Medical Image Analysis (2025)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

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

Divyanshu Mishra[a,*], Pramit Saha[a], He Zhao[b], Netzahualcoyotl Hernandez-Cruz[a], Olga Patey[c], Aris T. Papageorghiou[c], J. Alison Noble[a]

[a]*Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, OX3 7DQ, United Kingdom*
[b]*Institute of Life Course and Medical Sciences University of Liverpool, William Henry Duncan Building, 6 West Derby Street, Liverpool, L7 8TX*
[c]*Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, OX3 9DU, United Kingdom*

ARTICLE INFO

Article history:

Keywords:
Fetal Ultrasound,
Video Understanding,
Video Transformers,
Multi-Modal Video Understanding
Visual-Query

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 utilise 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.

© 2025 Elsevier B. V. All rights reserved.

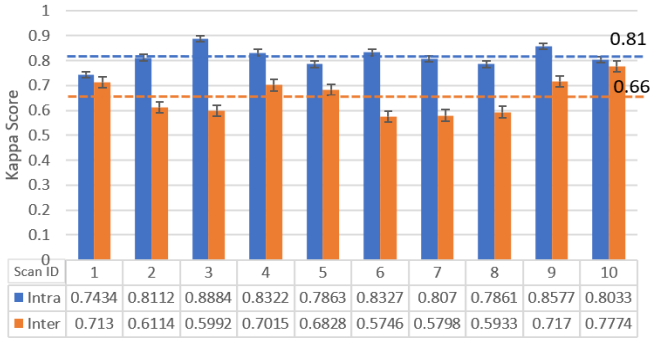
*Corresponding author at: Department of Engineering Science, University of Oxford, Oxford, United Kingdom.
e-mail: divyanshu.mishra@eng.ox.ac.uk (Divyanshu Mishra)

1. Introduction

Text query-based localization tasks, including video-temporal grounding Yang et al. (2022); Zeng et al. (2020); Lin et al. (2023), video moment retrieval Liu et al. (2022b); Moon et al. (2023), 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 Gordo et al. (2016); Lee et al. (2023); Sain et al. (2023); Pan et al. (2023) and (b) visual query-based 2D localization (VQ2D) Jiang et al. (2023); Goyal et al. (2023); Xu et al. (2023, 2022). However, both lines of work output a single image or utilise datasets with more distinct classes as shown in Fig. 2. In the case of ultrasound video, retrieving a video clip rather than a single frame is more challenging as observed in Fig. 2 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 Carvalho et al. (2013); Salomon et al. (2022)). However, this process is time-consuming. Integrating a video-clip



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

Fig. 1. 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

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.2. 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. 1 that depicts a study where two cardiologists were asked to annotate the same 10 fetal heart videos. The kappa score McHugh (2012) 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 Rahmatullah et al. (2011); Yaqub et al. (2015); Baumgartner et al. (2017); Chen et al. (2015b,a, 2017) or self-supervised Fu et al. (2023) 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 Men et al. (2023). That approach has not yet been generalized to multiple standard planes. A video-based ultrasound analysis method to de-

tect clinical high-quality video clips containing standard planes is proposed in Zhao et al. (2022, 2023). 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 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.
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 Grauman et al. (2022) dataset and evaluate our model's performance on it to allow benchmarking with respect to our approach by others in the future.

2. Related Work

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 Rahmatullah et al. (2011); Cai et al. (2018); Lee et al. (2021); Baumgartner et al. (2016, 2017); Schlemper et al. (2018). 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, Zhao et al. (2022, 2023) leverage the temporal information of ultrasound videos and develop a video-based anomaly detection method. This method identifies clinical high-quality video clips containing standard planes by using reconstruction error as the indicator. Men et al. (2023) 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.

2.2. Clinical Workflow Analysis:

In the context of automated fetal ultrasound clinical workflow analysis, first introduced by Sharma et al. (2021) 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. Sharma et al. (2021) introduced a (2D + T) CNN architecture, termed Sono-2Dt-CNN, which was used for automated video classification. Building upon this approach, Yasrab et al. (2022) 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

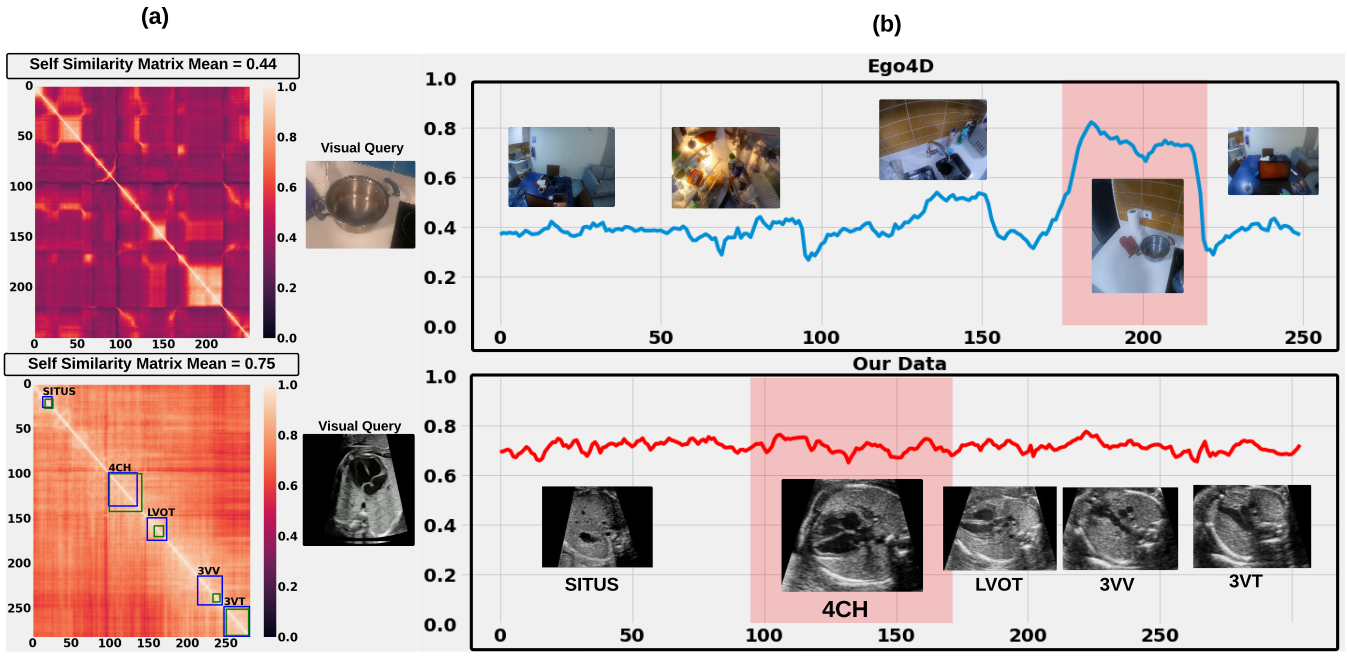


Fig. 2. (a) Self-similarity matrix for a randomly chosen video from Ego4D (top, mean=0.44) Grauman et al. (2022) and our clinical video dataset (bottom, mean=0.75), which reveals higher task difficulty for our video clip localization task. The uncertainty in annotations of two expert cardiologists is 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.

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.

2.3. Visual Query 2D Localization (VQ2D)

Introduced in Ego4D's episodic memory benchmark Grauman et al. (2022), 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 Grauman et al. (2022) and its improved variant Xu et al. (2023) treat spatial and temporal reasoning separately. In contrast, single-stage approaches—e.g., MINOTAUR Goyal et al. (2023) (multi-task) and VQLOC Jiang et al. (2023) (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.

2.4. Video Temporal Grounding

Video Temporal Grounding (VTG) introduced in Gao et al. (2017); Hendricks et al. (2018) aims to localize a target moment by predicting start and end frame numbers according to a given natural language query. Earlier methods focussed on generating moment proposals using sliding windows Gao et al. (2017); Liu et al. (2018), proposal generation networks Xiao et al. (2021); Xu et al. (2019) or utilised predefined anchors Liu et al. (2021); Zhang et al. (2019). 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 Chen et al. (2020); Liu et al. (2022a); Mun et al. (2020); Yuan et al. (2019); Zhang et al. (2020); Yang et al. (2022) 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 utilises a single image as a visual query rather than a text query.

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 Wang et al.; Fan et al.; Li et al. (2022) 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

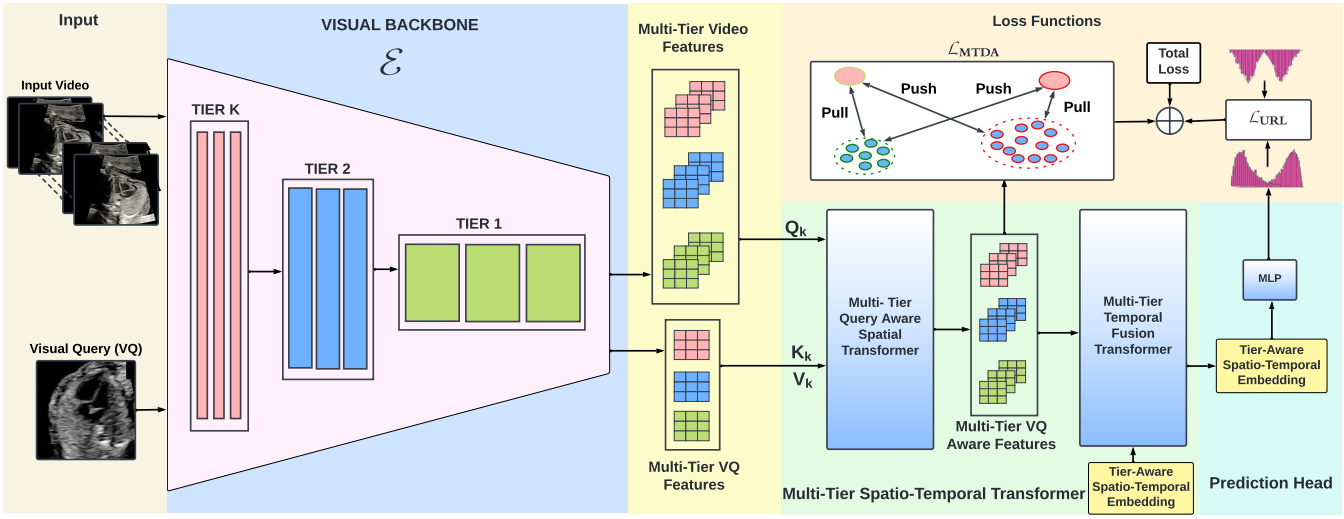


Fig. 3. Main architecture of TIER-LOC. The input video v and visual query q are passed to the visual backbone to give multi-Tier features. These features are fused spatially using the Multi-Tier Query Aware Spatial Transformer. The Tier-specific features are passed to a) $\mathcal{L}_{MTDA}$ to learn 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.

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.

2.6. Metric Learning

Existing metric learning methods Ge (2018); Chopra et al. (2005); Schroff et al. (2015); Boudiaf et al. (2020) 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.

3. Methods

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.

3.2. TIER-LOC Overall Architecture

Our proposed model, as depicted in Fig. 3, 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. 3. 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 3.4.

3.3. Multi-Tier Spatio-Temporal Transformer:

3.3.1. 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 Yang et al. (2022) and moment retrieval Lei et al. (2021) 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 Moon et al. (2023). 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 Vaswani et al. (2017). 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 Vaswani et al. (2017) 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. 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.

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

3.3.2. 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 Vaswani et al. (2017) 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. 2. This helps in fusing the spatial and temporal information available across the Tiers into the Tier-Aware Spatio-Temporal Embedding which is then utilised by an MLP to make the final predictions.

$$E_{T_A} = FFN\left(\text{softmax}\left(\frac{Q_v\left(K_v^T\right)}{\sqrt{d_k}}\right)V_v\right) \tag{2}$$

3.4. Loss Functions

3.4.1. 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 the challenges are further amplified by high spatial similarity among video frames (see Fig. 2). 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 utilises 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 utilise 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. 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. 3 where τ^+ is the positive temperature parameter.

$$\mathcal{L}_{PAC} = -\log \frac{\sum_{k=1}^{K-1}\sum_{i=1}^{M_1}\exp\left(sim(f'^{+}_{vq_a}, f'^{+}_{vq_{k,i}})/\tau^+\right)}{\sum_{k=1}^{K-1}\sum_{j=1}^{M_2}\exp\left(sim(f'^{+}_{vq_a}, f'^{-}_{vq_{k,j}})/\tau^+\right)} \tag{3}$$

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. 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. 4 where τ^- is temperature for $\mathcal{L}_{NAC}$.

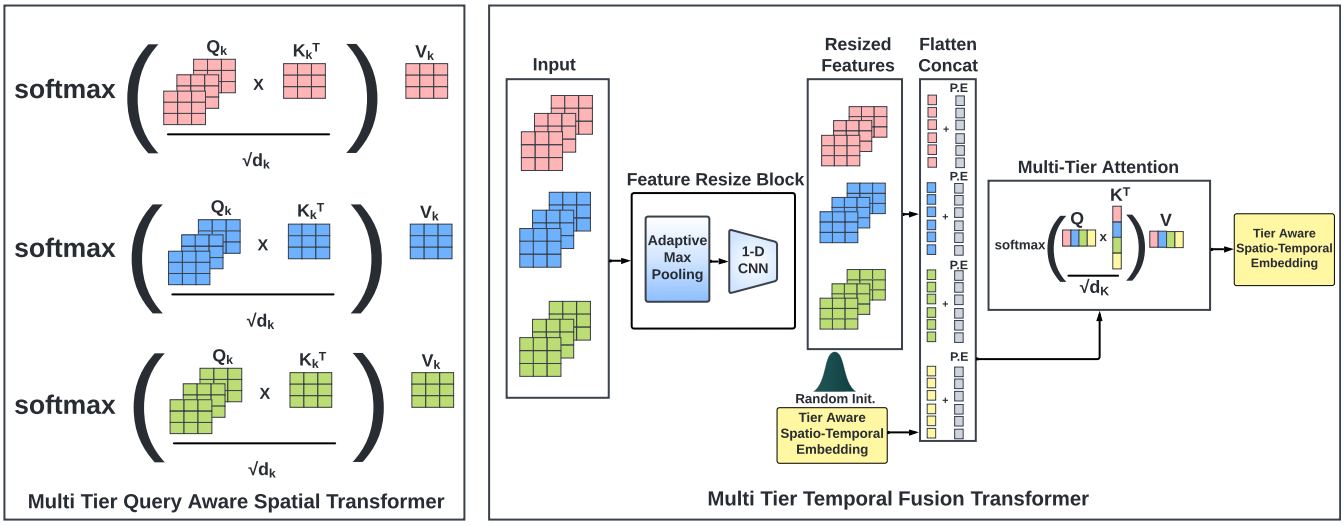


Fig. 4. Fig (left) shows the spatial feature fusion mechanism where Tier-specific video and VQ features are spatially fused to give Tier-specific query-aware features. Figure 4 (right) shows how the Tier-specific query-aware features are first resized, flattened and enriched with positional information.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

$$\mathcal{L}_{NAC} = -\log \frac{\sum_{k=1}^{K-1}\sum_{i=1}^{M_2} \exp\left(sim(f'^{-}_{vq_a}, f'^{-}_{vq_{k,i}})/\tau^{-}\right)}{\sum_{k=1}^{K-1}\sum_{j=1}^{M_1} \exp\left(sim(f'^{-}_{vq_a}, f'^{+}_{vq_{k,j}})/\tau^{-}\right)} \tag{4}$$

The final loss $\mathcal{L}_{MTDA}$ is denoted in Eqn. 5 where w_p and w_n are tunable weights for $\mathcal{L}_{PAC}$ and $\mathcal{L}_{NAC}$ respectively.

$$\mathcal{L}_{MTDA} = w_p * \mathcal{L}_{PAC} + w_n * \mathcal{L}_{NAC} \tag{5}$$

3.4.2. 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 . Finally, we optimise 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 and 8 respectively.

$$T_s(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu_s)^2/2\sigma^2}, \; T_e(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu_e)^2/2\sigma^2} \tag{6}$$

$$\begin{aligned} KL_s(P_s||T_s) &= \sum_x P_s(x)\log(\frac{P_s(x)}{T_s(x)}), \\ KL_e(P_e||T_e) &= \sum_x P_e(x)\log(\frac{P_e(x)}{T_e(x)}) \end{aligned} \tag{7}$$

$$\mathcal{L}_{URL} = KL_s + KL_e \tag{8}$$

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. 9.

$$\mathcal{L} = \mathcal{L}_{MTDA} + \mathcal{L}_{URL} \tag{9}$$

4. Experiments and Results

4.1. Dataset and Implementation

Following Mishra et al. (2024), we use ultrasound datasets from two research studies PULSE (Perception Ultrasound by Learning Sonographer Experience) Drukker et al. (2021), and CAIFE (Development of Clinical Artificial Intelligence Models in Fetal Echocardiography for the Detection of Congenital Heart Defects). Details are described next.

4.1.1. 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 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 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

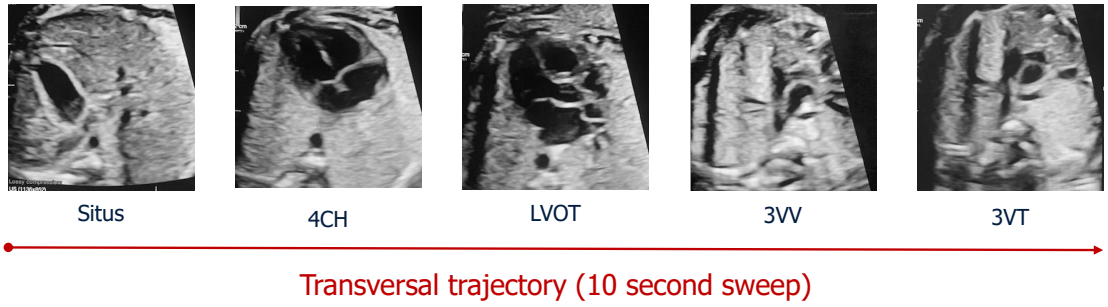


Fig. 5. Technique for scanning fetal heart through transversal trajectory. (I) Situs view of the upper abdomen is visualized first. (II) The four-chamber view is obtained through an axial scanning plane across the fetal chest by moving and tilting the transducer in a cephalad direction. Further, cephalad 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.

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.

4.1.2. PULSE dataset:

The PULSE dataset Drukker et al. (2021) consists of free-hand ultrasound (US) videos including participants aged > 18 years in their second trimester of pregnancy (≈ 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 × 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.

4.1.3. Ego4D VQ-VCL

For the VQ-VCL task, open-source computer vision datasets are absent. Hence, we utilize the existing Ego4D Grauman et al. (2022) 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. 3 and 4) and weighting factors in their sum (Eq. 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 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 1. Grid search hyperparameter ranges used in hyperparameter optimization.

4.1.4. Metrics

To quantify the performance of TIER-LOC compared to previous work on temporal video-grounding Wang et al. (2023); Moon et al. (2023) and our baselines Goyal et al. (2023); Jiang et al. (2023), 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.

4.1.5. Quantative Results

We compare our model TIER-LOC with ResNet3D CNN Hara et al. (2017), Cosine Similarity Supervised 2D CNN, TubeDETR Yang et al. (2022), VQLOC Jiang et al. (2023) and MomentDETRLei et al. (2021) as shown in Table 2.

1.ResNet 3D Hara et al. (2017) We concatenate video frames and the visual query along the channel dimension and pass it to

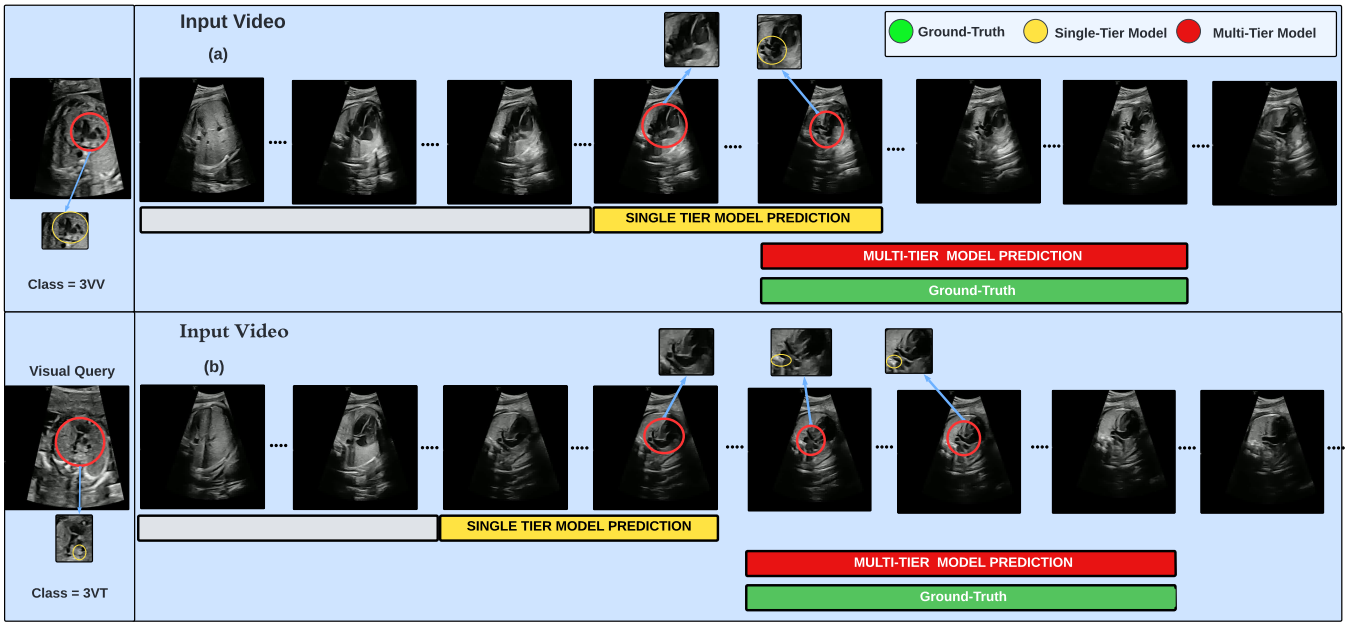


Fig. 6. Figure showing the importance of the Multi-Tier Feature Fusion Transformer. We show that the model with a Single TIER transformer fails to detect minute local patterns (shown in patches above). For instance, in (a), the only difference between a 3VV view and the view before (LVOT) is the appearance 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.

Table 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
TubeDETRYang et al. (2022)	12.72	2.00	2.00	10.22	20.00
MomentDETRLei et al. (2021)	14.89	0.00	8.00	25.00	39.72
Resnet 3D Hara et al. (2017)	19.79	6.00	6.00	23.22	47.17
VQLOC Jiang et al. (2023)	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 Drukker et al. (2021)					
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
TubeDETRYang et al. (2022)	7.07	4.76	4.76	4.76	14.42
MomentDETRLei et al. (2021)	10.34	2.04	6.93	16.60	21.50
Resnet 3D Hara et al. (2017)	17.89	14.29	17.14	22.04	**28.16**
VQLOC Jiang et al. (2023)	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 Grauman et al. (2022)					
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 Hara et al. (2017)	10.72	3.57	8.33	12.70	20.24
VQLOC Jiang et al. (2023)	25.35	7.14	19.44	32.94	44.84
MomentDETR Lei et al. (2021)	38.44	15.08	25.40	52.78	66.67
TubeDETR Yang et al. (2022)	38.59	18.65	32.94	61.51	71.03
TIER-LOC (Ours)	**43.01**	**23.81**	**34.92**	**70.63**	**73.02**

the model. We train the model with our L_{URL} loss.

2. Cosine Similarity Supervision CNN: We use ResNet 101 He et al. (2016) 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 Yang et al. (2022) 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 Jiang et al. (2023) 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 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 Moment-DETR, 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 Drukker et al. (2021) data (refer Table 2) and Ego4D VQ-VCL (refer Table 2) with our model TIER-LOC outperforming the best-performing baseline by 4.00% and 4.42 % respectively.

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(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(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. 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. 7) , which the single-tier model misses. Our model's ability to detect this feature leads to a significant boost in localization performance.

4.3. Ablation Study

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

4.3.1. 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

Table 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

Table 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**

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 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.

4.3.2. Importance of different Loss functions

In Table 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 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 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**

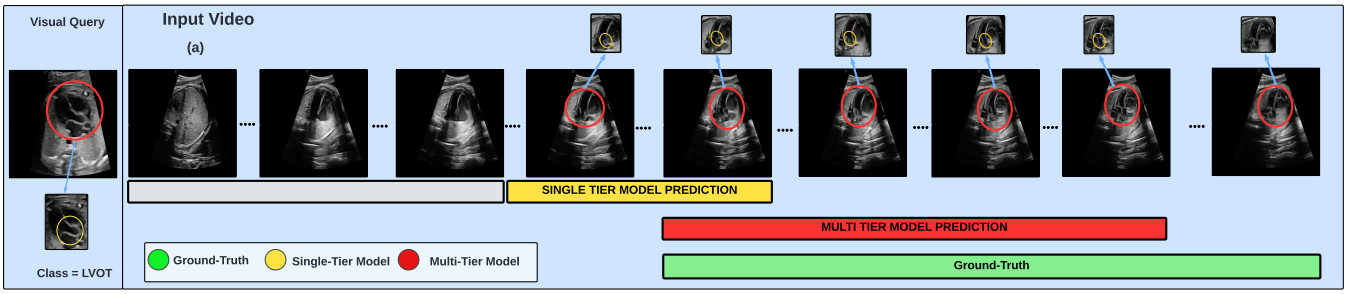


Fig. 7. This figure compares the predictions of a single-tier model with those of our multi-tier model for LVOT visual query. Our multi-tier model excels at detecting subtle differences, resulting in precise boundary predictions. In the given example, the single-tier model struggles to differentiate 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. 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**

4.3.3. 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 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.

4.3.4. Sequential vs Parallel Fusion

In existing works utilising multi-scale features Wang et al.; Fan et al. 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 where parallel fusion outperforms sequential fusion by 10% mtIoU.

4.3.5. 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 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 He et al. (2016) 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 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

4.3.6. 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 Yang et al. (2022); Devlin et al. (2018); Lei et al. (2021) and Cross-Attention feature fusion techniques. In Table 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 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**

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.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge financial support InnoHK-funded Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE) and from UKRI grant EP/X040186/1, UK EPSRC (Engineering and Physical Research Council) grant EP/T028572/1 (VisualAI). PS is supported by a UK EPSRC Doctoral Training Partnership award.

References

Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D., 2017. Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE transactions on medical imaging 36, 2204–2215.

Baumgartner, C.F., Kamnitsas, K., Matthew, J., Smith, S., Kainz, B., Rueckert, D., 2016. 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. pp. 203–211.

Boudiaf, M., Rony, J., Ziko, I.M., Granger, E., Pedersoli, M., Piantanida, P., Ayed, I.B., 2020. A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses, in: European conference on computer vision, Springer. pp. 548–564.

Cai, Y., Sharma, H., Chatelain, P., Noble, J.A., 2018. 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. pp. 871–879.

Carvalho, J.S., Allan, L., Chaoui, R., Copel, J., DeVore, G., Hecher, K., Lee, W., Munoz, H., Paladini, D., Tutschek, B., et al., 2013. Isuog practice guidelines (updated): sonographic screening examination of the fetal heart .

Chen, H., Dou, Q., Ni, D., Cheng, J.Z., Qin, J., Li, S., Heng, P.A., 2015a. 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. pp. 507–514.

Chen, H., Ni, D., Qin, J., Li, S., Yang, X., Wang, T., Heng, P.A., 2015b. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. IEEE journal of biomedical and health informatics 19, 1627–1636.

Chen, H., Wu, L., Dou, Q., Qin, J., Li, S., Cheng, J.Z., Ni, D., Heng, P.A., 2017. Ultrasound standard plane detection using a composite neural network framework. IEEE transactions on cybernetics 47, 1576–1586.

Chen, L., Lu, C., Tang, S., Xiao, J., Zhang, D., Tan, C., Li, X., 2020. Rethinking the bottom-up framework for query-based video localization, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 10551–10558.

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification, in: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), IEEE. pp. 539–546.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Drukker, L., Sharma, H., Droste, R., Alsharid, M., Chatelain, P., Noble, J.A., Papageorghiou, A.T., 2021. Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video. Scientific Reports 11, 14109.

Fan, H., Xiong, B., et al., . Multiscale vision transformers, in: CVPR 2021.

Fu, Z., Jiao, J., Yasrab, R., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2023. Anatomy-aware contrastive representation learning for fetal ultrasound, in: Karlinsky, L., Michaeli, T., Nishino, K. (Eds.), Computer Vision – ECCV 2022 Workshops, Springer Nature Switzerland, Cham. pp. 422–436.

Gao, J., Sun, C., Yang, Z., Nevatia, R., 2017. Tall: Temporal activity localization via language query, in: Proceedings of the IEEE international conference on computer vision, pp. 5267–5275.

Ge, W., 2018. Deep metric learning with hierarchical triplet loss, in: Proceedings of the European conference on computer vision (ECCV), pp. 269–285.

Gordo, A., Almazán, J., Revaud, J., Larlus, D., 2016. 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. pp. 241–257.

Goyal, R., Mavroudi, E., Yang, X., Sukhbaatar, S., Sigal, L., Feiszli, M., Torresani, L., Tran, D., 2023. Minotaur: Multi-task video grounding from multimodal queries. arXiv preprint arXiv:2302.08063 .

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al., 2022. Ego4d: Around the world in 3,000 hours of egocentric video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18995–19012.

Hara, K., Kataoka, H., Satoh, Y., 2017. Learning spatio-temporal features with 3d residual networks for action recognition, in: Proceedings of the IEEE international conference on computer vision workshops, pp. 3154–3160.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B., 2018. Localizing moments in video with temporal language. arXiv preprint arXiv:1809.01337 .

Jiang, H., Ramakrishnan, S.K., Grauman, K., 2023. Single-stage visual query localization in egocentric videos. arXiv preprint arXiv:2306.09324 .

Lee, L.H., Gao, Y., Noble, J.A., 2021. Principled ultrasound data augmentation for classification of standard planes, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 729–741.

Lee, S., Lee, S., Seong, H., Kim, E., 2023. Revisiting self-similarity: Structural embedding for image retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23412–23421.

Lei, J., Berg, T.L., Bansal, M., 2021. Detecting moments and highlights in videos via natural language queries. Advances in Neural Information Processing Systems 34, 11846–11858.

Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C., 2022. Mvitv2: Improved multiscale vision transformers for classification and detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4804–4814.

Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z., 2023. Univtg: Towards unified video-language temporal grounding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2794–2804.

Liu, D., Qu, X., Dong, J., Zhou, P., Cheng, Y., Wei, W., Xu, Z., Xie, Y., 2021. Context-aware biaffine localizing network for temporal sentence grounding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11235–11244.

Liu, D., Qu, X., Wang, Y., Di, X., Zou, K., Cheng, Y., Xu, Z., Zhou, P., 2022a. Unsupervised temporal video grounding with deep semantic clustering, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1683–1691.

Liu, M., Wang, X., Nie, L., Tian, Q., Chen, B., Chua, T.S., 2018. Cross-modal moment localization in videos, in: Proceedings of the 26th ACM international conference on Multimedia, pp. 843–851.

Liu, Y., Li, S., Wu, Y., Chen, C.W., Shan, Y., Qie, X., 2022b. 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, pp. 3042–3051.

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochemia medica 22, 276–282.

Men, Q., Zhao, H., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2023. Towards standard plane prediction of fetal head ultrasound with domain adaption, in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–5.

Mishra, D., Saha, P., Zhao, H., Patey, O., Papageorghiou, A.T., Noble, J.A., 2024. 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, Springer Nature Switzerland.

Moon, W., Hyun, S., Park, S., Park, D., Heo, J.P., 2023. Query-dependent video representation for moment retrieval and highlight detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23023–23033.

Mun, J., Cho, M., Han, B., 2020. Local-global video-text interactions for temporal grounding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10810–10819.

Pan, T., Xu, F., Yang, X., He, S., Jiang, C., Guo, Q., Qian, F., Zhang, X., Cheng, Y., Yang, L., et al., 2023. Boundary-aware backward-compatible representation via adversarial learning in image retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15201–15210.

Rahmatullah, B., Papageorghiou, A., Noble, J.A., 2011. 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. pp. 35–42.

Sain, A., Bhunia, A.K., Chowdhury, P.N., Koley, S., Xiang, T., Song, Y.Z., 2023. 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, pp. 2765–2775.

Salomon, L., Alfirevic, Z., Berghella, V., Bilardo, C., Chalouhi, G., Costa, F.D.S., Hernandez-Andrade, E., Malinger, G., Munoz, H., Paladini, D., et al., 2022. Isuog practice guidelines (updated): performance of the routine mid-trimester fetal ultrasound scan. Ultrasound in obstetrics & gynecology: the official journal of the International Society of Ultrasound in Obstetrics and Gynecology 59, 840–856.

Schlemper, J., Oktay, O., Chen, L., Matthew, J., Knight, C., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention-gated networks for improving ultrasound scan plane detection. arXiv preprint arXiv:1804.05338 .

Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823.

Sharma, H., Drukker, L., Chatelain, P., Droste, R., Papageorghiou, A.T., Noble, J.A., 2021. Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos. Medical Image Analysis 69, 101973.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Wang, L., Mittal, G., Sajeev, S., Yu, Y., Hall, M., Boddeti, V.N., Chen, M., 2023. Protege: Untrimmed pretraining for video temporal grounding by video temporal grounding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6575–6585.

Wang, W., et al., . Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: CVPR 2021.

Xiao, S., Chen, L., Zhang, S., Ji, W., Shao, J., Ye, L., Xiao, J., 2021. Boundary proposal network for two-stage natural language video localization, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2986–2994.

Xu, H., He, K., Plummer, B.A., Sigal, L., Sclaroff, S., Saenko, K., 2019. Multilevel language and vision integration for text-to-clip retrieval, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 9062–9069.

Xu, M., Fu, C.Y., Li, Y., Ghanem, B., Perez-Rua, J.M., Xiang, T., 2022. Negative frames matter in egocentric visual query 2d localization. arXiv preprint arXiv:2208.01949 .

Xu, M., Li, Y., Fu, C.Y., Ghanem, B., Xiang, T., Pérez-Rúa, J.M., 2023. 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, pp. 2593–2603.

Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C., 2022. Tubedetr: Spatio-temporal video grounding with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16442–16453.

Yaqub, M., Kelly, B., Papageorghiou, A.T., Noble, J.A., 2015. Guided random forests for identification of key fetal anatomy and image categorization in ultrasound scans, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham. pp. 687–694.

Yasrab, R., Fu, Z., Zhao, H., Lee, L.H., Sharma, H., Drukker, L., Papageorgiou, A.T., Noble, J.A., 2022. A machine learning method for automated description and workflow analysis of first trimester ultrasound scans. IEEE Transactions on Medical Imaging 42, 1301–1313.

Yuan, Y., Mei, T., Zhu, W., 2019. To find where you talk: Temporal sentence localization in video with attention based location regression, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 9159–9166.

Zeng, R., Xu, H., Huang, W., Chen, P., Tan, M., Gan, C., 2020. Dense regression network for video grounding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10287–10296.

Zhang, D., Dai, X., Wang, X., Wang, Y.F., Davis, L.S., 2019. 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, pp. 1247–1257.

Zhang, H., Sun, A., Jing, W., Zhou, J.T., 2020. Span-based localizing network for natural language video localization. arXiv preprint arXiv:2004.13931 .

Zhao, H., Zheng, Q., Teng, C., Yasrab, R., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2022. Towards unsupervised ultrasound video clinical quality assessment with multi-modality data, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 228–237.

Zhao, H., Zheng, Q., Teng, C., Yasrab, R., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2023. Memory-based unsupervised video clinical quality assessment with multi-modality data in fetal ultrasound. Medical Image Analysis 90, 102977.