The proposed systems are first prototypes and have not yet been integrated or evaluated in real-time sonography workflows. Further studies involving prospective testing, human–AI interaction analysis, and usability evaluation are necessary before translation into clinical practice.

10.3 Future Directions

Building on the methods and findings of this thesis, several directions are proposed for future research:

- **Larger and more diverse clinical studies:** Expanding multi-center collaborations to include a wider range of hospitals, patient demographics, and ultrasound manufacturers will help validate the generalisability and fairness of the proposed models.
- **Real-time and resource-efficient models:** Further research on lightweight architectures, parameter-efficient fine-tuning, and edge-device optimisation will be essential for deployment on portable scanners and clinical machines.
- **Cross-modal and physiology-aware learning:** Incorporating complementary modalities such as Doppler imaging or M-mode signals, along with explicit modelling of cardiac motion phases, may enhance the physiological relevance of predictions.
- **Explainability and human–AI collaboration:** Developing interpretable visualisations, temporal saliency maps, and feedback interfaces can help sonographers understand model outputs, improving trust and adoption in clinical settings.
- **Clinical workflow evaluation:** Prospective studies evaluating the impact of AI assistance on scanning time, diagnostic accuracy, and inter-operator consistency will be crucial for regulatory approval and eventual clinical deployment.

10.4 Conclusion

This thesis presents a coherent progression from unsupervised detection of cardiac frames to transformer-based localisation of standard views and self-supervised modelling of normal and abnormal cardiac motion. By framing congenital heart disease detection as a video understanding problem, the research moves beyond

static image analysis to exploit the full temporal dynamics of ultrasound imaging. Across this progression, the modelling emphasis shifts from specialized, task-specific solutions focused on individual cardiac views toward learning richer, more general representations that capture variation across views and cardiac motion. Through DCDM, STAN-LOC, TIER-LOC, MCAT, STUD, and DISCOVR, this work contributes new methodological foundations for automated analysis of fetal cardiac videos. Collectively, these methods demonstrate that modelling the temporal and spatial richness of ultrasound data can enhance both the accuracy and efficiency of prenatal screening. While further validation and integration studies are needed, the approaches developed in this thesis represent an important step toward intelligent, video-based systems that aim to support sonographers in detecting congenital heart disease more effectively and consistently in diverse clinical environments.

11
Statement of Joint Authorship

A statement of joint first authorship is provided for each joint first–author paper included in this thesis. These statements describe the candidate's and co-authors' individual research contributions to the published work. For each publication, a complete statement of authorship has been prepared and signed by the candidate and the supervisor.

Statement of Authorship for the paper "Self-supervised Normality Learning and Divergence Vector-guided Model Merging for Zero-shot Congenital Heart Disease Detection in Fetal Ultrasound Videos" in Chapter 9.

Paper title	Self-supervised Normality Learning and Divergence Vector-guided Model Merging for Zero-shot Congenital Heart Disease Detection in Fetal Ultrasound Videos
Authors	Pramit Saha* , **Divyanshu Mishra*** , Netzahualcoyotl Hernandez-Cruz , Olga Patey , Aris Papageorghiou , Yuki M. Asano , and J. Alison Noble. (* equal contribution)
Publication status	Published
Publication details	International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2025

Student Confirmation

Student name	Divyanshu Mishra	
Contribution to the paper	First-author contribution: • Conceived and developed the video-based component of the paper, including dataset preparation, model design, implementation, and optimization for CHD detection. • Wrote the dataset and video model sections, as well as the CHD detection motivation and task description, and presented the work at the conference.	
Signature and Date		Oct. 29th 2025

Supervisor Confirmation

By signing the Statement of Authorship, the supervisor is certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor name	Prof. J. Alison Noble	
Supervisor comments		
Signature and Date		Oct. 29th 2022

"We are all like fireworks: we climb, we shine and always go our separate ways and become further apart. But even when that time comes, let's not disappear like a firework and continue to shine forever."

— *Bleach*

References

[1] CDC. *Data and Statistics — cdc.gov*. https://www.cdc.gov/heart-defects/data/?CDC_AAref_Val=https://www.cdc.gov/ncbddd/heartdefects/data.html. [Accessed 11-03-2025].

[2] *Congenital disorders — who.int*. https://www.who.int/news-room/fact-sheets/detail/birth-defects. [Accessed 11-03-2025].

[3] Julene S Carvalho et al. "ISUOG Practice Guidelines (updated): sonographic screening examination of the fetal heart". In: (2013).

[4] Amber EL van Nisselrooij et al. "Why are congenital heart defects being missed?" In: *Ultrasound in Obstetrics & Gynecology* 55.6 (2020), pp. 747–757.

[5] Lior Drukker et al. "Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video". In: *Scientific Reports* 11.1 (2021), p. 14109.

[6] Divyanshu Mishra et al. "Dual conditioned diffusion models for out-of-distribution detection: Application to fetal ultrasound videos". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 216–226.

[7] Divyanshu Mishra et al. "STAN-LOC: Visual Query-Based Video Clip Localization for Fetal Ultrasound Sweep Videos". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 742–752.

[8] Divyanshu Mishra et al. "TIER-LOC: Visual Query-based Video Clip Localization in fetal ultrasound videos with a multi-tier transformer". In: *Medical Image Analysis* (2025), p. 103611.

[9] Divyanshu Mishra et al. "MCAT: Visual Query-Based Localization of Standard Anatomical Clips in Fetal Ultrasound Videos Using Multi-Tier Class-Aware Token Transformer". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 27. 2025, pp. 28267–28275.

[10] Pramit Saha et al. "Self-supervised Normality Learning and Divergence Vector-guided Model Merging for Zero-shot Congenital Heart Disease Detection in Fetal Ultrasound Videos". In: *arXiv preprint arXiv:2503.07799* (2025).

[11] Meghan S Zimmerman et al. "Global, regional, and national burden of congenital heart disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017". In: *The Lancet Child & Adolescent Health* 4.3 (2020), pp. 185–200.

[12] Gregory A Roth et al. "Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study". In: *Journal of the American College of Cardiology* 76.25 (2020), pp. 2982–3021.

[13] Shan An et al. “Simultaneous segmentation of four cardiac chambers in fetal echocardiography”. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 3122–3126.

[14] Kaiming He et al. “Mask r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[16] Ken CL Wong et al. “Multiview and multiclass image segmentation using deep learning in fetal echocardiography”. In: *Medical Imaging 2021: Computer-Aided Diagnosis*. Vol. 11597. SPIE. 2021, pp. 308–313.

[17] Bin Pu et al. “Mobileunet-fpn: a semantic segmentation model for fetal ultrasound four-chamber segmentation in edge computing environments”. In: *IEEE Journal of Biomedical and Health Informatics* 26.11 (2022), pp. 5540–5550.

[18] Andrew G Howard et al. “Mobilenets: Efficient convolutional neural networks for mobile vision applications”. In: *arXiv preprint arXiv:1704.04861* (2017).

[19] Tsung-Yi Lin et al. “Feature pyramid networks for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[20] Arijit Patra, Weilin Huang, and J Alison Noble. “Learning spatio-temporal aggregation for fetal heart analysis in ultrasound video”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer. 2017, pp. 276–284.

[21] Elisa Chotzoglou et al. “Learning normal appearance for fetal anomaly screening: Application to the unsupervised detection of Hypoplastic Left Heart Syndrome”. In: *arXiv preprint arXiv:2012.03679* (2020).

[22] Mihaela Rosca et al. “Variational approaches for auto-encoding generative adversarial networks”. In: *arXiv preprint arXiv:1706.04987* (2017).

[23] Sudhakar Sengan et al. “Echocardiographic Image Segmentation for Diagnosing Fetal Cardiac Rhabdomyoma During Pregnancy Using Deep Learning”. In: *IEEE Access* 10 (2022), pp. 114077–114091.

[24] Ade Iriani Sapitri et al. “Deep learning-based real time detection for cardiac objects with fetal ultrasound video”. In: *Informatics in Medicine Unlocked* 36 (2023), p. 101150.

[25] Lu Xu et al. “DW-Net: A cascaded convolutional neural network for apical four-chamber view segmentation in fetal echocardiography”. In: *Computerized Medical Imaging and Graphics* 80 (2020), p. 101690.

[26] Jeremy Tan et al. "Automated detection of congenital heart disease in fetal ultrasound screening". In: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis: First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings 1*. Springer. 2020, pp. 243–252.

[27] Christian F Baumgartner et al. "SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound". In: *IEEE transactions on medical imaging* 36.11 (2017), pp. 2204–2215.

[28] Jinbao Dong et al. "A generic quality control framework for fetal ultrasound cardiac four-chamber planes". In: *IEEE journal of biomedical and health informatics* 24.4 (2019), pp. 931–942.

[29] Yuhuan Lu et al. "A YOLOX-based Deep Instance Segmentation Neural Network for Cardiac Anatomical Structures in Fetal Ultrasound Images". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2022).

[30] Zheng Ge et al. "Yolox: Exceeding yolo series in 2021". In: *arXiv preprint arXiv:2107.08430* (2021).

[31] Rima Arnaout et al. "An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease". In: *Nature medicine* 27.5 (2021), pp. 882–891.

[32] Siti Nurmaini et al. "Accurate detection of septal defects with fetal ultrasonography images using deep learning-based multiclass instance segmentation". In: *IEEE Access* 8 (2020), pp. 196160–196174.

[33] Yuxin Gong et al. "Fetal congenital heart disease echocardiogram screening based on DGACNN: adversarial one-class classification combined with video transfer learning". In: *IEEE transactions on medical imaging* 39.4 (2019), pp. 1206–1222.

[34] Ishaan Gulrajani et al. "Improved training of wasserstein gans". In: *Advances in neural information processing systems* 30 (2017).

[35] Siti Nurmaini et al. "An improved semantic segmentation with region proposal network for cardiac defect interpretation". In: *Neural Computing and Applications* 34.16 (2022), pp. 13937–13950.

[36] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28 (2015).

[37] Sibo Qiao et al. "Flds: An intelligent feature learning detection system for visualizing medical images supporting fetal four-chamber views". In: *IEEE Journal of Biomedical and Health Informatics* 26.10 (2021), pp. 4814–4825.

[38] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection". In: *arXiv preprint arXiv:2004.10934* (2020).

[39] Masaaki Komatsu et al. "Detection of cardiac structural abnormalities in fetal ultrasound videos using deep learning". In: *Applied Sciences* 11.1 (2021), p. 371.

[40] Kaiwen Duan et al. "Centernet: Keypoint triplets for object detection". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6569–6578.

[41] Xizhou Zhu et al. “Flow-guided feature aggregation for video object detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 408–417.

[42] Yihong Chen et al. “Memory enhanced global-local aggregation for video object detection”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10337–10346.

[43] Shiyao Wang et al. “Fully motion-aware network for video object detection”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 542–557.

[44] Xingyu Chen, Junzhi Yu, and Zhengxing Wu. “Temporally identity-aware SSD with attentional LSTM”. In: *IEEE transactions on cybernetics* 50.6 (2019), pp. 2674–2686.

[45] Chen Zhang and Joohee Kim. “Video object detection with two-path convolutional LSTM pyramid”. In: *IEEE Access* 8 (2020), pp. 151681–151691.

[46] Haiping Wu et al. “Sequence level semantics aggregation for video object detection”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 9217–9225.

[47] Hanming Deng et al. “Object guided external memory network for video object detection”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 6678–6687.

[48] Masato Fujitake and Akihiro Sugimoto. “Temporal feature enhancement network with external memory for object detection in surveillance video”. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 7684–7691.

[49] Lu He et al. “End-to-end video object detection with spatial-temporal transformers”. In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 1507–1516.

[50] Han Wang et al. “PTSEFormer: Progressive Temporal-Spatial Enhanced TransFormer Towards Video Object Detection”. In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*. Springer. 2022, pp. 732–747.

[51] Qianyu Zhou et al. “TransVOD: end-to-end video object detection with spatial-temporal transformers”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[52] Xizhou Zhu et al. “Deformable detr: Deformable transformers for end-to-end object detection”. In: *arXiv preprint arXiv:2010.04159* (2020).

[53] Yiming Cui. “FAQ: Feature Aggregated Queries for Transformer-based Video Object Detectors”. In: *arXiv preprint arXiv:2303.08319* (2023).

[54] Zhiwu Qing et al. “Temporal context aggregation network for temporal action proposal refinement”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 485–494.

[55] Deepak Sridhar et al. “Class semantics-based attention for action detection”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 13739–13748.

[56] Dingfeng Shi et al. “Tridet: Temporal action detection with relative boundary modeling”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 18857–18866.

[57] Xin Li et al. “Deep concept-wise temporal convolutional networks for action localization”. In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 4004–4012.

[58] Zixin Zhu et al. “Enriching local and global contexts for temporal action localization”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 13516–13525.

[59] Guo Chen et al. “DCAN: Improving temporal action detection via dual context aggregation”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. 1. 2022, pp. 248–257.

[60] Victor Escorcia et al. “Daps: Deep action proposals for action understanding”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer. 2016, pp. 768–784.

[61] Chuming Lin et al. “Fast learning of temporal action proposal via dense boundary generator”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 07. 2020, pp. 11499–11506.

[62] Tianwei Lin et al. “Bmn: Boundary-matching network for temporal action proposal generation”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 3889–3898.

[63] Tianwei Lin et al. “Bsn: Boundary sensitive network for temporal action proposal generation”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[64] Xiaolong Liu et al. “Multi-shot temporal event localization: a benchmark”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 12596–12606.

[65] Tianwei Lin, Xu Zhao, and Zheng Shou. “Single shot temporal action detection”. In: *Proceedings of the 25th ACM international conference on Multimedia*. 2017, pp. 988–996.

[66] Wei Liu et al. “Ssd: Single shot multibox detector”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer. 2016, pp. 21–37.

[67] Shyamal Buch et al. “End-to-end, single-stream temporal action detection in untrimmed videos”. In: (2019).

[68] Fuchen Long et al. “Gaussian temporal awareness networks for action localization”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 344–353.

[69] Le Yang et al. “Revisiting anchor mechanisms for temporal action localization”. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 8535–8548.

[70] Chuming Lin et al. “Learning salient boundary feature for anchor-free temporal action localization”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 3320–3329.

[71] Feng Cheng and Gedas Bertasius. "TallFormer: Temporal Action Localization with a Long-Memory Transformer". In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*. Springer. 2022, pp. 503–521.

[72] Xiaolong Liu, Song Bai, and Xiang Bai. "An empirical study of end-to-end temporal action detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20010–20019.

[73] Xiaolong Liu et al. "End-to-end temporal action detection with transformer". In: *IEEE Transactions on Image Processing* 31 (2022), pp. 5427–5441.

[74] Dingfeng Shi et al. "React: Temporal action detection with relational queries". In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*. Springer. 2022, pp. 105–121.

[75] Jing Tan et al. "Relaxed transformer decoders for direct action proposal generation". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 13526–13535.

[76] Chen-Lin Zhang, Jianxin Wu, and Yin Li. "Actionformer: Localizing moments of actions with transformers". In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*. Springer. 2022, pp. 492–510.

[77] Kristen Grauman et al. "Ego4d: Around the world in 3,000 hours of egocentric video". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 18995–19012.

[78] Mengmeng Xu et al. "Where is my Wallet? Modeling Object Proposal Sets for Egocentric Visual Query Localization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2593–2603.

[79] Raghav Goyal et al. "MINOTAUR: Multi-task Video Grounding From Multimodal Queries". In: *arXiv preprint arXiv:2302.08063* (2023).

[80] Antoine Yang et al. "Tubedetr: Spatio-temporal video grounding with transformers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16442–16453.

[81] Ting Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *International Conference on Machine Learning (ICML)*. 2020, pp. 1597–1607.

[82] Kaiming He et al. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 9729–9738.

[83] Xinlei Chen et al. "Improved Baselines with Momentum Contrastive Learning". In: *arXiv:2003.04297* (2020).

[84] Xinlei Chen, Saining Xie, and Kaiming He. "An Empirical Study of Training Self-Supervised Vision Transformers". In: *arXiv:2104.02057* (2021).

[85] Debidatta Dwibedi et al. "With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations". In: *CVPR*. 2021.

[86] Jean-Bastien Grill et al. “Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[87] Xinlei Chen, Saining Xie, et al. “Exploring Simple Siamese Representation Learning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

[88] Jure Zbontar et al. “Barlow Twins: Self-Supervised Learning via Redundancy Reduction”. In: *ICML*. 2021.

[89] Adrien Bardes, Jean Ponce, and Yann LeCun. “VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning”. In: *ICLR*. 2022.

[90] Mathilde Caron et al. “Deep Clustering for Unsupervised Learning of Visual Features”. In: *ECCV*. 2018.

[91] Mathilde Caron et al. “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments”. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2020).

[92] Yuki M Asano, Christian Rupprecht, and Andrea Vedaldi. “Self-Labelling via Simultaneous Clustering and Representation Learning”. In: *ICLR*. 2020.

[93] Mathilde Caron et al. “Emerging Properties in Self-Supervised Vision Transformers”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 9650–9660.

[94] Maxime Oquab et al. “Dinov2: Learning robust visual features without supervision”. In: *arXiv preprint arXiv:2304.07193* (2023).

[95] Kaiming He et al. “Masked Autoencoders Are Scalable Vision Learners”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 16000–16009.

[96] Hang Bao et al. “BEiT: BERT Pre-Training of Image Transformers”. In: *International Conference on Learning Representations (ICLR)*. 2022.

[97] Zhiliang Peng et al. “BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers”. In: *arXiv:2208.06366* (2022).

[98] Jiahui Zhou et al. “iBOT: Image BERT Pre-Training with Online Tokenizer”. In: *ICLR*. 2022.

[99] Chen Wei et al. “Masked Feature Prediction for Self-Supervised Visual Pre-Training”. In: *CVPR*. 2022.

[100] Alexei Baevski et al. “data2vec: A General Framework for Self-Supervised Learning in Speech, Vision and Language”. In: *ICML*. 2022.

[101] Alexei Baevski et al. “Efficient Self-Supervised Learning with Contextualized Target Representations for Vision, Speech and Language (data2vec 2.0)”. In: *arXiv:2212.07525* (2023).

[102] Mahmoud Assran et al. “Self-Supervised Learning with a Joint-Embedding Predictive Architecture”. In: *CVPR*. 2023.

[103] Sagie Benaim et al. "Speednet: Learning the speediness in videos". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9922–9931.

[104] Rui Qian et al. "Spatiotemporal Contrastive Video Representation Learning". In: *CVPR*. 2021.

[105] Humam Alwassel et al. "Self-Supervised Learning by Cross-Modal Audio-Video Clustering". In: *NeurIPS*. 2020.

[106] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. "Audio-Visual Instance Discrimination with Cross-Modal Agreement". In: *CVPR*. 2021.

[107] Zhan Tong et al. "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training". In: *Advances in neural information processing systems* 35 (2022), pp. 10078–10093.

[108] Jian Wang et al. "Masked Autoencoders for Video with Dual Masking". In: *CVPR*. 2023.

[109] Rohit Girdhar et al. "OmniMAE: Single Model Masked Pretraining on Images and Videos". In: *CVPR*. 2023.

[110] Bingkun Huang et al. "Mgmae: Motion guided masking for video masked autoencoding". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 13493–13504.

[111] Rui Qian et al. "Masked Video Distillation: Rethinking Masked Feature Modeling for Self-Supervised Video Pre-Training". In: *CVPR*. 2023.

[112] Quentin Garrido et al. "SIGMA: Sinkhorn-Guided Masked Video Modeling". In: *arXiv:2407.17518* (2024).

[113] Inseop Hwang et al. "EVEREST: Efficient Masked Video Pre-training by Removing Redundant Tokens". In: *ICML*. 2024.

[114] Michael S. Ryoo, Rohit Girdhar, et al. "TokenLearner: What Can 8 Learned Tokens Do for Images and Videos?" In: *NeurIPS*. 2021.

[115] Daniel Bolya et al. "Token Merging: Your ViT But Faster". In: *arXiv:2210.09461* (2022).

[116] Adrien Bardes et al. *V-JEPA: Video Joint-Embedding Predictive Architecture*. GitHub repository and OpenReview preprint. `https://github.com/facebookresearch/jepa`. 2024.

[117] Mido Assran, Adrien Bardes, et al. "V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning". In: *arXiv:2506.09985* (2025).

[118] Yixiong Chen et al. "USCL: Pretraining Deep Ultrasound Image Diagnosis Model through Video Contrastive Representation Learning". In: *MICCAI*. 2021.

[119] Gregory Holste et al. "Efficient deep learning-based automated diagnosis from echocardiography with contrastive self-supervised learning". In: *Communications Medicine* (2024).

[120] Yiman Liu et al. "EDMAE: An Efficient Decoupled Masked Autoencoder for Standard View Identification in Pediatric Echocardiography". In: *Biomedical Signal Processing and Control* (2023). Also available as arXiv:2302.13869.

[121] Qingbo Kang et al. “Deblurring Masked Autoencoder is Better Recipe for Ultrasound Image Recognition”. In: *MICCAI*. 2023.

[122] Aimon Rahman and Vishal M. Patel. “UltraMAE: Multi-modal Masked Autoencoder for Ultrasound Pre-training”. In: *Proceedings of the 7th International Conference on Medical Imaging with Deep Learning (MIDL)*. Vol. 250. Proceedings of Machine Learning Research. 2024, pp. 1196–1206.

[123] Munawar Kim et al. “SurgMAE: Masked Autoencoders for Long Surgical Video Analysis”. In: *NeurIPS Workshop on Self-Supervised Learning*. 2023.

[124] Zhao Wang et al. “Foundation Model for Endoscopy Video Analysis via Large-Scale Self-Supervised Pre-Train”. In: *MICCAI*. 2023.

[125] Olga Patey et al. “Prenatal detection of congenital heart defects using the deep learning-based image and video analysis: protocol for Clinical Artificial Intelligence in Fetal Echocardiography (CAIFE), an international multicentre multidisciplinary study”. In: *BMJ Open* 15.6 (2025), e101263. URL: https://doi.org/10.1136/bmjopen-2025-101263.

[126] JS Carvalho et al. “ISUOG Practice Guidelines (updated): fetal cardiac screening”. In: *Ultrasound in Obstetrics & Gynecology* 61.6 (2023), pp. 788–803.

[127] Mary T Donofrio, Anita J Moon-Grady, Larry K Hornberger, et al. “Diagnosis and treatment of fetal cardiac disease: a scientific statement from the American Heart Association”. In: *Circulation* 129.21 (2014), pp. 2183–2242. URL: https://doi.org/10.1161/01.cir.0000437597.44550.5d.

[128] Olga Patey et al. “Prenatal detection of congenital heart defects using the deep learning-based image and video analysis: protocol for Clinical Artificial Intelligence in Fetal Echocardiography (CAIFE), an international multicentre multidisciplinary study”. In: *BMJ open* 15.6 (2025), e101263.

[129] David Ouyang et al. “Video-based AI for beat-to-beat assessment of cardiac function”. In: *Nature* 580.7802 (2020), pp. 252–256.

[130] Grant Duffy et al. “High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning”. In: *JAMA cardiology* 7.4 (2022), pp. 386–395.

[131] Bálint Magyar et al. “RVENet: a large echocardiographic dataset for the deep learning-based assessment of right ventricular function”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 569–583.

[132] Mohammad Sabokrou et al. “Adversarially learned one-class classifier for novelty detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3379–3388.

[133] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. “Exploring the limits of out-of-distribution detection”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 7068–7081.

[134] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[135] Jingkang Yang et al. “Generalized out-of-distribution detection: A survey”. In: *arXiv preprint arXiv:2110.11334* (2021).

[136] Zhi Zhou et al. "Step: Out-of-distribution detection in the presence of limited in-distribution labeled data". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 29168–29180.

[137] Terrance DeVries and Graham W Taylor. "Learning confidence for out-of-distribution detection in neural networks". In: *arXiv preprint arXiv:1802.04865* (2018).

[138] Théo Guénais et al. "Bacoun: Bayesian classifers with out-of-distribution uncertainty". In: *arXiv preprint arXiv:2007.06096* (2020).

[139] Dan Hendrycks and Kevin Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks". In: *arXiv preprint arXiv:1610.02136* (2016).

[140] Jie Ren et al. "Likelihood ratios for out-of-distribution detection". In: *Advances in neural information processing systems* 32 (2019).

[141] Haowen Xu et al. "Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications". In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 187–196.

[142] Xiaoran Chen and Ender Konukoglu. "Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders". In: *arXiv preprint arXiv:1806.04972* (2018).

[143] Yibo Zhou. "Rethinking reconstruction autoencoder-based out-of-distribution detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 7379–7387.

[144] Thomas Schlegl et al. "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery". In: *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*. Springer. 2017, pp. 146–157.

[145] Eric Nalisnick et al. "Do deep generative models know what they don't know?" In: *arXiv preprint arXiv:1810.09136* (2018).

[146] Hyunsun Choi, Eric Jang, and Alexander A Alemi. "Waic, but why? generative ensembles for robust anomaly detection". In: *arXiv preprint arXiv:1810.01392* (2018).

[147] Stanislav Fort. "Adversarial vulnerability of powerful near out-of-distribution detection". In: *arXiv preprint arXiv:2201.07012* (2022).

[148] Yoav Wald et al. "On calibration and out-of-domain generalization". In: *Advances in neural information processing systems* 34 (2021), pp. 2215–2227.

[149] Mark S Graham et al. "Denoising Diffusion Models for Out-of-Distribution Detection". In: *arXiv preprint arXiv:2211.07740* (2022).

[150] Martin Arjovsky and Léon Bottou. "Towards principled methods for training generative adversarial networks". In: *arXiv preprint arXiv:1701.04862* (2017).

[151] David Bau et al. "Seeing what a gan cannot generate". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 4502–4511.

[152] Prafulla Dhariwal and Alexander Nichol. "Diffusion models beat gans on image synthesis". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 8780–8794.

[153] Ling Yang et al. "Diffusion models: A comprehensive survey of methods and applications". In: *arXiv preprint arXiv:2209.00796* (2022).

[154] Chenlin Meng et al. "Sdedit: Guided image synthesis and editing with stochastic differential equations". In: *International Conference on Learning Representations*. 2021.

[155] Chitwan Saharia et al. "Palette: Image-to-image diffusion models". In: *ACM SIGGRAPH 2022 Conference Proceedings*. 2022, pp. 1–10.

[156] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.

[157] Luping Liu et al. "Pseudo numerical methods for diffusion models on manifolds". In: *arXiv preprint arXiv:2202.09778* (2022).

[158] Julian Wyatt et al. "Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 650–656.

[159] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[160] Amir Hertz et al. "Prompt-to-prompt image editing with cross attention control". In: *arXiv preprint arXiv:2208.01626* (2022).

[161] Daniel Rebain et al. "Attention Beats Concatenation for Conditioning Neural Fields". In: *arXiv preprint arXiv:2209.10684* (2022).

[162] Katerina Margatina, Christos Baziotis, and Alexandros Potamianos. "Attention-based conditioning methods for external knowledge integration". In: *arXiv preprint arXiv:1906.03674* (2019).

[163] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models". In: *arXiv preprint arXiv:2010.02502* (2020).

[164] Dongha Lee, Sehun Yu, and Hwanjo Yu. "Multi-class data description for out-of-distribution detection". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 1362–1370.

[165] LJ Salomon et al. "ISUOG Practice Guidelines (updated): performance of the routine mid-trimester fetal ultrasound scan". In: *Ultrasound in obstetrics & gynecology: the official journal of the International Society of Ultrasound in Obstetrics and Gynecology* 59.6 (2022), pp. 840–856.

[166] Mary L McHugh. "Interrater reliability: the kappa statistic". In: *Biochemia medica* 22.3 (2012), pp. 276–282.

[167] Albert Gordo et al. "Deep image retrieval: Learning global representations for image search". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer. 2016, pp. 241–257.

[168] Seongwon Lee et al. "Revisiting Self-Similarity: Structural Embedding for Image Retrieval". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 23412–23421.

[169] Aneeshan Sain et al. "Clip for all things zero-shot sketch-based image retrieval, fine-grained or not". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2765–2775.

[170] Tan Pan et al. "Boundary-aware Backward-Compatible Representation via Adversarial Learning in Image Retrieval". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15201–15210.

[171] Mengmeng Xu et al. "Negative frames matter in egocentric visual query 2d localization". In: *arXiv preprint arXiv:2208.01949* (2022).

[172] Hanwen Jiang, Santhosh Kumar Ramakrishnan, and Kristen Grauman. "Single-Stage Visual Query Localization in Egocentric Videos". In: *arXiv preprint arXiv:2306.09324* (2023).

[173] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[174] Jie Lei, Tamara L Berg, and Mohit Bansal. "Detecting moments and highlights in videos via natural language queries". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 11846–11858.

[175] WonJun Moon et al. "Query-dependent video representation for moment retrieval and highlight detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 23023–23033.

[176] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. "Learning spatio-temporal features with 3d residual networks for action recognition". In: *Proceedings of the IEEE international conference on computer vision workshops*. 2017, pp. 3154–3160.

[177] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[178] Runhao Zeng et al. "Dense regression network for video grounding". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 10287–10296.

[179] Kevin Qinghong Lin et al. "Univtg: Towards unified video-language temporal grounding". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 2794–2804.

[180] Ye Liu et al. "Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 3042–3051.

[181] Bahbibi Rahmatullah, Aris Papageorghiou, and J Alison Noble. "Automated selection of standardized planes from ultrasound volume". In: *Machine Learning in Medical Imaging: Second International Workshop, MLMI 2011, Held in Conjunction with MICCAI 2011, Toronto, Canada, September 18, 2011. Proceedings 2*. Springer. 2011, pp. 35–42.

[182] Mohammad Yaqub et al. "Guided Random Forests for Identification of Key Fetal Anatomy and Image Categorization in Ultrasound Scans". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab et al. Cham: Springer International Publishing, 2015, pp. 687–694.

[183] Hao Chen et al. "Standard plane localization in fetal ultrasound via domain transferred deep neural networks". In: *IEEE journal of biomedical and health informatics* 19.5 (2015), pp. 1627–1636.

[184] Hao Chen et al. "Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*. Springer. 2015, pp. 507–514.

[185] Hao Chen et al. "Ultrasound standard plane detection using a composite neural network framework". In: *IEEE transactions on cybernetics* 47.6 (2017), pp. 1576–1586.

[186] Zeyu Fu et al. "Anatomy-Aware Contrastive Representation Learning for Fetal Ultrasound". In: *Computer Vision – ECCV 2022 Workshops*. Ed. by Leonid Karlinsky, Tomer Michaeli, and Ko Nishino. Cham: Springer Nature Switzerland, 2023, pp. 422–436.

[187] Qianhui Men et al. "Towards Standard Plane Prediction of Fetal Head Ultrasound with Domain Adaption". In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2023, pp. 1–5.

[188] He Zhao et al. "Towards unsupervised ultrasound video clinical quality assessment with multi-modality data". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 228–237.

[189] He Zhao et al. "Memory-based unsupervised video clinical quality assessment with multi-modality data in fetal ultrasound". In: *Medical Image Analysis* 90 (2023), p. 102977.

[190] Yifan Cai et al. "Multi-task sonoeyenet: detection of fetal standardized planes assisted by generated sonographer attention maps". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer. 2018, pp. 871–879.

[191] Lok Hin Lee, Yuan Gao, and J Alison Noble. "Principled ultrasound data augmentation for classification of standard planes". In: *International Conference on Information Processing in Medical Imaging*. Springer. 2021, pp. 729–741.

[192] Christian F Baumgartner et al. "Real-time standard scan plane detection and localisation in fetal ultrasound using fully convolutional neural networks". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer. 2016, pp. 203–211.

[193] Jo Schlemper et al. "Attention-gated networks for improving ultrasound scan plane detection". In: *arXiv preprint arXiv:1804.05338* (2018).

[194] Harshita Sharma et al. "Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos". In: *Medical Image Analysis* 69 (2021), p. 101973.

[195] Robail Yasrab et al. "A machine learning method for automated description and workflow analysis of first trimester ultrasound scans". In: *IEEE Transactions on Medical Imaging* 42.5 (2022), pp. 1301–1313.

[196] Jiyang Gao et al. "Tall: Temporal activity localization via language query". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5267–5275.

[197] Lisa Anne Hendricks et al. "Localizing moments in video with temporal language". In: *arXiv preprint arXiv:1809.01337* (2018).

[198] Meng Liu et al. "Cross-modal moment localization in videos". In: *Proceedings of the 26th ACM international conference on Multimedia*. 2018, pp. 843–851.

[199] Shaoning Xiao et al. "Boundary proposal network for two-stage natural language video localization". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 4. 2021, pp. 2986–2994.

[200] Huijuan Xu et al. "Multilevel language and vision integration for text-to-clip retrieval". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 9062–9069.

[201] Daizong Liu et al. "Context-aware biaffine localizing network for temporal sentence grounding". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 11235–11244.

[202] Da Zhang et al. "Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 1247–1257.

[203] Long Chen et al. "Rethinking the bottom-up framework for query-based video localization". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 10551–10558.

[204] Daizong Liu et al. "Unsupervised temporal video grounding with deep semantic clustering". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 2. 2022, pp. 1683–1691.

[205] Jonghwan Mun, Minsu Cho, and Bohyung Han. "Local-global video-text interactions for temporal grounding". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 10810–10819.

[206] Yitian Yuan, Tao Mei, and Wenwu Zhu. "To find where you talk: Temporal sentence localization in video with attention based location regression". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 9159–9166.

[207] Hao Zhang et al. "Span-based localizing network for natural language video localization". In: *arXiv preprint arXiv:2004.13931* (2020).

[208] Wenhai Wang et al. "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions". In: *CVPR 2021*.

[209] Haoqi Fan et al. "Multiscale vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 6824–6835.

[210] Yanghao Li et al. "Mvitv2: Improved multiscale vision transformers for classification and detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 4804–4814.

[211] Weifeng Ge. "Deep metric learning with hierarchical triplet loss". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 269–285.

[212] Sumit Chopra, Raia Hadsell, and Yann LeCun. "Learning a similarity metric discriminatively, with application to face verification". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. IEEE. 2005, pp. 539–546.

[213] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823.

[214] Malik Boudiaf et al. "A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses". In: *European conference on computer vision*. Springer. 2020, pp. 548–564.

[215] Divyanshu Mishra et al. " STAN-LOC: Visual Query-based Video Clip Localization for Fetal Ultrasound Sweep Videos ". In: *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Vol. LNCS 15004. Springer Nature Switzerland, Oct. 2024.

[216] Lan Wang et al. "ProTeGe: Untrimmed Pretraining for Video Temporal Grounding by Video Temporal Grounding". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 6575–6585.

[217] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL-HLT*. 2019, pp. 4171–4186.

[218] Netzahualcoyotl Hernandez-Cruz et al. "A comprehensive scoping review on machine learning-based fetal echocardiography analysis". In: *Computers in Biology and Medicine* 186 (2025), p. 109666.

[219] Jong Hak Moon et al. "Multi-modal understanding and generation for medical images and text via vision-language pre-training". In: *IEEE Journal of Biomedical and Health Informatics* 26.12 (2022), pp. 6070–6080.

[220] Pramit Saha et al. "Examining Modality Incongruity in Multimodal Federated Learning for Medical Vision and Language-based Disease Detection". In: *arXiv preprint arXiv:2402.05294* (2024).

[221] Pramit Saha et al. "FedPIA–Permuting and Integrating Adapters leveraging Wasserstein Barycenters for Finetuning Foundation Models in Multi-Modal Federated Learning". In: *arXiv preprint arXiv:2412.14424* (2024).

[222] Pramit Saha et al. "F^{3}OCUS–Federated Finetuning of Vision-Language Foundation Models with Optimal Client Layer Updating Strategy via Multi-objective Meta-Heuristics". In: *arXiv preprint arXiv:2411.11912* (2024).

[223] Ted E Scott et al. "Increasing the detection rate of congenital heart disease during routine obstetric screening using cine loop sweeps". In: *Journal of Ultrasound in Medicine* 32.6 (2013), pp. 973–979.

[224] Denise Van Der Linde et al. “Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis”. In: *Journal of the American College of Cardiology* 58.21 (2011), pp. 2241–2247.

[225] Asma Khalil and Kypros H Nicolaides. “Fetal heart defects: potential and pitfalls of first-trimester detection”. In: *Seminars in fetal and neonatal medicine*. Vol. 18. 5. Elsevier. 2013, pp. 251–260.

[226] JS Carvalho et al. “Improving the effectiveness of routine prenatal screening for major congenital heart defects”. In: *Heart* 88.4 (2002), pp. 387–391.

[227] Rolf Becker and R-D Wegner. “Detailed screening for fetal anomalies and cardiac defects at the 11–13-week scan”. In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 27.6 (2006), pp. 613–618.

[228] Tim Van Mieghem et al. “Methods for prenatal assessment of fetal cardiac function”. In: *Prenatal Diagnosis: Published in Affiliation With the International Society for Prenatal Diagnosis* 29.13 (2009), pp. 1193–1203.

[229] Guang Yu et al. “Deep anomaly discovery from unlabeled videos via normality advantage and self-paced refinement”. In: *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 2022, pp. 13987–13998.

[230] Seongheon Park et al. “Normality guided multiple instance learning for weakly supervised video anomaly detection”. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2023, pp. 2665–2674.

[231] Mengyang Zhao et al. “Lgn-net: local-global normality network for video anomaly detection”. In: *arXiv preprint arXiv:2211.07454* (2022).

[232] Sunghyun Ahn et al. “VideoPatchCore: An Effective Method to Memorize Normality for Video Anomaly Detection”. In: *Proceedings of the Asian Conference on Computer Vision*. 2024, pp. 2179–2195.

[233] Parisasadat Shojaei, Elena Vlahu-Gjorgievska, and Yang-Wai Chow. “Security and privacy of technologies in health information systems: A systematic literature review”. In: *Computers* 13.2 (2024), p. 41.

[234] DaDong Jiang et al. “TimeFormer: Capturing Temporal Relationships of Deformable 3D Gaussians for Robust Reconstruction”. In: *arXiv preprint arXiv:2411.11941* (2024).

[235] AJ Piergiovanni, Weicheng Kuo, and Anelia Angelova. “Rethinking video vits: Sparse video tubes for joint image and video learning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2214–2224.

[236] Mathilde Caron et al. “Emerging properties in self-supervised vision transformers”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660.

[237] Mitchell Wortsman et al. “Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time”. In: *International conference on machine learning*. PMLR. 2022, pp. 23965–23998.

[238] Gabriel Ilharco et al. “Editing models with task arithmetic”. In: *The Eleventh International Conference on Learning Representations*.

[239] Prateek Yadav et al. "Ties-merging: Resolving interference when merging models". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 7093–7115.

[240] Le Yu et al. "Language models are super mario: Absorbing abilities from homologous models as a free lunch". In: *Forty-first International Conference on Machine Learning*. 2024.

[241] Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. "Model stock: All we need is just a few fine-tuned models". In: *European Conference on Computer Vision*. Springer. 2024, pp. 207–223.

[242] Jhih-Ciang Wu et al. "Self-supervised sparse representation for video anomaly detection". In: *European Conference on Computer Vision*. Springer. 2022, pp. 729–745.

[243] RM Lang, LP Badano, V Mor-Avi, et al. "Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging". In: *J. Am. Soc. Echocardiogr* 28.1 (2015), pp. 1–39.

[244] Yutong Chen et al. "Self-supervised Spatial-Temporal Normality Learning for Time Series Anomaly Detection". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2024, pp. 145–162.

[245] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. "Masked autoencoders as spatiotemporal learners". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 35946–35958.

[246] David Fan et al. "Motion-guided masking for spatiotemporal representation learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 5619–5629.

[247] Kaiming He et al. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.

[248] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation Learning with Contrastive Predictive Coding". In: *arXiv preprint arXiv:1807.03748*. 2018.

[249] Mehdi Noroozi and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles". In: *European conference on computer vision*. Springer. 2016, pp. 69–84.

[250] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations". In: *arXiv preprint arXiv:1803.07728* (2018).

[251] Richard Zhang, Phillip Isola, and Alexei A Efros. "Colorful image colorization". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer. 2016, pp. 649–666.

[252] Zhirong Wu et al. "Unsupervised feature learning via non-parametric instance discrimination". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3733–3742.

[253] Ting Chen et al. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PmLR. 2020, pp. 1597–1607.

[254] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. "Shuffle and learn: unsupervised learning using temporal order verification". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer. 2016, pp. 527–544.

[255] Dejing Xu et al. "Self-supervised spatiotemporal learning via video clip order prediction". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 10334–10343.

[256] Dahun Kim, Donghyeon Cho, and In So Kweon. "Self-supervised video representation learning with space-time cubic puzzles". In: *Proceedings of the AAAI conference on artificial intelligence*. 2019, pp. 8545–8552.

[257] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. "Self-supervised video representation learning by pace prediction". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer. 2020, pp. 504–521.

[258] Mohammadreza Salehi et al. "SIGMA: Sinkhorn-Guided Masked Video Modeling". In: *European Conference on Computer Vision*. Springer. 2024, pp. 293–312.

[259] Jianbo Jiao et al. "Self-supervised representation learning for ultrasound video". In: *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*. IEEE. 2020, pp. 1847–1850.

[260] Hadrien Reynaud et al. "EchoFlow: A Foundation Model for Cardiac Ultrasound Image and Video Generation". In: *arXiv preprint arXiv:2503.22357* (2025).

[261] Cleveland Clinic. *Ejection Fraction: What It Is, Types and Normal Range*. Accessed: 2025-05-01. 2023. URL: `https://my.clevelandclinic.org/health/articles/16950-ejection-fraction`.

[262] Sarah Leclerc et al. "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography". In: *IEEE transactions on medical imaging* 38.9 (2019), pp. 2198–2210.

[263] Rui Wang et al. "Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 6312–6322.

[264] Fernando Pérez-García et al. "Rad-dino: Exploring scalable medical image encoders beyond text supervision". In: *arXiv preprint arXiv:2401.10815* (2024).

[265] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. "Learning memory-guided normality for anomaly detection". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 14372–14381.

[266] Dong Gong et al. "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1705–1714.

[267] Anas Al-Lahham et al. “A coarse-to-fine pseudo-labeling (c2fpl) framework for unsupervised video anomaly detection”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 6793–6802.

[268] Walid Al-Dhabyani et al. “Dataset of breast ultrasound images”. In: *Data in brief* 28 (2020), p. 104863.

[269] Jiancheng Yang et al. “Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification”. In: *Scientific Data* 10.1 (2023), p. 41.

[270] Xinyu Sun et al. “Masked motion encoding for self-supervised video representation learning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 2235–2245.

[271] RL Knowles and RM Hunter. “Screening for congenital heart defects: external review against programme appraisal criteria for the UK NSC”. In: (2014).

[272] World Health Organization. *Cardiovascular diseases (CVDs): Fact Sheet*. `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`. Accessed: October 2023. 2021.