Searching audiovisual media with natural language queries

Andreea-Maria Oncescu

St John's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Supervised by

Dr. Samuel Albanie, Dr. João F. Henriques,

Prof. Andrew Zisserman

Hilary 2024

Abstract

In the modern era, individuals are flooded with information. Developments in storage technology have enabled vast information from sources like entertainment media and historical archives to be preserved cheaply at scale. Consequently, there is a pressing need for approaches that efficiently search through this content.

In this thesis, we leverage developments in deep learning and various data sources to advance capabilities of retrieval systems in understanding the interactions between text, audio and video content. We do this through three core contributions.

First, we focus on text-video retrieval with natural queries, collecting and benchmarking a high-quality dataset with detailed, well-localised descriptions and corresponding long videos. We showcase the advantages of using multiple experts e.g. object and audio classification models to improve text-video retrieval performance.

Second, we propose new benchmarks for semantic text-audio retrieval using free-form text. We employ state-of-the-art multimodal text-video retrieval models for this task and investigate how useful visual support is for finding the correct audio file. Additionally, we propose a large free-form text-audio dataset to aid with training of large text-audio models. Lastly, we investigate if text-audio retrieval models understand temporal ordering of sound events. We then propose a new contrastive loss term to guide the model to focus on temporal cues.

Finally, we employ Large Language Models' (LLMs) understanding of the world to leverage large text-video datasets for text-audio understanding. We show that LLMs are capable of proposing plausible descriptions for video soundtracks, starting from the visual-based descriptions of the video content. This is important, as it can be used to scale up current text-audio retrieval datasets.

Keywords – multimodal retrieval, large language models, deep learning

This thesis is submitted to the Department of Engineering Science, The University of Oxford, in fulfilment of the requirements for the degree of Doctor of Philosophy. This thesis is entirely my own work, and except where otherwise stated, describes my own research.

Andreea-Maria Oncescu, Apr 2024.

Acknowledgement

I would like to thank my supervisors Samuel Albanie, João F. Henriques and Andrew Zisserman for their support during these 4 years. I am grateful to them for giving me the chance to join the Visual Geometry Group and for providing me with the opportunity to work on such interesting projects. Special thanks to Sam for providing an intensive course to get me up to speed with everything as I was starting my DPhil. Also, I am very grateful to my unofficial supervisor, A. Sophia Koepke, for her guidance and support from the early days when I was learning how to use tmux, until the last days when I was writing my thesis. Lastly, special thanks to Alice, my previous supervisor, for teaching me good research standards and helping me figure out what I wanted to do.

I would like to thank my family for their continuous support and encouragement throughout my (unending) studies. They always tried to help me in any way they could, and I am very grateful for that. Thanks to my mom for always listening to me complaining about being stuck on some problem and to Costin for helping with writing, coding and for always being there for me.

Thank you to my friends Alex, Julia and Emil for when we went to Prague and you all used up your Charles Bridge statue wishes on me finishing my DPhil. It worked!!

To my waking early support groups, the breakfast club: Felipe, Anna and Dave and the coffee meetings with Ola, Georiga, Ester and Nina (in no particular order). Thank you for getting me out of bed and into the office to work on my thesis.

I am very grateful to Daffy for giving me the opportunity to work with him at Meta in New York for a summer. Thank you and Huiyu and Bernie for pushing me to get out of my research confort zone and for nudging me to network. And of course, my time in New York would not have been the same without my fellow interns and friends, Desi, Daniel, Sagar, Leo, Efi, Karan, Pascal, Keren and the rest of the Meta people. Lastly, thank you to Utsav for being a great host and showing me around the city.

I would also like to thank my colleagues at the Visual Geometry Group and at University of Oxford for their support. Sharing common struggles and deadlines made the journey easier. To name a few, thanks to Subha, Jelka, Owen, Henrique, Tom, Marian, David, Paul and Deevanshi.

And how can I forget about the main support group, Eliza and Andreea. Thank you for always agreeing with me and relating to my problems. I dare to say this group was instrumental in keeping me going.

Last but definitely not least, I would like to thank Emil for his support and encouragement throughout and for making me laugh even when I was already crying. I am very grateful for all the time you spent reading my drafts, code and listening to me complain about my work. I am very lucky to have you in my life.

Finally, I would like to thank my examiners Prof. Esa Rahtu and Dr. Iro Laina for the fruitful conversation and the feedback provided during the viva and the EPSRC for funding my DPhil.

Contents

1 Introduction and Background **9**

1.1 Contributions outline 13

1.2 Publications 16

2 QuerYD: A Video Dataset with High-Quality Text and Audio Narrations **18**

2.1 Introduction 20

2.2 Related work 21

2.3 The QuerYD dataset 23

2.4 Video understanding tasks 26

2.5 Conclusion 28

3 Audio Retrieval with Natural Language Queries: A Benchmark Study **30**

3.1 Introduction 32

3.2 Related work 35

3.3 SoundDescs dataset 37

3.3.1 Dataset collection 38

3.3.2 Data analysis 38

3.3.3 Dataset examples 41

3.3.4 Rights to use . . . 42
3.4 Methods, datasets, and benchmark . . . 42
3.4.1 Methods . . . 42
3.4.2 Datasets . . . 44
3.4.3 Benchmark . . . 46
3.5 Experiments . . . 46
3.6 Conclusion . . . 53
3.7 Supplementary material . . . 53

4 Dissecting Temporal Understanding in Text-to-Audio Retrieval 56
4.1 Introduction . . . 57
4.2 Related work . . . 60
4.3 Temporal understanding in text-to-audio retrieval . . . 62
4.3.1 AudioCaps dataset . . . 62
4.3.2 Model performance on AudioCaps . . . 64
4.4 Temporal understanding in controlled setting . . . 71
4.4.1 Data generation . . . 71
4.4.2 SynCaps experiments . . . 72
4.4.3 Proposing new loss . . . 72
4.5 Conclusion . . . 73
4.6 Supplementary Material . . . 74
4.6.1 AudioCaps validation data . . . 74
4.6.2 Additional experiments on AudioCaps . . . 74
4.6.3 Why AudioCaps uniform data . . . 75
4.6.4 Employing the text-text contrastive loss in other datasets . . . 76
4.6.5 Verify correctness of LLM evaluations of AudioCaps descriptions . . . 76

5 A sound approach: Using large language models to generate audio

descriptions for egocentric text-audio retrieval **77**
5.1 Introduction . . . 79
5.2 Related work . . . 80
5.3 Datasets and approach . . . 81
5.3.1 Datasets and tasks . . . 81
5.3.2 Audio description generation methodology . . . 82
5.3.3 Our proposed text-audio retrieval benchmarks . . . 83
5.3.4 Determining the audio relevancy using LLMs . . . 84
5.4 Experiments . . . 84
5.5 Conclusion . . . 87
5.6 Supplementary Material . . . 89
5.6.1 EpicSounds . . . 89
5.6.2 Prompts . . . 90
5.6.3 Multimodal text-video retrieval . . . 90
5.6.4 Leveraging models finetuned on Clotho for tasks with original visual descriptions . . . 93
5.6.5 Intra-video vs inter-video results . . . 94
5.6.6 Evaluation on different subsets of AudioEgoMCQ according to informativeness . . . 94
5.6.7 Why does WavCaps-Cl give better results on AudioEpicMIR and AudioEgoMCQ whilst the WavCaps-AC model is better on EpicSoundsRet? . . . 95

6 Summary and extensions **96**

References **102**

A Audio Retrieval with Natural Language Queries **118**
A.1 Introduction . . . 119

A.2 Related work 121
A.3 Methods, datasets and benchmarks 123
A.4 Experiments 127
A.5 Conclusion 131
A.6 Supplementary material 132
A.6.1 Pre-trained experts 132
A.6.2 Optimisation details and hyperparameter choices 133

B Statement of Authorship **135**

Chapter 1

Introduction and Background

From ancient times, the establishment of rules and laws was crucial for allowing larger groups of individuals to coexist in harmony. However, as the number of individuals grew, it became harder to effectively manage these more complex societies. The capacity to store detailed information about individuals and their contributions to society started exceeding the limits of human memory [Harari 2014]. As a result, taking written notes became more and more important, allowing leaders to keep track of people's information, such as birth dates, taxes paid to the state or wages owed by the state. It also allowed for coordinating people's work more efficiently, improving the productivity of the society [Diamond 1999]. Moreover, writing played a pivotal role in the preservation and dissemination of knowledge, culture, and religion, laying the foundation for advanced civilizations and fostering cultural and technological advancements across generations.

Having all these records written down led to the creation of libraries and archives [Molina 2016], which, in turn, required efficient methods for searching through these records. Initially, searches were conducted manually by checking each entry in the archive. As the size of the archives and libraries grew, searching this way became unsustainable. This resulted in the creation of catalogues, which kept short information about records such as titles and descriptions, improving search efficiency [Casson 2001]. Fast forward to the internet era, the amount of information has grown exponentially, not only in text but also in images, sounds, and videos. Consequently, the need for efficient search mechanisms has become more critical than ever.

In current times, platforms such as YouTube, Instagram, TikTok, and WeChat are inundated with staggering quantities of content, while news organizations like the BBC, Reuters, and Associated Press continue to amass vast archives of historically significant material. This data represents a wealth of information for artists to discover new ideas, for social media users to share, discuss (and remember) personal experiences, and for historians and journalists to find evidence for stories and events. However, navigating through all these extensive collections of multimedia (in the form of images, sounds and videos) is a significant challenge. The central goal of this thesis is to help the next generation of computational algorithms to search through audiovisual media extremely efficiently.

The technical foundation of this thesis is an approach known as 'content-based search'. That is, rather than reading the video title, the algorithm 'watches' the video for itself and tries to determine whether the video is relevant to the user's query. For example, given a query 'a video showing a tiger on a refrigerator', the algorithm would scan the video and try to recognise the appearance of a tiger itself (and the refrigerator), rather than relying on the video uploader to include this information in the title or metadata tags.

Building on the principles of content-based search, this research also extends into the domain of text-audio search. Before proceeding, it is important to distinguish here between general audio sounds and speech. This thesis focuses on the former, exploring the task of semantic search for general audio data. At the time of starting this research, the way to semantically search through audio files was only possible using tags [Slaney 2002b; Chechik et al. 2008] such as "train", "bells", "dog barks" instead of more detailed text queries such as "dog barking in a moving train". The ability to efficiently search through audio data can be very useful for creative applications by allowing artists and content producers to find highly specific and relevant audio files. An example is the fast-developing domain of creating podcasts and the need to find the right sounds to accompany the vocal content.

The task of finding the video or audio that best matches a given text query presents multiple challenges. One challenge is not having sufficient high-quality datasets. While some datasets exist, they are either poorly curated (e.g they could contain noise in the form of misaligned text descriptions for a given video [Miech et al. 2019b; Han et al. 2022]), or they are difficult to use or distribute due to copyright

issues [Rohrbach et al. 2015]. Another challenge is having the models to actually search through the data efficiently and effectively. Lastly, having benchmarks to compare the performance of different models is crucial for guiding future model development.

This thesis is split into three themes. The first theme explores the collection of high-quality narration text-video datasets. This is crucial for advancing tasks such as: text-video retrieval where given a description we want to find the video content that best matches the text query [C. Zhu et al. 2023], video moment localization where we want to find the time interval in the video where a given event takes place [M. Liu et al. 2023], and video captioning where given a video we want to generate a description for the video content [Jain et al. 2022; Iashin and Rahtu 2020]. Text-video datasets come in different shapes and sizes, depending on the downstream task the collectors have in mind. These datasets can be curated for many different tasks such as action/motion classification, where short class labels (e.g. drinking, sword fighting) are provided e.g. [Kuehne et al. 2011; Laptev et al. 2008]. For tasks such as video captioning or text-video retrieval, having more descriptive captions helps the models better differentiate between videos. These datasets can have brief descriptions of the main event (e.g. "A woman is adding flour to meat." [D. L. Chen and Dolan 2011]) or multiple (denser) event descriptions that overlap (e.g. "A woman walks to the piano and briefly talks to the the elderly man. The woman starts singing along with the pianist. Another man starts dancing to the music, gathering attention from the crowd." [Krishna et al. 2017]). However, obtaining dense descriptions for video requires a lot of human effort, making these datasets more limited in size. To address this challenge, the pivotal work by [Miech et al. 2019b] shows that increasing the amount of text-video pairs available for pre-training greatly improves downstream retrieval performance, even when many text-video pairs are misaligned (noisy). This naturally raises the question of how important aligned pairs of text and video are for downstream performance. [Han et al. 2022] address this question by showing that better aligning the same text-video pairs during training, results in increased performance on tasks such as text-video retrieval and classification. These findings underscore the demand for not only more data but also for datasets of higher quality.

The second theme focuses on the task of text-audio retrieval. The task of searching

through audio databases has been studied as early as 1995, when [Ghias et al. 1995] proposed an approach to search through a database using a query in the form of humming a melody. A year later, [Wold et al. 1996] paved the way for searching through an audio database by using words that describe characteristics of the audio (e.g. 'animal') and/or audio file/s as queries. The features in both these works are manually extracted, to describe features percieved by humans such as pitch and loudness. In more recent years, the field has matured, with features being extracted using Convolutional Neural Networks (CNNs) (e.g. [Hershey et al. 2016]) and, recently, audio-based Transformers (e.g. [Gong et al. 2021]). However, most of the benchmarks available are still limited to using class labels or keywords as queries. These works perform retrieval on newly curated datasets [Chechik et al. 2008] or employ classification datasets for this task [Elizalde et al. 2019]. This constrains how explicit we can be when searching for audio content. Thus, having benchmarks that take in free-form text descriptions for this task would be beneficial for advancing the field of text-audio retrieval. Lastly, we want our models to not only understand the content of the audio files, but also the temporal ordering of sound events. This is crucial for retrieving the correct audio file when the query contains temporal cues. [H.-H. Wu et al. 2023] show in their paper that text-audio models do not make use of temporal cues yet. Starting from this, we investigate if more recent models have this understanding and if existing benchmarks are suitable for evaluating this task.

In the third and last theme of this thesis, the focus is on leveraging Large Language Models (LLMs) and large video-text databases to improve the performance of text-audio retrieval. Large Language Models have achieved impressive capabilities on solving very general language tasks such as long-range text understanding [Paperno et al. 2016], world knowledge [Kwiatkowski et al. 2019; Joshi et al. 2017], code generation [M. Chen et al. 2021] or multi-task understanding [Hendrycks et al. 2021b; Hendrycks et al. 2021a] (e.g. history, mathematics questions). The recent progress in this field has been made possible by the availability of large text datasets e.g. [Y. Zhu et al. 2015; Penedo et al. 2023] and also the powerful Transformer architecture [Vaswani et al. 2017], resulting in models such as BERT [Devlin et al. 2018], LLama2 [Touvron et al. 2023b], GPT-4 [OpenAI 2023]. Given LLMs' exposure to varied information and the vast amount of available

text-video datasets, as described in the first theme, we now investigate the potential of LLMs to act as external knowledge bases for bridging modalities such as video and audio, through text descriptions. More specifically, we show that LLMs can generate plausible video soundtrack (audio) descriptions starting from visual labels or descriptions of the video. This capability can then be used to generate large (albeit noisy) text-audio datasets.

1.1 Contributions outline

Theme 1: Text-video data collection (Chapter 2) As part of the text-video data collection theme, we propose collecting and curating a new dataset called QuerYD (Chapter 2). We believe this makes for a good benchmark for the task of text-video retrieval as its goal is to provide Audio Descriptions[1] of visual content for users with visual impairments. Therefore, the descriptions are highly related to the video content and are very dense as they need to convey enough information for someone who cannot see the video to understand the action. Furthermore, the dataset showcases a wide range of diversity, featuring examples from various domains including cooking, sports, and music, as well as talk shows, cartoons, and interviews. QuerYD has a non-uniform distribution of video durations, which sets it apart from the many fixed duration 10-second long video datasets e.g. Kinetics-700 [Kay et al. 2017]. Lastly, the dataset can be easily distributed as it is contains publicly available videos rather than copyrighted movies (e.g. [Rohrbach et al. 2015]).

Theme 2: Free-form semantic text search through audios (Chapter 3, Chapter 4) To leverage the advancements in models' understanding of text and audio, in Chapter 3 we propose benchmarks for the more difficult task of free-form text-audio retrieval. To do so, we employ a model that can take in multiple modalities (e.g. video frames, audio content, text) and adapt it to this new task. We propose using the only two audio captioning based datasets AudioCaps [C. D. Kim et al. 2019] and Clotho [Drossos et al. 2020] available at the time for this

[1] In this chapter, Audio Description refers to an audio-based narration designed to provide accessibility of video content for users with visual impairments. However, it is important to note that in the subsequent chapters, the term 'audio description' (in lowercase) will be used within a different context – specifically, it refers to a written content description of audio files.

task.

However, the data for training and evaluating the text-audio retrieval models is limited, especially if we compare it with the plethora of text-video data available for research currently. This motivated us to collect a new dataset called SoundDescs. These contributions are detailed in Chapter 3.

Additionally, to encourage research in the field of text-audio retrieval using free-form queries, we have created an online demo[2] where people can search for the audio content they are interested in, using their own words, rather than pre-defined labels. The demo also allows researchers to visualise success and failure cases of a text-audio retrieval model, inspiring future research in improving text-audio search results.

In Chapter 4 we take a closer look at a current state-of-the-art text-audio retrieval model. Here, we want to evaluate to which extent this model understands temporal ordering of sound events. To do so, we first analyse how well-suited the current text-audio retrieval datasets are for this task in terms of both training and evaluation. We then propose a synthetic dataset, aiming to decouple the model's capacity to understand temporal content from the data being noisy or not diverse enough (e.g. only having sentences containing 'followed by' but no sentences containing 'before'). We conclude that both the data is not well suited for this task and the model does not make use of temporal cues even in a more controlled synthetic setting. Lastly, we investigate using an additional loss function during training to guide the model to pay attention to temporal cues.

Theme 3: Leveraging LLMs and text-video datasets for text-audio retrieval (Chapter 5) Text-audio retrieval datasets available today [Y. Wu et al. 2023; Mei et al. 2023] are still 20 times smaller than, for example, the HowTo100m text-video dataset [Miech et al. 2019b]. At the same time, state-of-the-art models for text-audio retrieval and text-video retrieval both leverage similarly sized Transformer-based architectures [Vaswani et al. 2017; Dosovitskiy et al. 2021; Gong et al. 2021]. Therefore, based on studies such as [Zhai et al. 2022; Koepke et al. 2022] in the video and audio domain, having more text-audio data to train on has the potential to further enhance text-audio retrieval models' performance.

[2]https://meru.robots.ox.ac.uk/audio-retrieval/

Consequently, in Chapter 5 we propose to leverage large text-video datasets and Large Language Models to generate noisy-but-large text-audio datasets in the egocentric setting. We show that LLMs can indeed help with the text-audio retrieval task by generating audio descriptions starting from video-based labels.

In Chapter 5 we additionally show that LLMs can be used to evaluate the difficulty of identifying a visual action given the corresponding audio soundtrack. This is useful, for example, in curating text-audio datasets from text-video datasets where we use the visual text label as audio description and the video soundtrack as audio. Then, using this proposed method, we can remove text-audio datapoints that are considered to be noisy. Finally, we investigate how useful the audio component is for the task of text-video retrieval in the egocentric setting.

1.2 Publications

The format of this thesis is an integrated one. The chapters of the thesis are therefore based on peer-reviewed publications together with a chapter that contains a manuscript that is currently undergoing review.

A-M. Oncescu, J.F. Henriques, Y. Liu, A. Zisserman and S. Albanie: QuerYD: A Video Dataset with High-Quality Text and Audio Narrations, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021.

A.S. Koepke∗, A-M. Oncescu∗, J.F. Henriques, Z. Akata and S. Albanie: Audio Retrieval with Natural Language Queries: A Benchmark Study, in *IEEE Transactions on Multimedia*, 2022.

A-M. Oncescu, J.F. Henriques and A.S. Koepke: Dissecting Temporal Understanding in Text-to-Audio Retrieval, under submission at *Association for Computing Machinery's International Conference on Multimedia*, 2024.

A-M. Oncescu, J.F. Henriques, A. Zisserman, S. Albanie and A.S. Koepke: A Sound Approach: Using Large Language Models to Generate Audio Descriptions for Egocentric text-audio retrieval, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024.

The following paper is included in the appendix:

A-M. Oncescu∗, A.S. Koepke∗, J.F. Henriques, Z. Akata and S. Albanie: Audio Retrieval with Natural Language Queries, in *INTERSPEECH*, 2021.

The appendix includes the original paper that laid the groundwork for the more complete study detailed in the Journal article, "Audio Retrieval with Natural Language Queries: A Benchmark Study" (Chapter 3). While the journal article contains the core findings and methodologies of the original submission, the appendix provides a more complete understanding of the research process. It contains more information about the original experimental setup which helps showcase how to better tune the hyperparameters to improve benchmark results.

Lastly, a demo[3] was presented at the IEEE Workshop on Applications of Signal

[3]https://meru.robots.ox.ac.uk/audio-retrieval/

Processing to Audio and Acoustics (WASPAA 2021) as a follow-up to the INTERSPEECH paper (included in the appendix).

Chapter 2

QuerYD: A Video Dataset with High-Quality Text and Audio Narrations

This paper has been accepted for publication at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2021.

QuerYD: a video dataset with high-quality text and audio narrations

Andreea-Maria Oncescu João F. Henriques
Yang Liu Andrew Zisserman Samuel Albanie
Visual Geometry Group
University of Oxford
Oxford, UK

Abstract

We introduce QuerYD, a new large-scale dataset for retrieval and event localisation in video. A unique feature of our dataset is the availability of two audio tracks for each video: the original audio, and a high-quality spoken description of the visual content. The dataset is based on YouDescribe [Research and Center 2013], a volunteer project that assists visually-impaired people by attaching voiced narrations to existing YouTube videos. This ever-growing collection of videos contains highly detailed, temporally aligned audio and text annotations. The content descriptions are more relevant than dialogue, and more detailed than previous description attempts, which can be observed to contain many superficial or uninformative descriptions. To demonstrate the utility of the QuerYD dataset, we show that it can be used to train and benchmark strong models for retrieval and event localisation. Data, code and models are made publicly available, and we hope that QuerYD inspires further research on video understanding with written and spoken natural language.

Index Terms— Audio description, retrieval

2.1 Introduction

The development of new datasets has been instrumental in the immense progress of modern machine learning, working hand-in-hand with methodological improvements. New, larger datasets allow incremental gains in performance with no change in methodology [Hammar et al. 2018], and prevent progress stalling from overfitting to saturated benchmarks [Recht et al. 2018]. In this work, we propose a dataset that supports investigations into the relationship between video and natural language.

For this task, obtaining rich and diverse training data is essential, an endeavour that is made difficult by the fact that "describing a video" is a relatively underconstrained objective. Dataset designers that rely on crowdsourced annotations, usually through paid micro-work platforms such as Amazon Mechanical Turk, must specify elaborate guidelines and multi-stage verification mechanisms to attempt to remove low-effort solutions, given the monetary incentive [X. Wang et al. 2019].

We propose a new dataset for retrieval and localisation, sourced from the *YouDescribe* community[1] that contributes audio descriptions for YouTube videos. The videos cover diverse domains and the descriptions are precisely localised in time. They exhibit richer vocabulary than existing publicly available text-video description datasets (sec. 2.4).

The aim of the annotators is to *communicate* the contents of videos to visually-impaired people, resulting in diverse and high quality audio descriptions. Annotators add spoken narrations to videos from YouTube, subject to the time-constraint of describing the action in near real-time. The proposed dataset expands previous datasets across four axes: (1) *Modality.* In addition to the original audio track of each video, QuerYD contains a separate audio track containing spoken narrations describing the action. This modality is highly complementary to the standard audio, the content of which can be unrelated to the visual content (e.g. background music, character dialogue about off-screen events, and aspects of the action that do not produce sound). In contrast, spoken narrations have a one-to-one relation to the video's content, encoded as audio. (2) *Quantity.* QuerYD is large-scale, containing over 200 hours of video and 74 hours of audio descriptions.

[1] http://youdescribe.ski.org

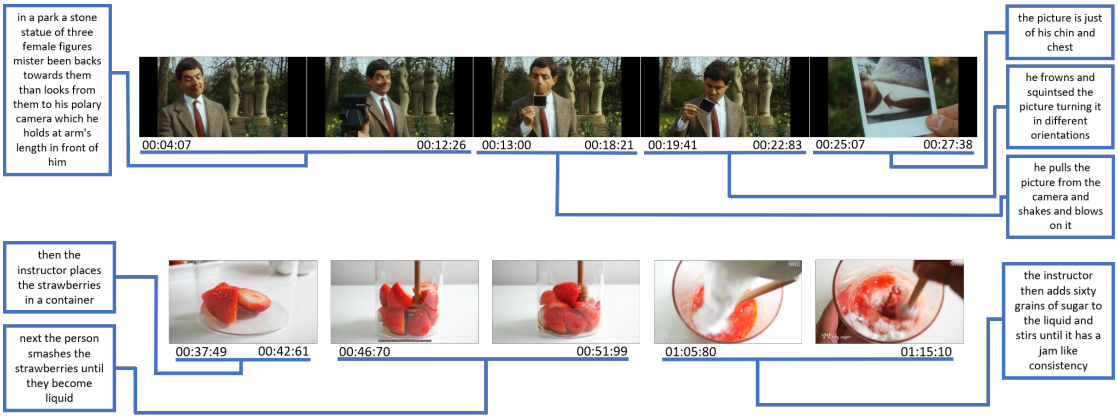


Figure 2.1: Qualitative examples from the QuerYD dataset. Audio descriptions for video segments were provided and transcribed by voluntary annotators. Some frames are individually described while other annotations take place over a sequence of frames. The time localisation of the audio descriptions is presented in the format *minutes:seconds:miliseconds*.

It also contains a higher density of descriptions than other datasets (as measured by words-per-second). (3) *Quality.* Since the descriptions in QuerYD are created by volunteers who aim for high-quality descriptions of videos for the visually-impaired, and narrations are rated by other users, there is an added incentive for quality when compared to micro-work platforms. We demonstrate this difference empirically, observing a larger vocabulary size, larger number of sentences per video, and sentence lengths, when compared to other datasets (which generally contain text captions instead of audio narrations). (4) *Scalability.* QuerYD is not a static dataset, since it is based on an ever-growing collection of audio descriptions, continually evolving with new narrations added every day. By periodically updating this dataset, we will empower future researchers to obtain more up-to-date snapshots of the audio descriptions data, ensuring that it stays relevant as the data demand grows.

In summary, our contributions are: (1) we propose QuerYD, a new dataset for video retrieval and localisation; (2) we provide an analysis of the data, demonstrating that it possesses a richer diversity of text descriptions than prior work; (3) we provide baseline performances with existing state-of-the-art models for text-video understanding tasks.

2.2 Related work

Researchers in the computer vision and natural language processing communities have been striving to bridge the gap between videos and natural language. We have

seen significant progress in many tasks, such as text-to-video retrieval, text-based action or event localisation, and video captioning. This success has resulted both from advances in deep learning as well as from the availability of video description datasets.

Video-Text Retrieval Benchmarks. In the past, video-text retrieval datasets have addressed controlled settings [Barbu et al. 2012; Kojima et al. 2002], specific domains such as cooking [Rohrbach et al. 2014; Das et al. 2013], and acted and edited material in movies [Rohrbach et al. 2015; Yao et al. 2015; Bain et al. 2020]. Some new video datasets [J. Xu et al. 2016; D. L. Chen and Dolan 2011; Anne Hendricks et al. 2017; Krishna et al. 2017; X. Wang et al. 2019] have been collected from YouTube and Flicker, which are open-domain and realistic. However, in each case, obtaining text annotations for these datasets is time-consuming and expensive and therefore difficult to do at large scale. Some recent works obtained text annotations by automatically transcribing narrations instead, and collected relatively large-scale datasets, such as CrossTask [Zhukov et al. 2019], YouCook2 [L. Zhou et al. 2018], HowTo100M [Miech et al. 2019b]. However, these datasets are all sourced from instructional videos related to pre-defined tasks. Moreover, collecting annotations from narration introduces some incoherence (noise) in the video-text pairs, due to the narrator talking about things unrelated to the video, or describing something before or after it happens.

In this work, we explore a different source for text annotations – Audio Description (AD). AD provides descriptions of the visual content in a video which allows visually-impaired people to understand the video. Compared to manual annotations and narration transcriptions from instructional videos, ADs describe precisely what is shown on the screen and are temporally aligned to the video. Related to our work, LSMDC [Rohrbach et al. 2015] also considers AD as a source of supervision. However, as opposed to our work, LSMDC focuses on movies and short clips (3s on average). Also of relevance is the Epic Kitchens dataset [Damen et al. 2018], which employs narration to provide descriptions, but differs from our approach through its focus on egocentric videos of kitchen-based activities. Our dataset instead aims to cover a more general range of domains.

Event Localisation Benchmarks. Event localisation is a task that aims to retrieve a specific temporal segment from a video given a natural language

text description. It requires the event localisation methods to determine not only what occurs in a video but also when. Some datasets have been collected for this task in the past few years, including Charades-STA [J. Gao et al. 2017], DiDemo [Anne Hendricks et al. 2017], and the Activity-Net [Krishna et al. 2017] dataset. However, the average length of the video in these datasets is less than 180s and the number of events per video is less than 4, which makes the event localisation less challenging. In contrast, our QuerYD dataset consists of longer videos and more distinct moments from each unedited video footage paired with descriptions that can uniquely localise the moment. It is also worth noting that instead of using manual annotations which can suffer from ambiguity in the event definition and disagreement of the start and end times, AD provide more fine-grained and precise temporal segment annotations.

2.3 The QuerYD dataset

In this section, we first give a high-level overview of the QuerYD dataset. Afterwards, we provide qualitative examples and quantitative analyses of the videos and descriptions that comprise the dataset, compared to others in the literature.

Dataset overview. The QuerYD dataset comprises 207 hours of video accompanied by 74 hours of Audio Description (AD) transcriptions, resulting in 31,441 descriptions. The videos, which are sourced from YouTube, cover a diverse range of visual content shown in Fig. 2.2. Of the total transcriptions, 13,019 are localised within video content (with precise start and end times), and the remaining 18,422 descriptions are *coarsely* localised (they are assigned a single time in a video, but their temporal extent is not annotated). The training, validation and test partitions of the dataset, obtained by uniform sampling of videos, are shown in Tab. 2.1.

Dataset analysis. We analysed several aspects of the QuerYD dataset: the diversity of vocabulary (Fig. 2.3), linguistic complexity (Tab. 2.2), as well as the

Partition	Train	Validation	Test
Untrimmed videos	1,815	388	390
Total descriptions	22,008	4,716	4,717
Localised descriptions	9,113	1,952	1,954

Table 2.1: The partitions of the proposed QuerYD dataset.

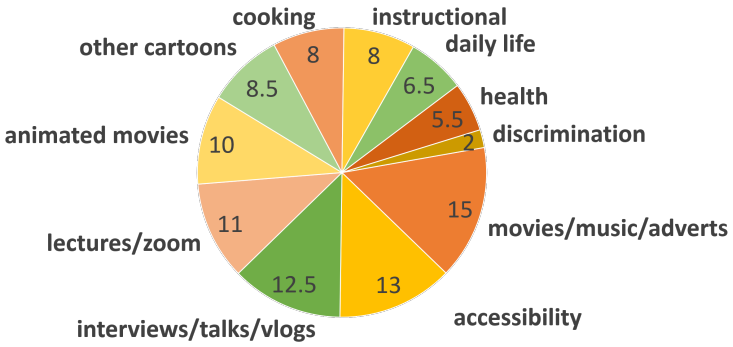


Figure 2.2: Detailed classification of the QuerYD video content by category, and associated proportion (%).

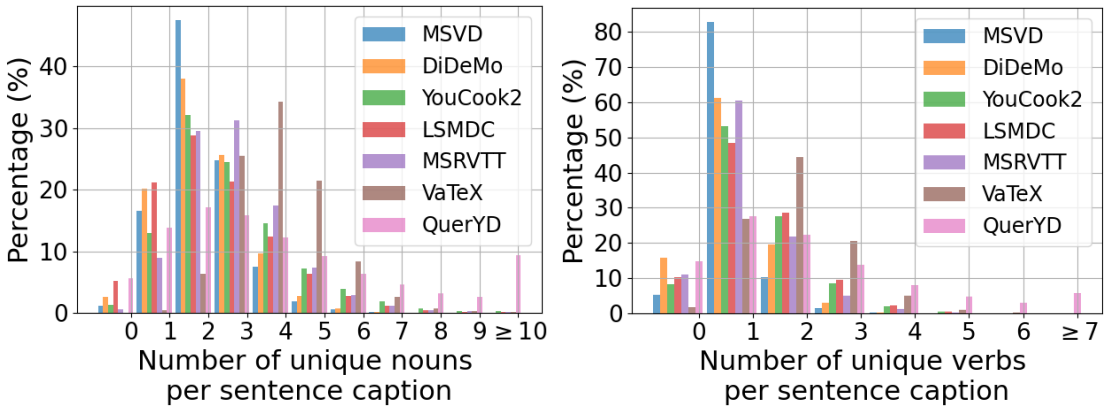


Figure 2.3: A comparison of the linguistic complexity and diversity of QuerYD with popular text-video datasets. Note that QuerYD has a higher proportion of unique nouns and verbs.

video duration and the distribution of localised segments (Tab. 2.2).

We compare the QuerYD dataset with other datasets containing annotated videos such as LSMDC [Rohrbach et al. 2015], VaTeX [X. Wang et al. 2019], YouCook2 [L. Zhou et al. 2018]. Because the QuerYD videos are chosen by voluntary annotators from YouTube, the topics vary greatly with a vocabulary more varied than that of more specific datasets, as can be seen in Fig. 2.3 and Table 2.2. The vocabulary size is calculated by first assigning words to their corresponding part-of-speech and then lemmatising the tokens using the Spacy library [AI 2016]. As well as having more varied video topics, the QuerYD dataset has a wider range of video lengths as demonstrated in Tab. 2.2, with 9 videos longer than 40 minutes. The average video time is 278 seconds for untrimmed video, while the average of the video segments is of 7.72 seconds. This results in better temporal localisation since the audio description provided describes only the current scene.

Qualitative examples. In this section, we explore some qualitative examples

Dataset	#Videos	Video Length	#Clips	Clip Length	#Sentence per video	Sent Length per clip	Vocabulary				
							Total	Noun	Verb	Adj	Adverb
DiDeMo	9605	25s	37185	6.3s	5.86 ± 0.34	7.50 ± 3.23	7865	3475	1316	841	339
ACT	20,000	180s	54926	36.18s	3.56 ± 1.67	13.58 ± 6.44	12413	5218	2162	1590	534
QuerYD	2593	278s	13019	7.72s	12.25 ± 15.27	19.91 ± 22.89	28515	8825	3551	3128	907

Table 2.2: Comparison of the event localisation datasets.

from QuerYD. Examples of videos and captions are given in Fig. 2.1 together with the timestamps of the corresponding audio descriptions. In the top picture there are four rich descriptions covering 27 seconds of one video. These descriptions have a high content of verbs and nouns, reflecting the statistics in Fig. 2.3. The bottom image is another example of detailed descriptions of a short video with extremely fine-grained localisation.

Descriptions localisation. To evaluate how well localised audio descriptions are, 100 audio descriptions were randomly selected and start and end times were carefully annotated. Seven descriptions were discarded: 2 due to irrelevant descriptions and 5 due to missing videos. Based on these values, the mean IOU metric for time intervals was found to be 0.62. When calculating the time difference in seconds between description's start and end times and the corresponding annotated values, a mean difference of seven seconds was obtained.

Data collection. QuerYD is gathered from user-contributed descriptions provided by the *YouDescribe* community, who contribute audio descriptions to videos hosted on YouTube to assist the visually-impaired [Research and Center 2013]. A portion of these audio descriptions is further accompanied by user-provided transcriptions. To handle cases in which such a transcription is not provided, we use the Google Speech-to-Text API [Google n.d.] to transcribe audio descriptions. Unlike speech transcription from speech "in the wild", audio descriptions are recorded in a clean environment and contain only the voice of the speaker, leading to high quality transcriptions. The *YouDescribe* privacy policy [Research and Center 2013] ensures that all captions are public with the consent of annotators. Lastly, we also group the localised descriptions with their corresponding clips for the localisation task and eliminate empty audio descriptions.

2.4 Video understanding tasks

In this section we demonstrate the application of QuerYD to two video understanding tasks: paragraph-level video retrieval and clip localisation. We consider three models:

(1) The *E2EWS* (End-to-end Weakly Supervised) model proposed by [Miech et al. 2019a] is a cross-modal retrieval model that is trained using weak supervision from a large corpus of 100 million instructional videos (using the speech content as supervision). The model employs a S3D [S. Xie et al. 2018] video feature extractor and a lightweight text encoder. We use this model without any form of fine-tuning [Bain et al. 2020] on QuerYD, providing a calibration of task difficulty.
(2) The *MoEE* (Mixture of Embedded Experts) model proposed by [Miech et al. 2018] is made of a multi-modal video model in combination with a system of context gates that learn to fuse together different pretrained "experts" [Bain et al. 2020] (inspired by the classical Mixture of Experts model [Jordan and Jacobs 1994]) to form a robust cross-modal text-video embedding.
(3) The *CE* (Collaborative Experts) model learns a cross-modal embedding by fusing a collection of pretrained experts to form a video encoder [Bain et al. 2020]. It uses a relation network [Santoro et al. 2017] sub-architecture to combine together different modalities, and represents the state-of-the-art on several retrieval benchmarks. Except where otherwise noted, the MoEE and CE models adopt the same four pretrained experts for scene classification, action recognition, ambient sound classification and image classification described in [Y. Liu et al. 2019] (described in detail in the supplemental material).

Text-video retrieval. We first consider the task of paragraph-level video-retrieval with natural language queries (e.g. as studied in [Bowen Zhang et al. 2018; Y. Liu et al. 2019]). We report standard retrieval evaluation metrics, including median rank (MdR, lower is better), mean rank (MnR, lower is better) and recall at rank K (R@k, higher is better). We report the mean and standard deviation for three runs with different random seeds. We first compare the three different baseline models on our dataset and then investigate the importance of the use of different modalities.

Baselines. Our results, reported in Tab. 2.3, show that the CE model [Y. Liu et

	Text ⟹ Video					Video ⟹ Text				
Method	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
E2EWS [Miech et al. 2019b]	$13.5_{\pm 0.0}$	$27.5_{\pm 0.0}$	$34.5_{\pm 0.0}$	$35.0_{\pm 0.0}$	$72.5_{\pm 0.0}$	$12.4_{\pm 0.0}$	$23.8_{\pm 0.0}$	$30.8_{\pm 0.0}$	$33.0_{\pm 0.0}$	$73.4_{\pm 0.0}$
MoEE [Miech et al. 2018]	$11.6_{\pm 1.3}$	$30.2_{\pm 3.0}$	$43.2_{\pm 3.1}$	$14.2_{\pm 1.6}$	$42.7_{\pm 2.6}$	$13.0_{\pm 3.1}$	$30.9_{\pm 2.0}$	$43.0_{\pm 2.8}$	$14.5_{\pm 1.8}$	$42.6_{\pm 1.5}$
CE [Y. Liu et al. 2019]	$13.9_{\pm 0.8}$	$37.6_{\pm 1.2}$	$48.3_{\pm 1.4}$	$11.3_{\pm 0.6}$	$35.1_{\pm 1.6}$	$13.7_{\pm 0.7}$	$35.2_{\pm 2.7}$	$46.9_{\pm 3.2}$	$12.3_{\pm 1.5}$	$35.8_{\pm 2.4}$

Table 2.3: Comparison of text-video retrieval methods trained with paragraph-level information on the QuerYD dataset.

	Text ⟹ Video					Video ⟹ Text				
Method	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
E2EWS [Miech et al. 2019b]	$3.4_{\pm 0.0}$	$9.8_{\pm 0.0}$	$14.4_{\pm 0.0}$	$274.0_{\pm 0.0}$	$523.8_{\pm 0.0}$	$4.2_{\pm 0.0}$	$10.4_{\pm 0.0}$	$13.9_{\pm 0.0}$	$268.5_{\pm 0.0}$	$522.0_{\pm 0.0}$

Table 2.4: Baseline for text-video retrieval methods on QuerYD. ↑ indicates that higher is better (similarly for ↓).

al. 2019] trained on our QuerYD dataset performs best, outperforming the MoEE [Miech et al. 2018] and E2EWS [Miech et al. 2019a] models (the latter being trained on a corpus of 100 million instructional videos, but used without any form of finetuning). Additionally, we propose a baseline for retrieval where the full QuerYD dataset is used as test set. The results when running the E2EWS [Miech et al. 2019a] model without finetuning on the full QuerYD dataset are shown in Table 2.4.

The importance of different modalities for QuerYD retrieval. In Tab. 2.5, we assess the importance of different pretrained experts for the retrieval performance of the CE model [Y. Liu et al. 2019] on QuerYD. We observe that the addition of each expert improves the performance, although to different extents: both object and action classification experts bring large gains in performance, while the ambient sound expert produces mixed results and has a smaller impact, indicating that this is a weaker cue, but one that still could benefit retrieval.

Clip localisation. We next study clip localisation. Corpus-level clip localisation without effective proposals is an extremely challenging task (e.g. [Escorcia et al. 2019] reports a performance of 0.85% recall at rank 1 on DiDeMo). For this

	Text ⟹ Video					Video ⟹ Text				
Method	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
SCENE	$8.7_{\pm 0.4}$	$26.3_{\pm 1.1}$	$37.1_{\pm 0.7}$	$22.2_{\pm 1.6}$	$52.3_{\pm 3.0}$	$9.1_{\pm 0.8}$	$25.4_{\pm 0.9}$	$35.3_{\pm 1.5}$	$23.2_{\pm 0.3}$	$52.6_{\pm 2.6}$
PREV. + AUDIO	$7.6_{\pm 2.7}$	$27.4_{\pm 1.4}$	$40.4_{\pm 0.9}$	$17.0_{\pm 1.7}$	$49.0_{\pm 1.9}$	$10.1_{\pm 1.2}$	$25.7_{\pm 1.5}$	$37.5_{\pm 1.2}$	$20.0_{\pm 1.3}$	$48.9_{\pm 2.0}$
PREV. + OBJECTS	$12.7_{\pm 1.7}$	$34.8_{\pm 1.7}$	$47.0_{\pm 1.3}$	$12.3_{\pm 0.6}$	$37.6_{\pm 2.1}$	$12.8_{\pm 1.3}$	$33.5_{\pm 2.8}$	$46.6_{\pm 1.0}$	$11.8_{\pm 0.8}$	$37.6_{\pm 1.9}$
PREV. + ACTION	$14.3_{\pm 0.3}$	$37.5_{\pm 1.3}$	$48.6_{\pm 0.8}$	$11.3_{\pm 0.6}$	$35.2_{\pm 1.8}$	$14.0_{\pm 0.3}$	$35.4_{\pm 2.9}$	$47.2_{\pm 2.8}$	$12.3_{\pm 1.5}$	$35.8_{\pm 2.4}$

Table 2.5: The influence of different pretrained experts for the performance of the CE model [Y. Liu et al. 2019] trained on QuerYD. The value and cumulative effect of different experts for scene classification (SCENE), ambient sound classification (AUDIO), image classification (OBJECT), and action recognition (ACTION). PREV. denotes the experts used in the previous row.

Method	Text ⟹ Video					Video ⟹ Text				
	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
E2EWS [Miech et al. 2019b]	$6.7_{\pm 0.0}$	$14.7_{\pm 0.0}$	$20.4_{\pm 0.0}$	$133.0_{\pm 0.0}$	$342.0_{\pm 0.0}$	$8.4_{\pm 0.0}$	$15.4_{\pm 0.0}$	$19.8_{\pm 0.0}$	$154.5_{\pm 0.0}$	$363.0_{\pm 0.0}$
MoEE [Miech et al. 2018]	$19.0_{\pm 0.8}$	$38.9_{\pm 1.0}$	$47.9_{\pm 0.7}$	$12.0_{\pm 1.0}$	$127.4_{\pm 5.9}$	$19.8_{\pm 0.2}$	$39.6_{\pm 0.6}$	$47.6_{\pm 0.1}$	$13.0_{\pm 0.0}$	$124.3_{\pm 5.5}$
CE [Y. Liu et al. 2019]	$18.2_{\pm 0.5}$	$38.1_{\pm 0.8}$	$46.8_{\pm 0.4}$	$13.3_{\pm 0.6}$	$127.5_{\pm 3.9}$	$18.1_{\pm 0.6}$	$37.3_{\pm 0.5}$	$45.9_{\pm 0.6}$	$14.0_{\pm 1.0}$	$123.9_{\pm 3.3}$

Table 2.6: Comparison of localisation methods trained with oracle temporal proposals information on the QuerYD dataset.

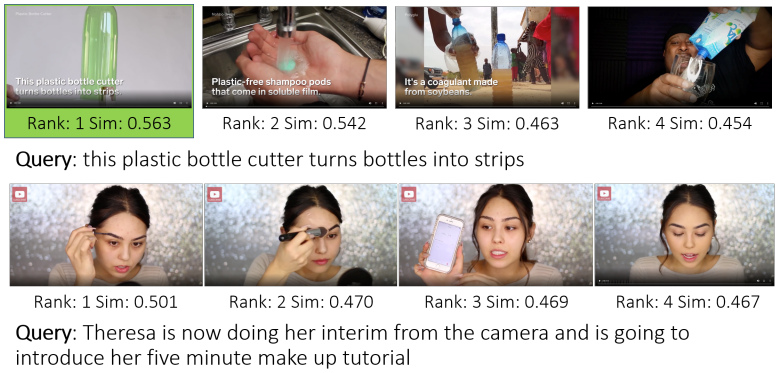


Figure 2.4: Success (top) and failure (bottom) cases of the CE model trained on QuerYD and tested on text-video retrieval.

reason, in this work we establish a baseline for the proposal oracle setting, which consists of having the temporal segment localisations (ground truth proposals) provided as input to each method.

Localisation baselines. Our results, reported in Tab. 2.6, show that the MoEE method performs best, followed closely by CE. The E2EWS method, which was trained with weak supervision from a much larger set of videos, performs worse. This may be due in part to the weak supervision, which does not enforce the network to be highly sensitive to timing. The model nevertheless establishes a solid baseline for performance without fine-tuning.

Failure cases. In Fig. 2.4 we show some failure cases of the CE model, which we observed to be the strongest model in the text-video retrieval task. One of the main failure modes is the ambiguity between segments that can plausibly correspond to the retrieval target, showing that at such a fine level of temporal granularity no model can perform perfectly.

2.5 Conclusion

In this work, we introduced the QuerYD dataset for video understanding. Through extensive analysis, we demonstrated that this dataset contains detailed annotations with rich, highly-relevant vocabulary. To demonstrate its applications and to boot-

strap the development of future methods, we evaluated several baselines for video retrieval and clip localisation with an oracle. By directing researchers' attention to this video-to-audio narration problem, and providing them with the tools to tackle it, we hope that *YouDescribe*'s goals of assisting the visually-impaired will be fully realised.

Acknowledgements. This work is supported by the EPSRC (VisualAI EP/T028572/1 and DTA Studentship), and the Royal Academy of Engineering (RF\201819\18\163). We are grateful to Sophia Koepke for her helpful comments and suggestions.

Chapter 3

Audio Retrieval with Natural Language Queries: A Benchmark Study

The paper has been accepted for publication at IEEE Transactions on Multimedia, 2022. This is an extension of the INTERSPEECH, 2021 accepted paper "Audio Retrieval with Natural Language Queries".

Audio Retrieval with Natural Language Queries: A Benchmark Study

A. Sophia Koepke*, Andreea-Maria Oncescu*, João F. Henriques, Zeynep Akata, and Samuel Albanie

Abstract

The objectives of this work are *cross-modal text-audio and audio-text retrieval*, in which the goal is to retrieve the audio content from a pool of candidates that best matches a given written description and vice versa. Text-audio retrieval enables users to search large databases through an intuitive interface: they simply issue free-form natural language descriptions of the sound they would like to hear. To study the tasks of text-audio and audio-text retrieval, which have received limited attention in the existing literature, we introduce three challenging new benchmarks. We first construct text-audio and audio-text retrieval benchmarks from the AudioCaps and Clotho audio captioning datasets. Additionally, we introduce the SoundDescs benchmark, which consists of paired audio and natural language descriptions for a diverse collection of sounds that are complementary to those found in AudioCaps and Clotho. We employ these three benchmarks to establish baselines for cross-modal text-audio and audio-text retrieval, where we demonstrate the benefits of pre-training on diverse audio tasks. We hope that our benchmarks will inspire further research into audio retrieval with free-form text queries. Code, audio features for all datasets used, and the SoundDescs dataset are publicly available at https://github.com/akoepke/audio-retrieval-benchmark.

Index Terms— Audio Retrieval, Text-based Retrieval, Datasets

3.1 Introduction

The vast and unabated growth of user-generated content in recent years has introduced a pressing need to search ever-growing databases of multimedia. Free-form natural language sentences (i.e. sequences of text that are written as they would be spoken) form an intuitive and powerful interface for composing search queries for these databases since they allow for expressing virtually any concept. Spanning multiple modalities, different retrieval strategies were developed for content as diverse as text (including web pages and books), images [Dong et al. 2016], and videos [Miech et al. 2018; Mithun et al. 2018]. Surprisingly, while search engines currently exist for these modalities (e.g. Google, Flickr and YouTube, respectively), unstructured audio is not accessible in the same way. The aim of this paper is to address this gap by curating the SoundDescs dataset which contains paired sound and natural language and by introducing benchmarks for text-audio retrieval.

It is important to distinguish content-based retrieval from that based on metadata, such as the title of a video or song or an audio tag. Metadata retrieval is feasible for manually-curated databases such as song or movie catalogues. However, content-based retrieval is more important in user-generated data, which often has little structure or metadata. There are methods to search for audio which matches an audio query [Manocha et al. 2018; Lallemand et al. 2012], but satisfying the requirement to input an example audio query can be difficult for a human (e.g. making convincing frog sounds is difficult). We, on the other hand, propose a framework which enables the searching of a sound database using detailed free-form natural language queries of the desired sound (e.g. "A man talking as music is playing followed by a frog croaking."). This enables the retrieval of audio data which will ideally match the temporal sequence of events in the query instead of

* Equal contribution.

A. S. Koepke (a-sophia.koepke@uni-tuebingen.de) and Z. Akata are with the Explainable Machine Learning group at the University of Tübingen. Z. Akata is also affiliated with the Max Planck Institute for Intelligent Systems, Tübingen and the Max Planck Institute for Informatics, Saarbrücken. A.-M. Oncescu (oncescu@robots.ox.ac.uk) and J. F. Henriques are with the Visual Geometry Group at the University of Oxford. S. Albanie is with the Department of Engineering at the University of Cambridge. © 2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

just a single class tag. Furthermore, natural language queries are a familiar user interface widely used in current search engines. Therefore, our proposed audio retrieval with free-form text queries could be a first step towards more natural and flexible audio-only search.

Text-based audio retrieval could also be beneficial for video retrieval. The majority of recent works that address the text-based video retrieval task focus heavily on the visual and text domains [Dong et al. 2016; Miech et al. 2018; Y. Liu et al. 2019; Gabeur et al. 2020]. Since audio and visual information inherently have natural semantic alignment for a significant portion of video data, text-based audio retrieval could also be used for querying video databases by only considering the audio stream of the video data. This would allow for video retrieval in the audio domain at reduced computational cost for cases in which audio and visual information correspond, with applications for low-power IoT devices, such as microphones in natural habitats, of particular interest for conservation and biology. Historical archives with extensive sound collections, such as the British Library Sounds, would be easier to search, facilitating historical research and public access. Furthermore, text-based retrieval could enable appealing creative applications, such as automatically finding (non-music) background sounds which correspond to input text. This could be especially useful given the growing popularity of audio podcasts and audiobooks which are often supplemented with (background) sounds that match their content.

Learning to retrieve audio, given natural language queries, requires data with paired text and sound. Audio captioning datasets naturally lend themselves to this task, since they contain audio and a matching text description for the sound. However, existing captioning datasets are limited in size and in the diversity of their audio content. Hence, we curated the novel SoundDescs dataset, which was sourced from the BBC Sound Effects database. SoundDescs contains text describing the sounds with significant variation with respect to the audio content and with a relatively large vocabulary used in the descriptions.

We introduce three new benchmarks for text-based audio retrieval, based on our proposed SoundDescs dataset, and the AudioCaps [C. D. Kim et al. 2019] and

https://sounds.bl.uk
https://sound-effects.bbcrewind.co.uk/

Clotho [Drossos et al. 2020] audio captioning datasets. AudioCaps consists of a subset of 10-second audio clips from the AudioSet dataset [Gemmeke et al. 2017] with additional human-written audio captions, while Clotho contains sounds sourced from the Freesound platform [Font et al. 2013], varying between 15 and 30 seconds in duration, and accompanied by crowd-sourced text captions. SoundDescs is considerably more varied in duration and audio content than the other two benchmarks. However, the text descriptions are of mixed quality, since they are obtained automatically from descriptions provided with the data.

In contrast to sound event class labels, audio captions contain detailed information about the sounds. A user searching for a particular sound would usually describe the sound using text similar to an audio caption. AudioCaps [C. D. Kim et al. 2019], Clotho [Drossos et al. 2020], and SoundDescs allow to leverage the matching text-audio pairs to train text-based audio retrieval frameworks. To establish baselines for this task, we adapt existing video retrieval frameworks for audio retrieval. We employ multiple pre-trained audio expert networks and show that using an ensemble of audio experts improves audio retrieval.

In summary, we make three contributions: (1) we introduce the SoundDescs dataset for text-based audio retrieval; (2) we introduce three new benchmarks for audio retrieval with natural language queries—to the best of our knowledge, these represent the first public benchmarks for this task; (3) we provide baseline performances with existing multi-modal video retrieval models that we adapt to text-based audio retrieval and show the benefits of using multiple datasets for pre-training.

This paper extends an initial Interspeech 2021 conference version of our work [Oncescu et al. 2021b] in two ways: (i) we introduce the new SoundDescs dataset for text-audio retrieval and provide an analysis of its characteristics (in Sec. 3.3), (ii) we provide more extensive baselines for the audio retrieval task with more detailed ablations across datasets and an additional retrieval architecture (Sec. 3.5). In particular, we explore the use of the Multi-Modal Transformer architecture [Gabeur et al. 2020].

3.2 Related work

Our work relates to several themes in the literature: *sound event recognition*, *audio captioning*, *audio-based retrieval*, *text-based video retrieval* and *text-domain audio retrieval*. We discuss each of these next.

Sound event recognition. There is a rich literature addressing the task of sound event recognition, which seeks to assign a given segment of audio to a corresponding semantic label. Examples include detecting audio events associated with sports [Xiong et al. 2003], urban sounds [Aucouturier et al. 2007], and distinguishing vocal and nonvocal events [Atrey et al. 2006]. Research in this area has been driven by challenges, such as DCASE [Stowell et al. 2015; Mesaros et al. 2017a], and by the collection of sound event datasets. These include TUT Acoustic scenes [Mesaros et al. 2017b], CHIME-Home [Foster et al. 2015], ESC-50 [Piczak 2015], FSDKaggle [Fonseca et al. 2019], and AudioSet [Gemmeke et al. 2017]. Of relevance to our approach, a number of prior works have employed deep learning for audio comprehension [Kong et al. 2018; Yu et al. 2018; Kong et al. 2019; Ford et al. 2019; Kong et al. 2020]. Our work differs from theirs in that we focus instead on the task of retrieval with natural language queries, rather than audio recognition.

Audio captioning. Audio captioning consists of generating a natural language description for a sound [Drossos et al. 2017]. This requires a more detailed understanding of the sound than simply mapping the sound to a set of labels (sound event recognition). Recently, several audio captioning datasets have been introduced, such as Clotho [Drossos et al. 2020] which was used in the DCASE automated audio captioning challenge 2020 [*DCASE2020 Challenge Task 6: Automated Audio Captioning* 2020], Audio Caption [M. Wu et al. 2019], and AudioCaps [C. D. Kim et al. 2019]. Drawing inspiration from work on video captioning [Venugopalan et al. 2015; L. Gao et al. 2017], multiple works have addressed automatic audio captioning on the AudioCaps and Clotho datasets [Koizumi et al. 2020; X. Xu et al. 2021a; Eren and Sert 2020; Mei et al. 2021; X. Liu et al. 2021]. In this work, we use the AudioCaps and Clotho datasets for cross-modal retrieval.

Audio-based retrieval. Multiple content-based audio retrieval frameworks, in particular query by example methods, leverage the similarity of sound features that represent different aspects of sounds (e.g. pitch, or loudness) [Foote 1997;

Wold et al. 1996; Helén and Virtanen 2007; Lallemand et al. 2012]. More recently, [Manocha et al. 2018] use a twin neural network framework to learn to encode semantically similar audio close together in the embedding space. [Q. Jin et al. 2012] address multimedia event detection using only audio data, while [Avgoustinakis et al. 2020] tackle near-duplicate video retrieval by audio retrieval. These are purely audio-based methods that are applied to video datasets, but without using visual information. [Hou and S. Zhou 2013] propose a two-step approach for video retrieval which uses audio (coarse) and visual (fine) information together.

Text-based video retrieval. More closely related to our work, a number of methods showed that embedding video and text jointly into a shared space (such that their similarity can be computed efficiently) is an effective approach [Dong et al. 2016; Mithun et al. 2018; Miech et al. 2018; Y. Liu et al. 2019; Wray et al. 2019; Gabeur et al. 2020; Croitoru et al. 2021] (though other formulations, such as computing similarities directly in visual space have also been explored [Dong et al. 2018]). One particular trend has been to combine cues from several "experts"—pre-trained models that specialise in different tasks (such as object recognition, action classification etc.) to inform the joint embedding. Recently, transformer-based architectures have demonstrated impressive results for text-based video retrieval [Gabeur et al. 2020; Bain et al. 2021; Luo et al. 2021]. In this work, we propose to adapt three expert-based methods: the Mixture of Embedded Experts method of [Miech et al. 2018], the Collaborative Experts model of [Y. Liu et al. 2019], and the Multi-Modal Transformer [Gabeur et al. 2020] by re-purposing them for the task of audio retrieval (described in more detail in Sec. 3.4).

Cross-domain audio retrieval. Methods that retrieve audio by matching associated text, such as metadata or sound event labels, have the implicit assumption that the text is relevant [Elizalde et al. 2018]. In contrast, [Slaney 2002b] is an early work that proposes to link audio and text representations in hierarchical semantic and acoustic spaces. [Slaney 2002a] builds on this using mixture-of-probability-expert models for each of the modalities. Chechik et al. [Chechik et al. 2008] propose a text-based sound retrieval framework which uses single-word audio tags as queries rather than caption-like natural language. Similarly, [Ikawa and Kashino 2018a; Ikawa and Kashino 2018b] learn shared latent spaces between onomatopoeias (words that mimic non-speech sounds) and sound for searching

audio using onomatopoeia queries and for generating sound words from audio. The creative approach of [Aytar et al. 2017] learns to align visual, audio, and text representations to enable cross-modal retrieval. Their framework is trained with captioned images and paired image-sound data (sourced from videos) and evaluated using the soundtrack of captioned videos. Other works have explored using images [Harwath et al. 2018] or video data [Yasuda et al. 2020; Surís et al. 2018; Zeng et al. 2020; C. Jin et al. 2021; Nagrani et al. 2018] as queries for retrieving audio. More recently, [Elizalde et al. 2019] use a twin network to learn a shared latent text and sound space for cross-modal retrieval. While they use class labels as text labels, we study unconstrained text descriptions as queries. Another highly related, concurrent line of works has explored the task of grounding sounds given a text description. [X. Xu et al. 2021b; Tang et al. 2021] presented results for the grounding task on the AudioGrounding dataset, proposed by [X. Xu et al. 2021b], which augments a subset of the AudioCaps dataset (approximately 10% of the full AudioCaps dataset) by adding fine-grained temporal grounding labels. In contrast to the audio grounding task, text-based audio retrieval does not require expensive temporal annotations and we can therefore leverage large databases which contain immensely varied content.

3.3 SoundDescs dataset

In this section, we introduce the SoundDescs dataset for text-audio retrieval. The SoundDescs dataset consists of 32,979 audio files accompanied by natural language descriptions. We present an overview of the SoundDescs dataset by describing how

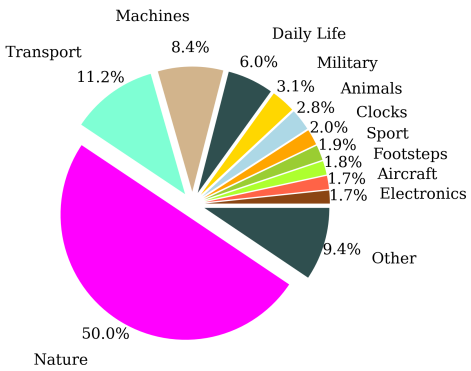


Figure 3.1: Pie chart showing the distribution of audio files in the SoundDescs dataset over different categories.

Table 3.1: Comparative overview of sound-language datasets. Comparing the number of files in the different datasets, including their training, validation, and test subsets, audio duration (dur.) and caption lengths (#words) measured in seconds and words respectively. Text is sourced from human caption annotations or audio descriptions provided with the sound data.

Dataset	Text source	language	duration(h)	#audios	#captions	max dur.(s)	avg dur.(s)	max #words	avg #words	train	val	test
AudioCaps [C. D. Kim et al. 2019]	Human captions	English	135.01	50535	55512	10.08	9.84	52	8.80	49291	428	816
Clotho [Drossos et al. 2020]	Human captions	English	22.55	3938	3938	30.00	22.44	21	11.32	2314	579	1045
Audio Caption [M. Wu et al. 2019]	Human captions	Chinese	10.3	3707	3707	N/A	10.00	54	11.14	3337	-	371
SoundDescs	Descriptions	English	1060.40	32979	32979	4475.89	115.75	65	15.28	23085	4947	4947

it was collected in Section 3.3.1, and by providing an analysis and comparison to related datasets which contain pairs of audio files and text descriptions in Section 3.3.2. Furthermore, we show some examples of the data in Section 3.3.3.

3.3.1 Dataset collection

The SoundDescs data was sourced from the BBC Sound Effects webpage. It contains audio files and corresponding textual descriptions of a wide range of sounds from the BBC Radiophonic workshop, the Blitz in London, special effects made for the BBC, and recordings from the Natural History Unit archive. The sounds are of high quality and were recorded for professional applications, such as radio and TV special effects. In some cases, additional information is provided which contains the size and number of channels of the audio file together with the date it was recorded and other sound tags. The audio files are associated with 23 categories, including but not limited to *nature*, *clocks*, *fire*, etc. We show the proportion of files from the different categories in Fig. 3.1. For this, we use the text tags accompanying the collected audio files. Since some of the files contain multiple text tags, we collected all of them in a bag-of-words fashion and then provided the frequencies with which they appear. We obtained 32,979 audio files sampled at 44.1 Hz which have a non-empty textual description (out of the 33,066 audio files on the BBC website). We propose to split the SoundDescs dataset into training/validation/test subsets by randomly selecting 70% of the files for training, and 15% each for validation and testing.

3.3.2 Data analysis

We compare the SoundDescs dataset to related datasets which contain matching sound and language data in Table 3.1. The two main novelties of the SoundDescs

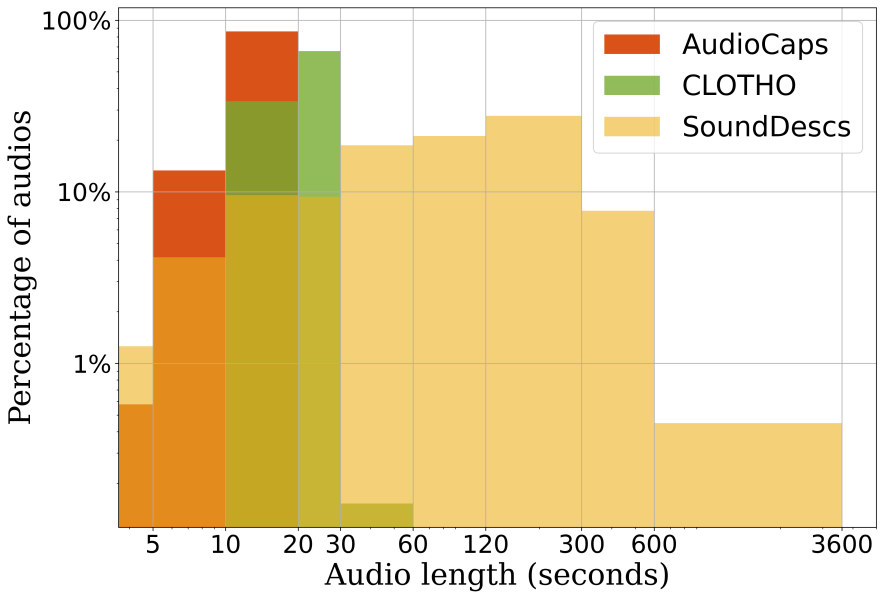


Figure 3.2: Histogram of audio files lengths for the AudioCaps, Clotho and SoundDescs datasets.

dataset compared to related sound-text datasets are the wide variation in duration of the audio files and the size of the vocabulary used for the descriptions. The audio captioning datasets AudioCaps [C. D. Kim et al. 2019] and Clotho [Drossos et al. 2020] contain audio files that are only 10-30 seconds long. As can be seen in Fig. 3.2, SoundDescs contains sounds with a much wider range of audio durations with 109 files lasting longer than 10 minutes. Processing audio files with such variations in duration is challenging, but it enables the detailed analysis of the performance of current text-audio and audio-text retrieval methods with respect to the audio duration. In addition to this, the SoundDescs dataset is larger than all related sound-language datasets with a total duration of 1060 hours, compared to 135 hours for AudioCaps. Since SoundDescs contains fewer audio files with associated captions than AudioCaps, the SoundDescs dataset presents a more challenging retrieval benchmark dataset. Furthermore, the average audio duration and average length of the text descriptions in SoundDescs is significantly higher than for AudioCaps or Clotho. The word length distributions for these datasets can be seen in Fig. 3.3. Interestingly, for the Clotho and SoundDescs datasets which contain audio and descriptions with varied lengths, there is no strong correlation between the audio and description lengths.

Lastly, the descriptions in the SoundDescs dataset use a larger vocabulary with

The numbers provided for the AudioCaps dataset in Table 3.1 correspond to the subset which does not have an overlap with the VGGSound dataset.

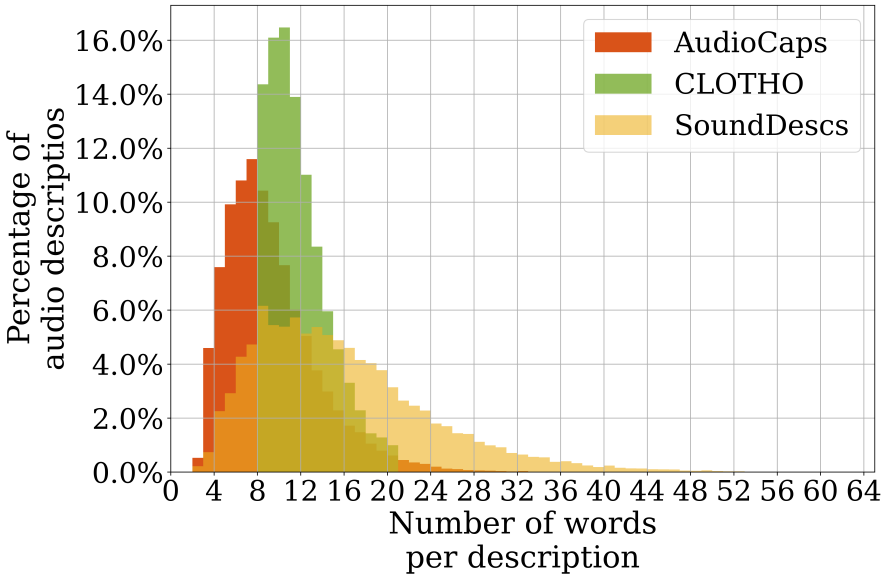


Figure 3.3: Distribution of the number of words per caption for the AudioCaps, Clotho and SoundDescs datasets.

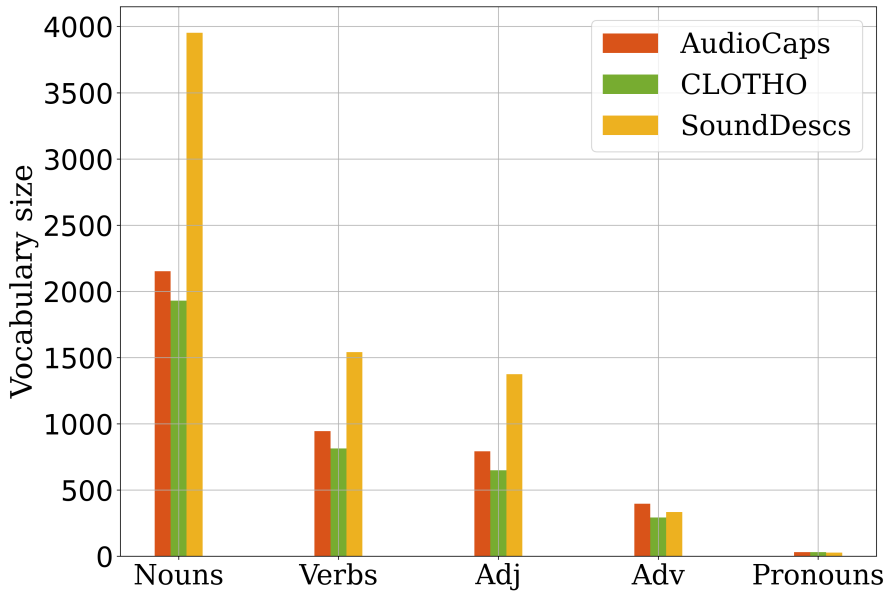


Figure 3.4: Vocabulary size of descriptions in the SoundDescs dataset, compared to the AudioCaps and Clotho datasets, divided into nouns, verbs, adjectives, adverbs, and pronouns.

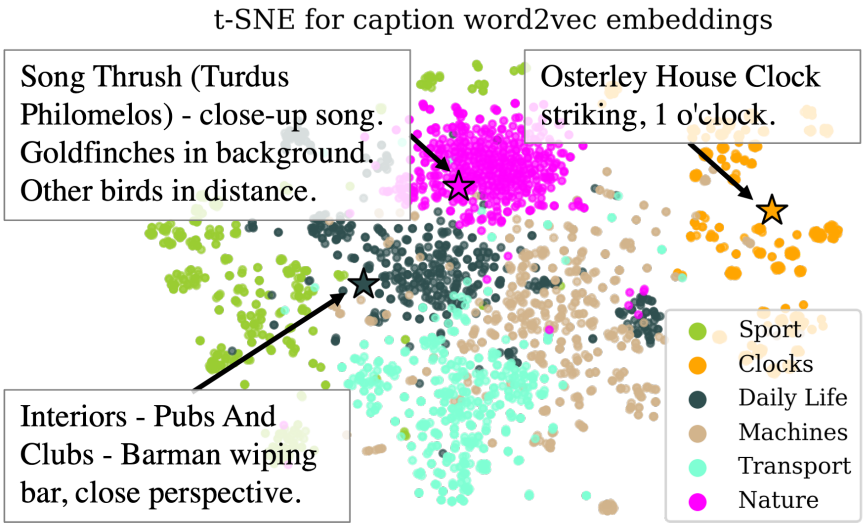


Figure 3.5: t-SNE visualisation for word2vec description embeddings on SoundDescs. Example descriptions corresponding to embeddings marked with a star are shown in text boxes.

respect to nouns, verbs, adjectives used, compared to the AudioCaps and Clotho datasets (see Fig. 3.4). In particular, the descriptions in SoundDescs contain almost 4000 distinct nouns as apposed to around 2000 in AudioCaps and Clotho. This is consistent with the wide array of topics (Fig. 3.1), which reflects a high diversity of environments in which the sound effects were recorded, and thus hugely varied sources of sounds are identified in the descriptions. A large contributor to this diversity in nouns is the high proportion of Nature sounds (Fig. 3.1), for which species names are often specified (an example is shown in Fig. 3.5).

Further details about the related datasets that we use in this work can be found in Section 3.5. A more in depth analysis of the SoundDescs's text descriptions can be found in the Appendix.

3.3.3 Dataset examples

In Fig. 3.5 we visualise the distribution of the descriptions in SoundDescs, showing the full descriptions for some examples. The averaged word2vec [Mikolov et al. 2013] vectors extracted from each description are embedded using t-SNE [Van der Maaten and Hinton 2008], and colour-coded according to the category (Fig. 3.1). We show 500 randomly chosen samples per class for 6 out of the 23 categories present in SoundDescs. We can observe that the descriptions cluster smoothly according to the categories, and that the descriptions are fairly specific and of

high quality, despite not always resembling full sentences.

3.3.4 Rights to use

Data collected for generating the new SoundDescs dataset is protected by the BBC RemArc licence. This allows the data to be used for non-commercial, personal or research purposes. We have released the list with urls along with code for downloading and constructing the SoundDescs dataset with our proposed train/val/test split.

3.4 Methods, datasets, and benchmark

In this section, we first formulate the problem of audio retrieval with natural language queries. Next, we describe three cross-modal embedding methods that we adapt for the task of audio retrieval (Section 3.4.1). Finally, we describe the five datasets used in our experimental study (Section 3.4.2), and the three benchmarks that we propose for evaluating performance on the audio retrieval task (Section 3.4.3).

Problem formulation. Given a natural language query (*i.e.* a written description of an audio event to be retrieved) and a pool of audio samples, the objective of text-audio (abbreviated to `t2a`) retrieval is to rank the audio samples according to their similarity to the query. We also consider the converse `a2t` task, viz. retrieving text with audio queries.

3.4.1 Methods

To tackle the problem of text-audio retrieval, we propose to learn cross-modal embeddings. Specifically, given a collection of N audio samples with corresponding textual descriptions, $\{(a_i, t_i) : i \in \{1, \ldots, N\}\}$, we aim to learn embedding functions, ψ_a and ψ_t, that project each audio sample a_i and text sample t_i into a shared space, such that $\psi_a(a_i)$ and $\psi_t(t_i)$ are close when the text describes the audio, and far apart otherwise. Writing s_{ij} for the cosine similarity of the audio embedding $\psi_a(a_i)$ and the text embedding $\psi_t(t_j)$, we learn the embedding functions

https://sound-effects.bbcrewind.co.uk/licensing

by minimising a contrastive ranking loss [Socher et al. 2014]:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1, j \neq i}^{B} [m + s_{ij} - s_{ii}]_+ + [m + s_{ji} - s_{ii}]_+ \tag{3.1}$$

where B denotes the batch size, m the *margin* (set as a hyperparameter), and $[\cdot]_+ = max(\cdot, 0)$ the hinge function.

We consider three recent state-of-the-art frameworks for learning such embedding functions ψ_a and ψ_t: Mixture-of-Embedded Experts (MoEE) [Miech et al. 2018], Collaborative-Experts (CE) [Y. Liu et al. 2019], and the Multi-Modal Transformer (MMT) [Gabeur et al. 2020]. All three frameworks were originally designed for text-video retrieval and construct their video encoder from a collection of "experts" (features extracted from networks pre-trained for object recognition, action classification, sound classification, etc.) which are computed for each video.

MoEE and CE aggregate experts along their temporal dimension with NetVLAD [Arandjelovic et al. 2016], and project them to a lower dimension via a self-gated linear map followed by L2-normalisation. Their text encoder first embeds each word token with word2vec [Mikolov et al. 2013] and aggregates the results with NetVLAD [Arandjelovic et al. 2016]. The result is projected by a sequence of self-gated linear maps (one for each expert) into a shared embedding space with the outputs of the video encoder. Finally, a scalar-weighted sum of the embedded experts in each joint space is used to compute the overall cosine similarity between the video and text (see [Miech et al. 2018] for more details). CE adopts the same text encoder as MoEE and similarly makes use of multiple video experts. However, rather than projecting them directly into independent spaces against the embedded text, CE first applies a "collaborative gating" mechanism, which filters each expert with an element-wise attention mask that is generated with a small MLP that ingests all pairwise combinations of experts (see [Y. Liu et al. 2019] for further details).

MMT, on the other hand, uses a multi-modal transformer encoder which refines the expert embeddings by passing through multiple multi-head self-attention layers. Differently from the temporal aggregation of expert features used by MoEE and CE, expert features are instead passed as input directly to the transformer encoder. This allows to iteratively focus on the relevant information from each

expert at multiple time steps by comparing content from different experts, rather than down-projecting the expert embeddings in a single step (as is the case for MoEE and CE). Each input expert uses an aggregated embedding which collects the information for that expert through the layers and serves as the output expert representation. The iterative attention across different modalities and time steps enables MMT to combine information across input experts, rather than comparing each expert individually to the text query embeddings. MMT leverages a pre-trained BERT model [Devlin et al. 2018] to extract text embeddings. Gated embedding functions are learned to obtain a text embedding for each video expert. The similarity s_{ij} between the aggregated audio and text embeddings is computed as the weighted sum of the similarity between each expert and the text embedding. Further information about MMT can be found in [Gabeur et al. 2020].

To adapt the MoEE, CE, and MMT frameworks for audio retrieval, we use the same text encoder structures as used for video retrieval ψ_t. We build audio encoders ψ_a by mimicking the structure of their video encoders, replacing the "video experts" with "audio experts" (described in Sec. 3.5).

3.4.2 Datasets

As the primary focus of our work, we study three *audio-centric datasets*—these are datasets which comprise audio streams (sometimes with accompanying visual streams) paired with natural language descriptions that focus explicitly on the content of the audio track. To explore differences between audio retrieval and video retrieval, we also consider two *visual-centric datasets*, that comprise audio and video streams paired with natural language which focus primarily (though not always exclusively) on the content of the video stream. Details of the five datasets we employ are given next.

1. SoundDescs *(audio-centric)* consists of sounds sourced from the BBC Sound Effects webpage and annotated with free-form text descriptions. It contains 23,085 training, 4947 validation and 4947 test samples. Further information on our newly introduced SoundDescs dataset is provided in Section 3.3.

2. AudioCaps [C. D. Kim et al. 2019] *(audio-centric)* is a dataset of sounds with

The sample list is publicly available at the project page [*Audio-Retrieval Project Page* n.d.].

event descriptions which was introduced for the task of audio captioning, with sounds sourced from the AudioSet dataset [Gemmeke et al. 2017]. Annotators were provided the audio tracks together with category hints (and with additional video hints if needed). We use a subset of the data, excluding a small number of samples for which either: (i) the YouTube-hosted source video is no longer available, (ii) the source video overlaps with the training partition of the VGGSound dataset [H. Chen et al. 2020]. Filtering to exclude samples affected by either issue leads to a dataset with 49,291 training, 428 validation and 816 test samples.

3. Clotho [Drossos et al. 2020] *(audio-centric)* is a dataset of described sounds that was also introduced for the task of audio captioning, with sounds sourced from the Freesound platform [Font et al. 2013]. During labelling, annotators only had access to the audio stream (i.e. no visual stream or meta tags) to avoid their reliance on contextual information for removing ambiguity that could not be resolved from the audio stream alone. The descriptions are filtered to exclude transcribed speech. The publicly available version of the dataset includes a `dev` set of 2893 audio samples and an `evaluation` set of 1045 audio samples. Every audio sample is accompanied by 5 written descriptions. We used a random split of the `dev` set into a training and validation set with 2,314 and 579 samples, respectively.

4. ActivityNet-Captions [Krishna et al. 2017] (visual-centric) consists of videos sourced from YouTube and annotated with dense event descriptions. It allocates 10,009 videos for training and 4,917 videos for testing (we use the public `val_1` split provided by [Krishna et al. 2017]). For this dataset, descriptions also tend to focus on the visual stream.

5. QuerYD [Oncescu et al. 2021a] (visual-centric) is a dataset of described videos sourced from YouTube and the YouDescribe [Research and Center 2013] platform. It is accompanied by *audio descriptions* that are provided with the explicit aim of conveying the video content to visually impaired users. Therefore, the provided descriptions focus heavily on the visual modality. We use the version of the dataset comprising trimmed videos with 9,114 training, 1,952 validation, and 1,954 test samples.

3.4.3 Benchmark

To facilitate the study of the text-based audio retrieval task, we introduce the SoundDescs dataset and propose to re-purpose the two *audio-centric* datasets described above, AudioCaps and Clotho, to provide benchmarks for text-based audio retrieval. The approach is inspired by precedents in the vision and language communities, where datasets, such as [J. Xu et al. 2016], that were originally introduced for the task of video captioning, have become popular benchmarks for text-based video retrieval [Miech et al. 2018; Wray et al. 2019; Y. Liu et al. 2019; Gabeur et al. 2020].

3.5 Experiments

In this section, we first compare text-audio retrieval (`t2a`) and audio-text retrieval (`a2t`) performance on audio-centric and visual-centric datasets. Next, we perform an ablation study on the contributions of different experts and present our baselines for the proposed AudioCaps, Clotho, and SoundDescs benchmarks. Finally we perform experiments to assess the influence of pre-training, audio segment duration and training dataset size, and give qualitative examples of retrieval results. Throughout the section, we use the standard retrieval metrics: recall at rank k (R@k) which measures the percentage of targets retrieved within the top k ranked results (higher is better), along with the median (medR) and mean (meanR) rank. For all metrics, we report the mean and standard deviation of three different randomly seeded runs.

Implementation details. We use pre-trained feature extractors to obtain audio and visual expert features. To encode the audio signal, we use two pre-trained audio feature extractors which we refer to as VGGish and VGGSound. We explain both in more detail in the following.

VGGish. These audio features are obtained with a VGGish model [Hershey et al. 2016], trained for audio classification on the YouTube-8M dataset [Abu-El-Haija et al. 2016]. To produce the input for the VGGish model, the audio stream of each video is re-sampled to a 16kHz mono signal, converted to an STFT with a window

Since the AudioCaps test set is a subset of the AudioSet training set (unbalanced), we do not use audio experts pre-trained on AudioSet.

size of 25ms and a hop size of 10ms with a Hann window, then mapped to a 64 bin log mel spectrogram. Finally, the features are parsed into non-overlapping 0.96s collections of frames (each collection comprises 96 frames, each of 10ms duration), which are mapped to a 128-dimensional feature vector.

VGGSound. These features are extracted using a ResNet-18 model [He et al. 2016] that has been pre-trained on the VGGSound dataset (model H) [H. Chen et al. 2020]. We modify the last average pooling layer to aggregate along the frequency dimension, but keep the full temporal dimension. This results in features of dimension t x 512, where t denotes the number of time steps.

For the results with visual experts in Table 3.3, we employed a subset of visual feature extractors used in [Y. Liu et al. 2019] which we refer to as Inst, Scene, and R2P1D. We we will describe each of those in more detail below.

Inst. These features are extracted using a ResNeXt-101 model [R. Xu et al. 2015] that has been pre-trained on Instagram hashtags [Mahajan et al. 2018] and fine-tuned on ImageNet [Deng et al. 2009] for the task of image classification. Features are obtained from frames extracted at 25 fps, where each frame is resized to 224 $\times$ 224 pixels. Embeddings are 2048-dimensional.

Scene. Scene features are extracted from 224$\times$224 pixel centre crops with a DenseNet-161 model [G. Huang et al. 2017] pre-trained on Places365 [B. Zhou et al. 2017]. Embeddings are 2208-dimensional.

R2P1D. Features are extracted with a 34-layer R(2+1)D model [Tran et al. 2018] trained on IG-65M [Ghadiyaram et al. 2019] which processes clips of 8 consecutive 112 $\times$ 112 pixel frames, extracted at 30 fps. Embeddings are 512-dimensional.

Training All models were trained using the contrastive ranking loss (Eqn.(3.1)), with m set to 0.2 for CE and MoEE, and 0.05 for MMT (those parameters were taken from [Y. Liu et al. 2019] and [Gabeur et al. 2020] respectively). CE and MoEE models were trained with a batch size of 128 for 20 epochs and the models that gave the best performance on the geometric mean of R@1, R@5, and R@10 were chosen as the final models. MMT was trained with a batch size of 32 for 50K steps.

For CE and MoEE, we used the Lookahead solver [M. R. Zhang et al. 2019] in

Table 3.2: Audio retrieval on audio-centric and visual-centric datasets. Performance is strongest on the audio-centric SoundDescs dataset and weakest on the visual-centric ActivityNet-Captions (ActNetCaps) dataset.

			Text $\Longrightarrow$ Audio		Audio $\Longrightarrow$ Text	
Dataset	Anno. Focus	Pool	R@1 ↑	R@10 ↑	R@1 ↑	R@10 ↑
AudioCaps [C. D. Kim et al. 2019]	audio	816	$18.5_{\pm 0.3}$	$62.0_{\pm 0.5}$	$20.7_{\pm 1.8}$	$62.9_{\pm 0.4}$
Clotho [Drossos et al. 2020]	audio	1045	$4.0_{\pm 0.2}$	$25.4_{\pm 0.5}$	$4.8_{\pm 0.4}$	$25.8_{\pm 1.7}$
SoundDescs	audio	4947	$25.4_{\pm 0.6}$	$64.1_{\pm 0.3}$	$24.2_{\pm 0.3}$	$62.5_{\pm 0.2}$
ActNetCaps [Krishna et al. 2017]	visual	4917	$1.4_{\pm 0.1}$	$8.5_{\pm 0.2}$	$1.1_{\pm 0.1}$	$7.9_{\pm 0.0}$
QuerYD [Oncescu et al. 2021a]	visual	1954	$3.7_{\pm 0.2}$	$17.3_{\pm 0.6}$	$3.8_{\pm 0.2}$	$16.8_{\pm 0.2}$

combination with RAdam [L. Liu et al. 2019] (implementation by [*Ranger optimiser* n.d.]) with an initial learning rate of 0.01 and weight decay of 0.001. We use a learning rate decay for each parameter group with a factor of 0.95 every epoch. MMT was trained using Adam and a learning rate of 0.00005, which was decayed by a multiplicative factor 0.95 every 1K optimisation steps.

For the NetVLAD module in CE and MoEE, we used 20 VLAD clusters and one ghost cluster [Zhong et al. 2018] for text, and 16 VLAD clusters for the audio features. On AudioCaps, we used a maximum of 52 word tokens, a maximum of 10 time frames for VGGish and a maximum of 32 time frames for VGGSound features. For Clotho, we used a maximum of 21 word tokens, a maximum of 31 time frames for VGGish, and 95 VGGSound features per sample (95 for both audio feature experts for MMT). For SoundDescs, the maximum number of word tokens was set to 46, and audio time frames used to 400 (both for VGGish and VGGSound).

Audio-centric vs. visual-centric queries. We first investigate audio retrieval with audio-centric and visual-centric queries (The audio-centric datasets contain audio-centric queries, and the video-centric datasets contain video-centric queries.). For this experiment, we use CE with a single expert (VGGish audio features). In Table 3.2, we observe that performance is strongest on the audio-centric datasets SoundDescs and AudioCaps, and it is weakest overall on the visual-centric ActivityNet-Captions dataset. This is expected, since visual-centric queries contain information that is not captured in the audio data. We note that the Clotho dataset is particularly challenging, with performance weaker (accounting for pool size) than the visual-centric QuerYD dataset. We hypothesise that this is for two reasons: (1) the significantly smaller training set size of Clotho, compared to all

Table 3.3: The influence of different experts on AudioCaps. A comparison of audio and visual experts (applied to the video from which the audio was sourced) using CE [Y. Liu et al. 2019]. Audio features are significantly more effective than visual features (which nevertheless provide some complementary signal as can be seen when jointly using audio and visual features).

	Text $\Longrightarrow$ Audio/Video		Audio/Video $\Longrightarrow$ Text	
Expert	R@1 ↑	R@10 ↑	R@1 ↑	R@10 ↑
Visual experts only				
Scene	$6.0_{\pm 0.0}$	$35.6_{\pm 0.8}$	$6.8_{\pm 0.6}$	$31.9_{\pm 1.3}$
Inst	$8.2_{\pm 0.3}$	$46.2_{\pm 0.5}$	$10.1_{\pm 0.8}$	$41.3_{\pm 0.6}$
R2P1D	$8.1_{\pm 0.4}$	$45.8_{\pm 0.2}$	$10.7_{\pm 0.1}$	$43.4_{\pm 1.9}$
Scene + Inst	$8.2_{\pm 0.3}$	$47.1_{\pm 0.2}$	$10.2_{\pm 1.2}$	$41.5_{\pm 1.3}$
Scene + R2P1D	$8.6_{\pm 0.1}$	$47.4_{\pm 0.2}$	$11.6_{\pm 0.4}$	$43.5_{\pm 0.8}$
R2P1D + Inst (*CE-Visual*)	$\mathbf{9.5_{\pm 0.6}}$	$\mathbf{50.0_{\pm 0.5}}$	$\mathbf{11.2_{\pm 0.1}}$	$\mathbf{45.2_{\pm 1.9}}$
Audio experts only				
VGGish	$18.5_{\pm 0.3}$	$62.0_{\pm 0.5}$	$20.7_{\pm 1.8}$	$62.9_{\pm 0.4}$
VGGSound	$22.4_{\pm 0.3}$	$69.2_{\pm 0.9}$	$27.0_{\pm 0.9}$	$72.5_{\pm 0.7}$
VGGish + VGGSound (*CE-Audio*)	$\mathbf{23.6_{\pm 0.6}}$	$\mathbf{71.4_{\pm 0.5}}$	$\mathbf{27.6_{\pm 1.0}}$	$\mathbf{74.7_{\pm 0.8}}$
Audio and visual experts				
CE-Visual + VGGish	$24.5_{\pm 0.8}$	$74.9_{\pm 1.0}$	$31.0_{\pm 2.2}$	$78.8_{\pm 1.2}$
CE-Visual + VGGSound	$27.6_{\pm 0.2}$	$78.0_{\pm 0.8}$	$32.7_{\pm 0.9}$	$82.4_{\pm 0.4}$
CE-Visual + *CE-Audio*	$\mathbf{28.0_{\pm 0.5}}$	$\mathbf{80.4_{\pm 0.3}}$	$\mathbf{35.8_{\pm 0.6}}$	$\mathbf{83.3_{\pm 0.6}}$

Table 3.4: Audio retrieval on AudioCaps, Clotho, and SoundDescs, using the CE, MoEE, and MMT frameworks. Retrieval performance metrics for text to audio and audio to text retrieval reported are recall at rank k (R@1, R@5, R@10), and the median and mean rank (medR and meanR). Audio experts used are obtained from the VGGish model [Gemmeke et al. 2017] and from a ResNet18 model pre-trained on VGGSound [H. Chen et al. 2020].

	Text $\Longrightarrow$ Audio						Audio $\Longrightarrow$ Text					
Dataset	R@1 ↑	R@5 ↑	R@10 ↑	R@50 ↑	medR ↓	meanR ↓	R@1 ↑	R@5 ↑	R@10 ↑	R@50 ↑	medR ↓	meanR ↓
AudioCaps												
CE	$23.6_{\pm 0.6}$	$56.2_{\pm 0.5}$	$71.4_{\pm 0.5}$	$92.3_{\pm 1.5}$	$4.0_{\pm 0.0}$	$18.3_{\pm 3.0}$	$27.6_{\pm 1.0}$	$60.5_{\pm 0.7}$	$74.7_{\pm 0.8}$	$94.2_{\pm 0.4}$	$4.0_{\pm 0.0}$	$14.7_{\pm 1.4}$
MoEE	$23.0_{\pm 0.7}$	$55.7_{\pm 0.3}$	$71.0_{\pm 1.2}$	$93.0_{\pm 0.3}$	$4.0_{\pm 0.0}$	$16.3_{\pm 0.5}$	$26.6_{\pm 0.7}$	$59.3_{\pm 1.4}$	$73.5_{\pm 1.1}$	$94.0_{\pm 0.5}$	$4.0_{\pm 0.0}$	$15.6_{\pm 0.8}$
MMT	$36.1_{\pm 3.3}$	$72.0_{\pm 2.9}$	$84.5_{\pm 2.0}$	$97.6_{\pm 0.4}$	$2.3_{\pm 0.6}$	$7.5_{\pm 1.3}$	$39.6_{\pm 0.2}$	$76.8_{\pm 0.9}$	$86.7_{\pm 1.8}$	$98.2_{\pm 0.4}$	$2.0_{\pm 0.0}$	$6.5_{\pm 0.5}$
Clotho												
CE	$6.7_{\pm 0.4}$	$21.6_{\pm 0.6}$	$33.2_{\pm 0.3}$	$69.8_{\pm 0.3}$	$22.3_{\pm 0.6}$	$58.3_{\pm 1.1}$	$7.0_{\pm 0.3}$	$22.7_{\pm 0.6}$	$34.6_{\pm 0.5}$	$67.9_{\pm 2.3}$	$21.3_{\pm 0.6}$	$72.6_{\pm 3.4}$
MoEE	$6.0_{\pm 0.1}$	$20.8_{\pm 0.7}$	$32.3_{\pm 0.3}$	$68.5_{\pm 0.5}$	$23.0_{\pm 0.0}$	$60.2_{\pm 0.8}$	$7.2_{\pm 0.5}$	$22.1_{\pm 0.7}$	$33.2_{\pm 1.1}$	$67.4_{\pm 0.3}$	$22.7_{\pm 0.6}$	$71.8_{\pm 2.3}$
MMT	$6.5_{\pm 0.6}$	$21.6_{\pm 0.7}$	$32.8_{\pm 2.1}$	$66.9_{\pm 2.0}$	$23.0_{\pm 2.6}$	$67.7_{\pm 3.1}$	$6.3_{\pm 0.5}$	$22.8_{\pm 1.7}$	$33.3_{\pm 2.2}$	$67.8_{\pm 1.5}$	$22.3_{\pm 1.5}$	$67.3_{\pm 2.9}$
SoundDescs												
CE	$31.1_{\pm 0.2}$	$60.6_{\pm 0.7}$	$70.8_{\pm 0.5}$	$86.0_{\pm 0.2}$	$3.0_{\pm 0.0}$	$63.6_{\pm 2.2}$	$30.8_{\pm 0.8}$	$60.3_{\pm 0.3}$	$69.5_{\pm 0.1}$	$85.4_{\pm 0.2}$	$3.0_{\pm 0.0}$	$63.2_{\pm 0.6}$
MoEE	$30.8_{\pm 0.7}$	$60.8_{\pm 0.3}$	$70.9_{\pm 0.5}$	$85.9_{\pm 0.6}$	$3.0_{\pm 0.0}$	$62.0_{\pm 3.8}$	$30.9_{\pm 0.3}$	$60.3_{\pm 0.4}$	$70.1_{\pm 0.3}$	$85.3_{\pm 0.6}$	$3.0_{\pm 0.0}$	$61.5_{\pm 3.2}$
MMT	$30.7_{\pm 0.4}$	$61.8_{\pm 1.0}$	$72.2_{\pm 0.8}$	$88.8_{\pm 0.4}$	$3.0_{\pm 0.0}$	$34.0_{\pm 0.6}$	$31.4_{\pm 0.8}$	$63.2_{\pm 0.7}$	$73.4_{\pm 0.5}$	$89.0_{\pm 0.3}$	$3.0_{\pm 0.0}$	$32.5_{\pm 0.4}$

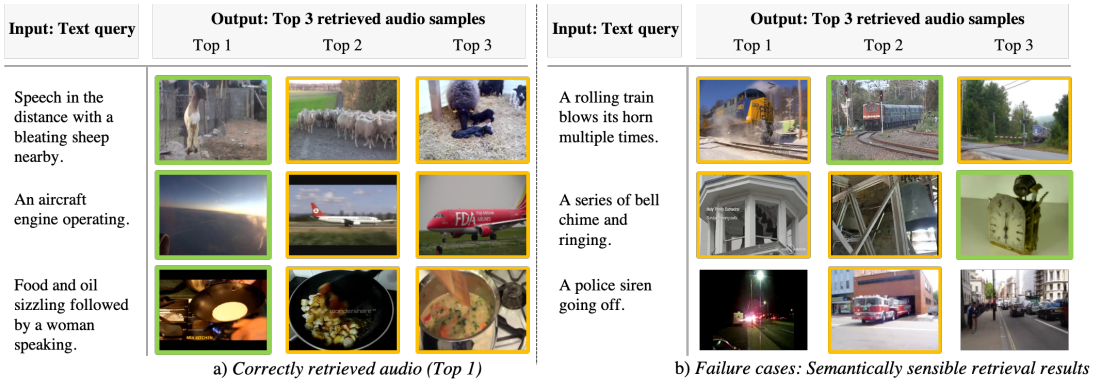


Figure 3.6: Qualitative results. Text-based audio retrieval results on AudioCaps using CE with VGGish and VGGSound features. For an input text query, we visualise the top 3 retrieved audio samples using a video frame from the corresponding videos (the audio can be heard at the project webpage [*Audio-Retrieval Project Page* n.d.]). We mark the audio samples which correspond to the query with green boxes. Successful retrievals are shown in a), failures in b). Note, in particular, the examples in b), where the model's top 1 retrieved audio is not the correct one, but the retrieved results nevertheless sound reasonable (visually convincing results are marked with yellow boxes).

other datasets, (2) Clotho was constructed such that the audio tag distribution resulted in varied audio content, making it a potentially more difficult benchmark. We note, however, that the QuerYD experiments suggest that computationally efficient video retrieval using only the audio stream can still be obtained, although at a lower accuracy.

Ablation study. We next conduct an ablation study to investigate the effectiveness of different audio and visual experts for audio retrieval on the AudioCaps dataset. We perform this experiment on the AudioCaps dataset, since the SoundDescs and Clotho datasets are audio-only datasets which implies that we cannot use any visual experts. We present the results in Table 3.3, where we observe that audio experts significantly outperform visual experts (pre-trained models for visual tasks like scene classification, which we compute from the video from which the audio was sourced). We note that the combination of audio and visual experts performs strongest overall, suggesting that the audio-centric queries contain information that is more accessible from the visual modality. The strongest audio-only retrieval is achieved by combining VGGish and VGGSound features—we therefore adopt this setting for the remaining experiments.

Additionally, we experimented with using speech features (word2vec [Mikolov et al. 2013] encodings of speech-to-text transcriptions [Google n.d.] of the audio stream).

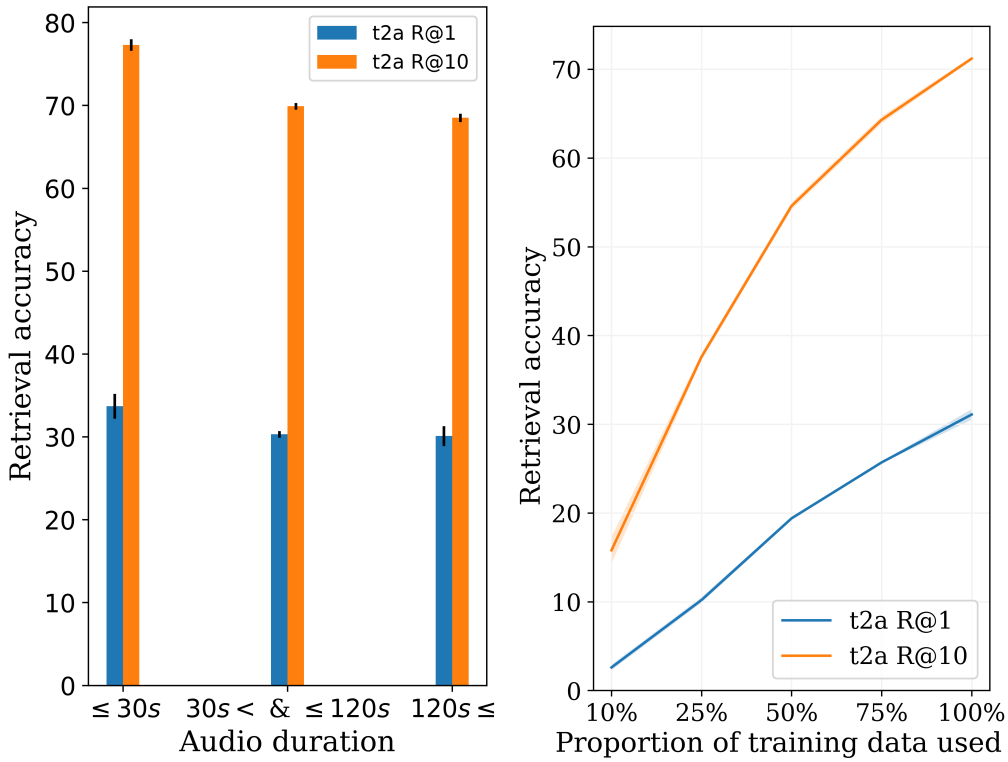
Table 3.5: Pre-training for audio retrieval. Text-audio and audio-text retrieval results for CE [Y. Liu et al. 2019] with VGGish and VGGSound features on the proposed SoundDescs, AudioCaps, and Clotho retrieval benchmarks. Pre-training on AudioCaps improves the performance on Clotho, and pre-training on SoundDescs slightly boosts the performance on AudioCaps.

	Text $\Longrightarrow$ Audio		Audio $\Longrightarrow$ Text	
Pre-training	R@1 ↑	R@10 ↑	R@1 ↑	R@10 ↑
AudioCaps				
None	$23.6_{\pm 0.6}$	$71.4_{\pm 0.5}$	$27.6_{\pm 1.0}$	$74.7_{\pm 0.8}$
SoundDescs	$24.6_{\pm 0.1}$	$72.2_{\pm 0.8}$	$27.8_{\pm 0.6}$	$75.2_{\pm 0.4}$
Clotho				
None	$6.7_{\pm 0.4}$	$33.2_{\pm 0.3}$	$7.0_{\pm 0.3}$	$34.6_{\pm 0.5}$
AudioCaps	$9.1_{\pm 0.3}$	$39.7_{\pm 0.4}$	$11.1_{\pm 1.1}$	$39.6_{\pm 1.1}$
SoundDescs	$6.4_{\pm 0.5}$	$32.5_{\pm 1.7}$	$6.1_{\pm 0.7}$	$31.4_{\pm 1.8}$
SoundDescs				
None	$31.1_{\pm 0.2}$	$70.8_{\pm 0.5}$	$30.8_{\pm 0.8}$	$69.5_{\pm 0.1}$
AudioCaps	$23.3_{\pm 0.7}$	$63.9_{\pm 0.5}$	$22.2_{\pm 0.4}$	$63.3_{\pm 0.3}$

However, this did not improve the retrieval performance. Upon further investigation, we found that the audio captions and the spoken words in the AudioCaps dataset do not have a significant overlap (corresponding to a METEOR [Banerjee and Lavie 2005] score of only 0.03 – perfect agreement between text sentences would give a score of 1).

Benchmark results. Incorporating the strongest combination of experts from the ablation study, we report our final baselines for text-audio and audio-text retrieval for three methods on the SoundDescs, AudioCaps, and Clotho datasets in Table 3.4. We observe that MMT outperforms CE and MoEE on the AudioCaps dataset for both text-audio and audio-text retrieval. For SoundDescs and Clotho, all three models yield comparable results.

We also report the performance after pre-training the CE model for retrieval on the AudioCaps or SoundDescs datasets and then fine-tuning on Clotho, AudioCaps, or SoundDescs in Table 3.5. Here, we observe that pre-training on SoundDescs brings a slight boost for AudioCaps and harms the performance on Clotho. This might be due to a larger domain gap between SoundDescs and Clotho compared to AudioCaps. Pre-training on AudioCaps improves the performance on Clotho, but is not beneficial for SoundDescs. Furthermore, we explored pre-training on Clotho and fine-tuning on the AudioCaps and SoundDescs datasets, but found negligible change in performance (likely due to the fact that the AudioCaps training set is significantly larger than that of Clotho).



(a) Audio duration.

(b) Training data size.

Figure 3.7: The influence of audio duration and training data scale on the text-audio retrieval performance on SoundDescs. Performance for the CE model is shown for the subsets of the SoundDescs test set according to different audio durations in a), and for different proportions of available training data in b).

Qualitative results. The qualitative results in Fig. 3.6 show examples in which the CE model with VGGish and VGGSound expert modalities (CE-Audio) is used to retrieve audio with natural language queries. The retrieved results mostly contain audio that is semantically similar to the input text queries. Observed failure cases arise from audio samples sounding very similar to one another despite being semantically distinct (e.g. the siren of a fire engine sounds very similar to a police siren).

Influence of audio segment duration. Next, we investigate the influence of audio segment duration on retrieval accuracy on the SoundDescs dataset. The retrieval accuracy for CE on different subsets of the test data according to their audio duration is shown in Fig. 3.7a. We show the performance for SoundDescs audio files with a duration of up to 30 seconds, those between 30 and 120 seconds, and those that are longer than 120 seconds. We observe that the retrieval performance is slightly weaker for longer audio segments than for those that last less than 30 seconds. However, the performance for very long audio files (longer than

120 seconds) is still solid.

Influence of training scale. Finally, we present experiments using different proportions of the SoundDescs dataset for training CE-Audio in Fig. 3.7b. As expected for deep learning frameworks, we observe that as more training data becomes available, the performance increases monotonically. We also observe that there is still clear room for improvement in terms of retrieval results simply by collecting additional training data, motivating further dataset construction work to support future research on this important task.

3.6 Conclusion

We introduced the novel SoundDescs dataset consisting of paired audio and text descriptions. Furthermore, we proposed three benchmarks for natural language based audio retrieval on the Clotho, AudioCaps, and SoundDescs datasets, and provided baseline results by adapting strong multi-modal video retrieval methods. Our results show that these methods are relatively well-suited for the audio retrieval task, however there is room for improvement, as expected for an under-explored problem. We hope that our proposed benchmarks will facilitate the development of future audio search engines, and make this large fraction of the world's produced media available for public use.

Acknowledgment

ASK and ZA were supported by the ERC (853489 - DEXIM), by the DFG (2064/1 – Project number 390727645), and by the BMBF (FKZ: 01IS18039A). AMO was supported by an EPSRC DTA Studentship. JFH is supported by the Royal Academy of Engineering (RF\201819\18\163). SA was supported by EPSRC EP/T028572/1 Visual AI. The authors would like to thank A. Zisserman for suggestions. SA would also like to thank Z. Novak and S. Carlson for support.

3.7 Supplementary material

In this appendix, we provide additional information about the SoundDescs dataset.

In Fig. 3.3, we presented a comparison of the description length distributions for the AudioCaps, Clotho, and SoundDescs datasets. Since descriptions in SoundDescs can contain one or more sentences, we also provide the distribution of sentence lengths in Fig. 3.8.

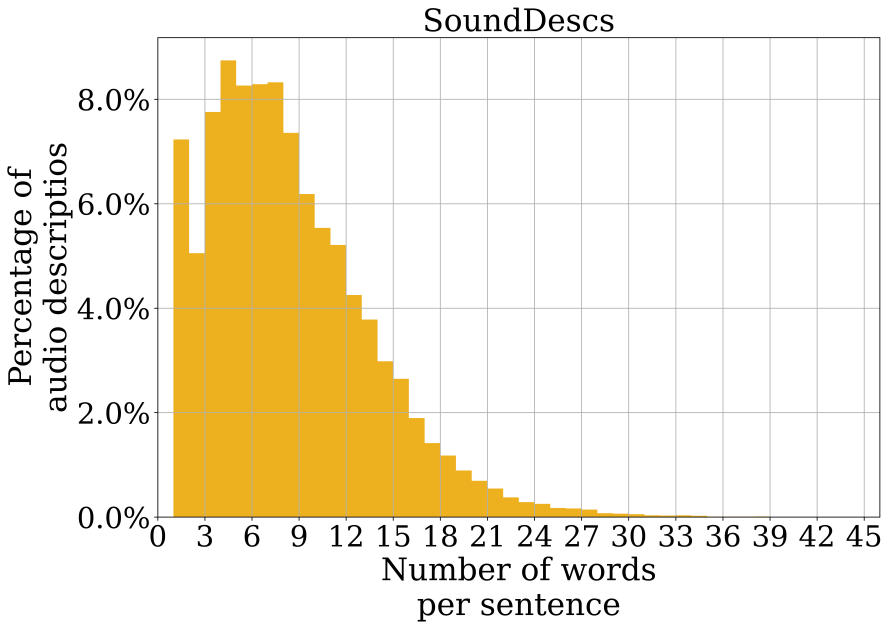


Figure 3.8: Sentence length distribution for SoundDescs.

To provide further insights into the text descriptions in the SoundDescs dataset, Figures 3.9 and 3.10 show the distribution of unique nouns and verbs per description (0 unique nouns/verbs indicates that none were present). SoundDescs contains noticeably more descriptions without any nouns and/or any verbs than AudioCaps and Clotho. There are many Nature instances in SoundDescs which contain names of species of birds that are labelled as Proper Nouns instead of Nouns. Descriptions shown to not contain any verbs are either wrongly tagged, or they do not describe actions, e.g. *Fountains: Rome - Sound of fountains, with street atmosphere.*

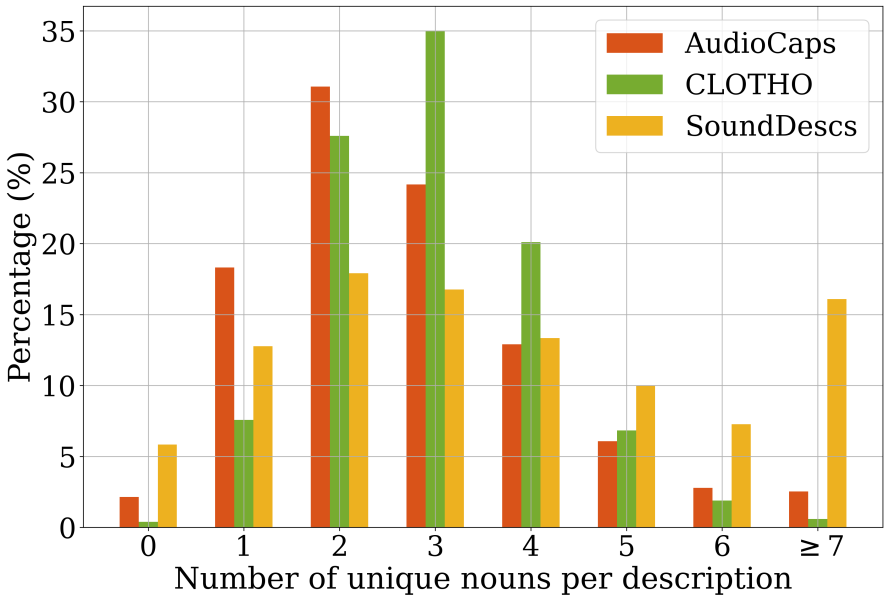


Figure 3.9: Distribution of unique nouns for descriptions in the SoundDescs, AudioCaps, and Clotho datasets.

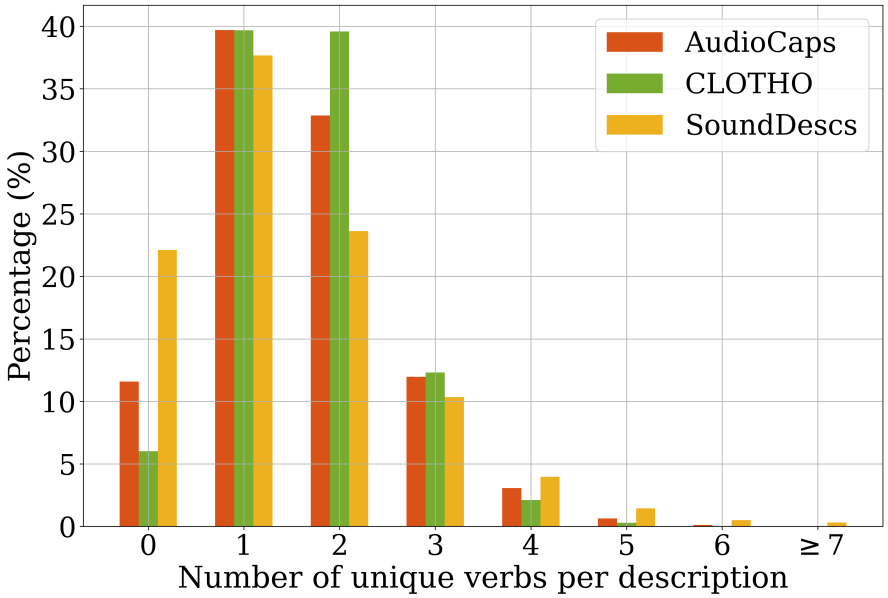


Figure 3.10: Distribution of unique verbs for descriptions in the SoundDescs, AudioCaps, and Clotho datasets.

Chapter 4

Dissecting Temporal Understanding in Text-to-Audio Retrieval

This paper has been accepted for publication at ACM Multimedia, 2024.

Dissecting Temporal Understanding in Text-to-Audio Retrieval

Andreea-Maria Oncescu[1] João F. Henriques[1] A. Sophia Koepke[2]
[1]University of Oxford [2]University of Tübingen

Abstract

Recent advancements in machine learning have fueled research on multimodal interactions, particularly in text-to-video and text-to-audio retrieval tasks. These tasks require models to develop an understanding of diverse content such as objects, sounds, and characters. Successful models must also grasp the temporal sequence and spatial arrangement of these elements. Specifically, for text-to-audio retrieval, the challenge of temporal ordering has only just started receiving attention. We dissect the temporal understanding capabilities of recent text-audio models on the AudioCaps dataset. Additionally, we introduce a synthetic audio dataset that provides a controlled setting for evaluating the temporal understanding of recent models. Lastly, we investigate a new loss function to encourage text-audio models to focus on the temporal ordering of events.

Index Terms— text-to-audio retrieval, temporal understanding

4.1 Introduction

The continued improvement of model capacities and the increase in data available have led to impressive results on multimodal tasks, such as text-image understanding [Radford et al. 2021] and text-audio understanding [Elizalde et al. 2023]. In the domain of text-audio understanding, tasks include text-to-audio retrieval [Oncescu et al. 2021b; Koepke et al. 2022; Yeh et al. 2023; Mei et al. 2023; Y. Wu

et al. 2023], audio captioning [Drossos et al. 2020; Mei et al. 2021; Deshmukh et al. 2024] and recently, text-to-audio generation [Kreuk et al. 2023; Yang et al. 2023; R. Huang et al. 2023]. Understanding details, such as temporal ordering of events, is important if we want our systems to give the best search results or generate reliable content for a text query. Recently, [H.-H. Wu et al. 2023] showed that text-audio models do not use temporal cues available in text-audio datasets.

In this work, we build on [H.-H. Wu et al. 2023] and examine limitations of current state-of-the-art text-audio models, particularly in their utilization of temporal information. Different from [H.-H. Wu et al. 2023] who use a Convolutional Neural Networks (CNNs) based audio encoder, our analysis uses the recent transformer-based audio encoder HTS-AT [K. Chen et al. 2022] that serves as a component of state-of-the-art text-to-audio retrieval models [Y. Wu et al. 2023; Mei et al. 2023]. We assess whether using an attention mechanism-based audio encoder, instead of a CNN-based one, improves the temporal understanding of text-audio models. Additionally, we investigate in detail the experimental designs and datasets used to determine if poor temporal understanding in current state-of-the-art models is caused by the training data or by the model architecture.

To determine whether commonly used text-audio datasets, such as AudioCaps [C. D. Kim et al. 2019], are suitable for training and evaluating current models' ability to comprehend time, we examine the relative distribution of audio descriptions that contain temporal cues, plotting their frequency in relation to the total number of descriptions. Our analysis shows that the AudioCaps dataset suffers from biases caused by the way humans describe events. That is, we tend to describe events in the order they appear. When first hearing the sounds of a dog barking and then the sound of a human speaking, we describe this as 'A dog barking followed by a human speaking' rather than 'A dog barking before a human speaks'. To try to address the lack of *some* temporal examples, in [H.-H. Wu et al. 2023], the authors generate new text-audio pairs starting from their existent text-audio data. They concatenate the audio files in a specific order, and then generate a description that reflects that e.g. if the generated sound is $Sound_1 + Sound_2$, the description is '<Original description of $Sound_1$> before <Original description of $Sound_1$>'. They thus further increase the training size of the data by 40%. In comparison to them, we rephrase existing text descriptions such that to preserve the content

but use a more uniform distribution of textual temporal cues. We investigate the impact of employing a more uniform set of training examples on the performance outcomes of models, comparing text-to-audio retrieval results on the original test data with those on rephrased test data.

Furthermore, we present an empirical evaluation of the correctness and completeness of AudioCaps descriptions. We leverage Large Language Models (LLMs) as oracles[1]. More specifically, we provide an LLM with the original descriptions and with the subset of AudioCaps for which we have temporally localised sounds (provided by [X. Xu et al. 2021b]). We ask the LLM to classify the sentences into correct – if the description contains all the sounds and the correct ordering, incomplete – if the description is missing sounds or is missing temporal context, and incorrect – if the description contradicts the provided grounded sounds. We observe that about 23% of the descriptions are incomplete or incorrect. This can contribute to models trained on AudioCaps not understanding temporal ordering.

To gain further insights into the temporal understanding capabilities, we propose a synthetic dataset that provides a controlled setting for analysing text-audio models. This dataset contains only 10 second long audios, keeping in line with the general setting that current models have been trained on. We show that the considered model struggles to use temporal cues in the synthetic dataset, too, confirming the findings from [H.-H. Wu et al. 2023] in a controlled setting. This is useful as it allows us to decouple the bad temporal performance of the model from the data not being suited for the task. Lastly, we propose a simple text-based contrastive loss function and show that it results in the model paying more attention to the temporal ordering of events. This also gives improvements in the overall retrieval results on the synthetic dataset.

In summary, we make the following contributions: (i) We show why existent text-to-audio retrieval datasets are not good indicators of a text-audio model's ability to understand temporal ordering, (ii) We propose a more uniform version of AudioCaps that is better suited for the temporal understanding task. We provide benchmarks and an analysis of the behaviour of current models on the original and more uniform versions of this dataset. This version keeps the audios intact

[1] An oracle in a computational context is a theoretical construct that provides perfect answers or decisions.

and only requires changing the text descriptions. (iii) We propose a synthetic dataset and use it to evaluate the model's understanding of time, (iv) We investigate an additional loss term to encourage the model to focus on text-based temporal cues.

4.2 Related work

Text-to-audio retrieval. Text-to-audio retrieval involves matching a textual query with its most relevant audio file. This task of searching through audio databases can be approached in multiple ways. One simple way is to match the text query with the title or the metadata of the audio file, provided it exists. However, for unlabelled databases, the aim is to find an audio file that has the content specified by the user through the given text query. This is called semantic search. For many years, text-audio semantic retrieval has used audio class labels made of individual or few words as text queries [Wold et al. 1996; Slaney 2002b; Ghias et al. 1995; Elizalde et al. 2019]. More recently, [Oncescu et al. 2021b; Koepke et al. 2022] proposed new benchmarks where the text query is a free-form text description rather than a pre-defined class label, allowing for more control over the retrieved audio content. Collecting new text-audio pairs for training and using state-of-the-art transformer-based audio encoders has proven beneficial on the text-to-audio retrieval benchmarks [Y. Wu et al. 2023; Mei et al. 2023; Yeh et al. 2023]. As the annotation of audio files with descriptions is time consuming, some of the text-audio pairs collected by [Y. Wu et al. 2023] and [Mei et al. 2023] contain short audio labels instead of descriptions. To overcome this, [Y. Wu et al. 2023] employed the T5 [Raffel et al. 2020] model to generate proper descriptions starting from audio labels, whilst [Mei et al. 2023] used ChatGPT [OpenAI n.d.]. [Mei et al. 2023] also used ChatGPT to clean audio descriptions from datasets such as BBC Sound Effects[2] by removing visual-based content. Another line of works considered metric learning objectives for text-to-audio retrieval [Mei et al. 2022; Xin et al. 2023]. Other concurrent research pushed the text-to-audio retrieval results even further by training models with additional modalities, such as video and speech [S. Chen et al. 2024; Y. Wang et al. 2024]. Recently, [Oncescu et al. 2024] introduced new text-to-audio retrieval benchmarks on egocentric video data.

[2]https://sound-effects.bbcrewind.co.uk/

Text-audio grounding. [X. Xu et al. 2021b] proposes a new set of data annotations for a subset of the AudioCaps dataset [C. D. Kim et al. 2019], with the aim of grounding each sound to a time interval. For this, annotators labelled the start and end times of all relevant sounds in each audio clip. [X. Xu et al. 2024a; X. Xu et al. 2023] investigated the task of weakly supervised text-to-audio grounding. The audio grounded dataset has also been used for learning to align sounds and text in an unsupervised manner [H. Xie et al. 2022]. [Bhosale et al. 2023] used the grounded sounds to propose new metrics for audio captioning. In this work, we use this subset to provide an empirical evaluation of the quality of existing AudioCaps captions. More specifically, we provide these grounded sounds and the corresponding AudioCaps descriptions to an LLM and ask it to evaluate if the descriptions are correct and complete.

Text-to-audio retrieval with synthetic data. A concurrent work [X. Xu et al. 2024b] claims that commonly used text-audio datasets only contain simple audio descriptions that are not always complete. In particular, they lack temporal cues, the number of times a sound can be heard, or details about sounds overlapping. [X. Xu et al. 2024b] propose a synthetic dataset for audio captioning, by merging 'atomic' sounds in a controlled way. We also generate a synthetic dataset and use it to analyse the temporal understanding capabilities of text-audio models.

Temporal understanding of text-audio models. [H.-H. Wu et al. 2023] show that text-audio models do not pay attention to temporal cues in text queries, such as 'followed by', or 'after'. One example of an experiment for checking if models understand time, is replacing temporal cues with words that represent a wrong ordering, e.g. by replacing 'then' with 'as'. Then, the performance of the model on the 'wrong' descriptions is evaluated. [H.-H. Wu et al. 2023] find that the model performs as well on these descriptions as when the temporal ordering in the text queries is correct. In their study, [H.-H. Wu et al. 2023] utilize Convolutional Neural Networks (CNNs) for audio processing. They identify a critical limitation of CNN-based models: the practice of applying temporal pooling across all embeddings can result in the loss of temporal information. To try mitigating this issue, they suggest augmenting the CNN architecture with several transformer layers to preserve temporal dynamics. In contrast, contemporary models built on transformers, inherently incorporate mechanisms to handle temporal data more effectively.

Different to [H.-H. Wu et al. 2023], we investigate the temporal understanding of a current transformer-based state-of-the-art audio-text retrieval model. In particular, we analyse if a transformer-based model also ignores temporal cues. We dive deep into the analysis of the descriptions in text-audio datasets in the context of temporal understanding. Additionally, we propose a synthetic text-to-audio retrieval dataset and perform temporal understanding experiments on it. Lastly, the approach proposed by [H.-H. Wu et al. 2023] for helping models better understand time does not improve the overall performance on different downstream retrieval benchmarks. In this work, we investigate a different way to guide the model to focus on temporal cues.

4.3 Temporal understanding in text-to-audio retrieval

4.3.1 AudioCaps dataset

The AudioCaps [C. D. Kim et al. 2019] dataset contains paired audio clips and text descriptions. The training set consists of one text description for each audio file. The validation and test sets contain five descriptions for each audio file. In this setting, which is employed by all benchmarks utilising AudioCaps, if any of the five text descriptions matches with the audio clip, this corresponds to 100% retrieval accuracy.

In this section, we take a closer look at the AudioCaps dataset, which is employed in all related text-to-audio retrieval works for training and evaluation. We want to better understand the temporal characteristics of this dataset to gauge if the data available for training text-to-audio retrieval models is a part of the problem of models not understanding temporal cues [H.-H. Wu et al. 2023].

First, we analyse the distribution of temporal conjunctions and prepositions in the audio captions in the AudioCaps dataset in Fig. 4.1. We observe that most conjunctions and temporal prepositions suggest future events, i.e. ‘Followed by’, ‘Then’. This is closely followed by the joint occurrence of audio events, i.e. ‘As’ and ‘While’. However, almost no examples contain the temporal prepositions ‘Before’ or ‘Preceded by’ which is reasonable as humans would not naturally describe

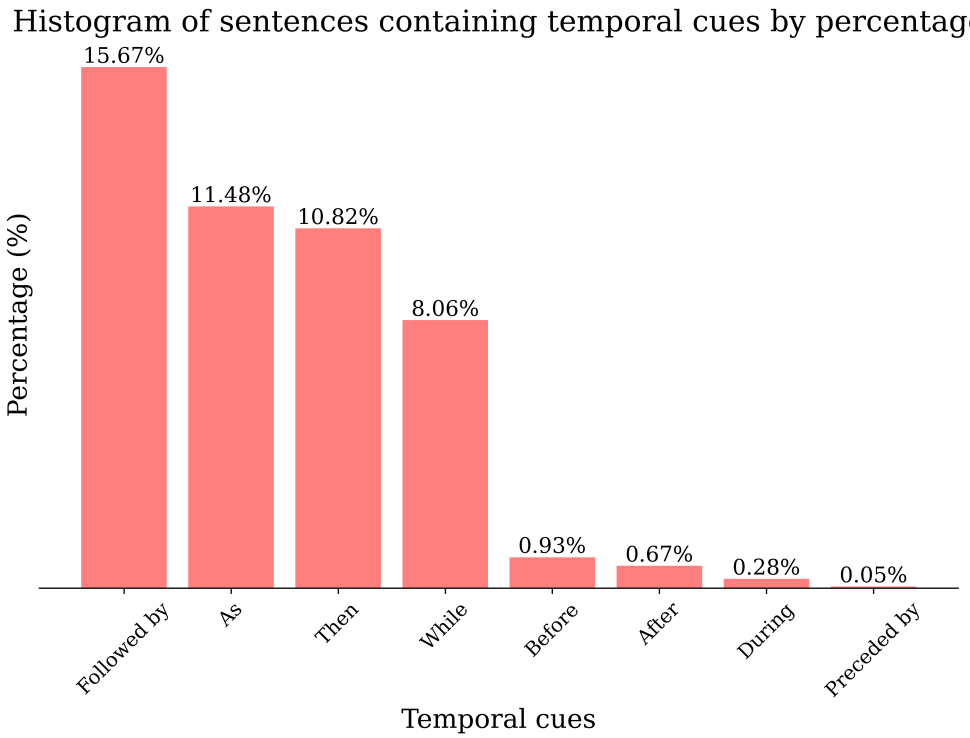


Figure 4.1: Distribution of temporal conjunctions and prepositions in the full AudioCaps dataset.

events in that order. A similar analysis is performed by [H.-H. Wu et al. 2023], where they provide percentage distribution for 'Followed by', 'Then', 'Before' and 'After'. In [H.-H. Wu et al. 2023] this distribution describes their training and test data, which are formed form a combination of multiple datasets amongst which AudioCaps [C. D. Kim et al. 2019] and Clotho [Drossos et al. 2020]. Here, we consider all words in the AudioCaps descriptions that represent temporal ordering. Given the distribution in Fig. 4.1, expecting a model trained on this data to understand the meaning of reverse temporal prepositions is unreasonable. At the same time, the test data also suffers from the same problem, therefore, using AudioCaps benchmarks for deciding if models understand temporal ordering is not optimal either.

Next, we *empirically* evaluate the correctness and completeness of AudioCaps descriptions by using the grounded sound time intervals provided by [X. Xu et al. 2021b]. Through manual inspection, we notice that many AudioCaps descriptions are composed of multiple sounds. For instance, a 10 second audio file with a bird singing from second 0 to second 6 and a dog barking from second 4 until second 10 can be described as 'Bird singing and/as dog barks'. Alternatively, this could be described as 'Bird singing followed by dog barking'. Both descriptions are correct,

however, a more complete version of these descriptions would be, for example, 'Bird singing, soon joined by a dog barking. Their sounds overlap briefly before the bird stops, while the dog continues barking.'. If the description is not complete, however, how could a model learn the difference between 'as' and 'followed by' when they describe the same audio clip?

To *empirically* evaluate the completeness and quality of the descriptions in AudioCaps with grounded sound sources, we use an LLM, specifically GPT-4 [OpenAI 2023]. We provide the LLM with the AudioCaps description, the grounded sources and their time intervals. We use one-shot prompting to give the model an example, such that it better understands the task. We then ask the model to provide an evaluation of 'correct', 'incomplete' and 'wrong' for the AudioCaps description based on the sound sources information. Details for our prompt are shown in Tab. 4.1. We process the outputs of the LLM, yielding proportions of correct, incomplete, and wrong descriptions in the subset of AudioCaps presented in Tab. 4.2. On average, 23% of the descriptions are incomplete or wrong. This percentage increases for descriptions containing *future* and *past* temporal cues. The use of *future* and *past* refers to the fact that if 'Sound 1' and 'Sound 2' are connected by a *future* temporal cue, then that means that 'Sound 1' comes first and is followed by 'Sound 2'. If a *past* cue is used, then 'Sound 1' comes after 'Sound 2', e.g. 'Bird sings after dog barks'. *Future* cues include 'Followed by', 'Before' and 'Then', e.g. 'Bird sings before dog barks'. For *past*, we consider 'Preceded by' and 'After'. Based on the significant proportion of incomplete or wrong descriptions, and the distribution of temporal textual cues, we conclude that AudioCaps is not well-suited for analysing if text-audio models understand temporal ordering.

4.3.2 Model performance on AudioCaps

In this section, we investigate the performance of a state-of-the-art model for text-to-audio retrieval on AudioCaps in detail.

Evaluation metrics.

Throughout all experiments, we use the standard evaluation metrics for retrieval: recall at rank k (R@k). This measures the percentage of targets retrieved within the top k ranked results. Higher numbers are better. We report results for text-

Table 4.1: Methodology for evaluating the quality of the grounded subset of AudioCaps using an LLM. The process includes setting the scene, one-shot prompting with an example, followed by the generation of new examples.

Prompts
Given descriptions of audio files and detailed temporal information about specific sounds within these files, where a sound may be present during multiple, distinct time intervals, your task is to evaluate the accuracy of each description with a primary focus on the timing and sequence of these sounds. Each audio file is 10 seconds long. For every description, assess its accuracy specifically in terms of how well it captures the chronological order and exact timing of sounds. Classify your evaluation into one of three categories: 'Correct', 'Incomplete', or 'Wrong'. If necessary, provide a corrected description that not only fixes inaccuracies related to timing but also maintains the original writing style of the description. Your analysis should critically examine the temporal details provided, ensuring your assessment is primarily guided by the accuracy of these temporal sequences. Keep in mind the following: Pay attention to whether the description matches the start and end times of sounds accurately. Consider if the sequence of described sounds follows the actual sequence in the audio file. Evaluate if the description misses any sounds within the specified time frames or includes sounds that do not occur within these times. Use similar vocabulary as the original audio description. Example: Input: Original audio description: A power tool motor running then revving Localized components and their start and end times: revving: 2.154, 10.02; a power tool motor running: 0.0, 10.02; Output: Evaluation: Incomplete Corrected description: A power tool motor running throughout, with revving starting early on and continuing alongside the motor's running sound until the end.

Table 4.2: Proportion of correct, incomplete and wrongly captioned AudioCaps data. First row contains the total numbers of grounded descriptions. The other rows show proportions for specific temporal cues.

Preposition	Correct	Incomplete	Wrong
Total (#)	3835	636	503
As (%)	75.3	13.9	10.8
Followed by (%)	60.6	15.3	24.1
Then (%)	62.1	15.9	22.0
While (%)	72.0	15.4	12.6
Before (%)	58.8	9.8	31.4
After (%)	54.5	6.1	39.4
Proceeded by (%)	50.0	0.0	50.0
During (%)	53.3	13.34	33.3
And (%)	75.6	13.5	10.9

to-audio (T $\rightarrow$ A) and audio-to-text retrieval (A $\rightarrow$ T). We report the mean of three runs that use different random seeds.

Model

We employ the state-of-the-art text-audio model by [Mei et al. 2023], utilising an HTS-AT audio encoder [K. Chen et al. 2022], and a pre-trained BERT encoder for text. After encoding audio and text, an MLP projects the embeddings into the same space. We use the model variant trained on the WavCaps [Mei et al. 2023]. In our experiments, we finetune the model for 40 epochs and use the same setup as [Mei et al. 2023]. The best model is selected based on the highest average validation retrieval accuracy R@1.

Loss function

We use the same loss as [Mei et al. 2023] - a normalised temperature scaled bidirectional cross-entropy loss (NT-Xent) [T. Chen et al. 2020]. We call this loss $\mathcal{L}_{at}$ with

$$s_{ij} = \frac{f(a_i) \cdot g(t_j)}{\|f(a_i)\|_2 \|g(t_j)\|_2}, \tag{4.1}$$

$$\mathcal{L}_{at} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \left(\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s_{ij}/\tau)} \right) + \log \left(\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s_{ji}/\tau)} \right) \right]. \tag{4.2}$$

Here $f(\cdot)$ is the audio encoder and $g(\cdot)$ the text encoder. s_{ij} is the cosine similarity, a_i is an audio input, t_j is a text input, B is the batch size, and τ is a temperature parameter. More details in [Mei et al. 2023].

Data used

For our analysis, we construct a more *uniform* version of the AudioCaps dataset with descriptions having a more balanced distribution of temporal conjunctions and prepositions. In particular, we rephrase AudioCaps descriptions to preserve the original meaning while varying the use of temporal conjunctions and prepositions. We investigate if this helps with temporal understanding. Specifically, we re-write the descriptions from the original AudioCaps dataset to generate the $AudioCaps^{uni}$ dataset with corresponding $Train^{uni}$, Val^{uni} and $Test^{uni}$ subsets. There are two approaches to re-writing the descriptions. One is to replace the temporal cues with something that has the same meaning e.g. 'Bird singing fol-

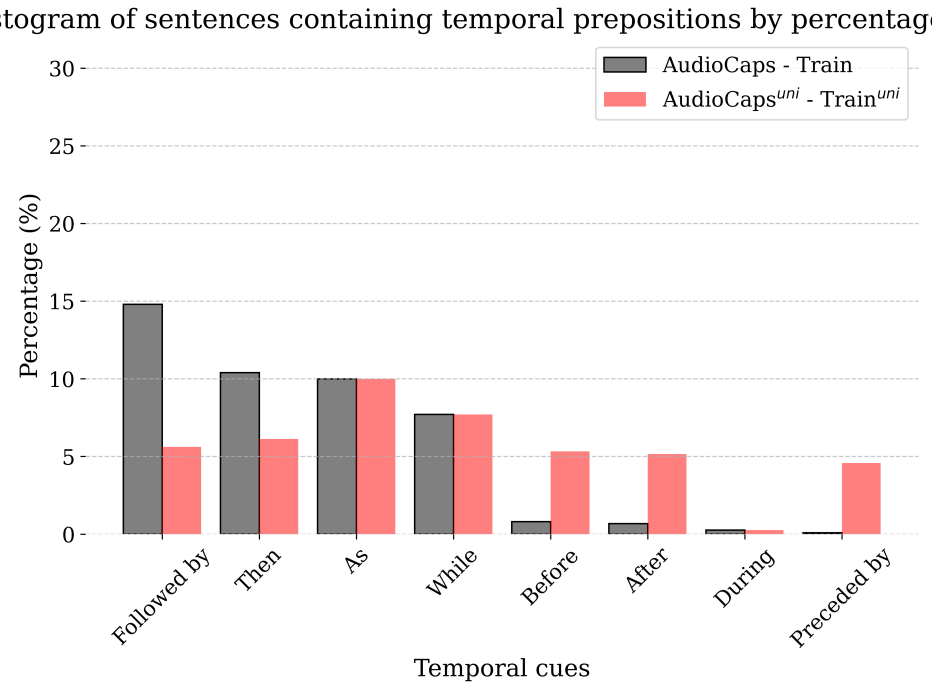


Figure 4.2: Distribution of temporal conjunctions and prepositions in AudioCaps training data. We compare the proportion of temporal textual cues in the original training dataset (Train) and the more uniform dataset (Train^{uni}).

lowed by dog barking' is equivalent to 'Bird singing before dog barking'. The second approach is to re-order the text location of events and also change the temporal cue e.g. 'Bird singing followed by dog barking' becomes 'Dog barking after bird singing'. We present the distribution of temporal cues in the original AudioCaps dataset and its uniform variant in Fig. 4.2 and Fig. 4.3. An analysis of the validation split can be found in the supplementary material.

In addition to the more uniform AudioCaps version, we create a test subset where at least one of the 5 text descriptions contains *future* and *past* temporal cues (as defined in Sec. 4.3.1). We do this to evaluate the performance of the model on sentences that actually contain temporal cues of interest. We call this subset *TempTest*.

Experiments

We consider three main experiments. First, we conduct the standard evaluation for text-to-audio retrieval on the AudioCaps test set. The performance on the standard test set serves as a point of reference for the training and evaluation on different variants of the AudioCaps data.

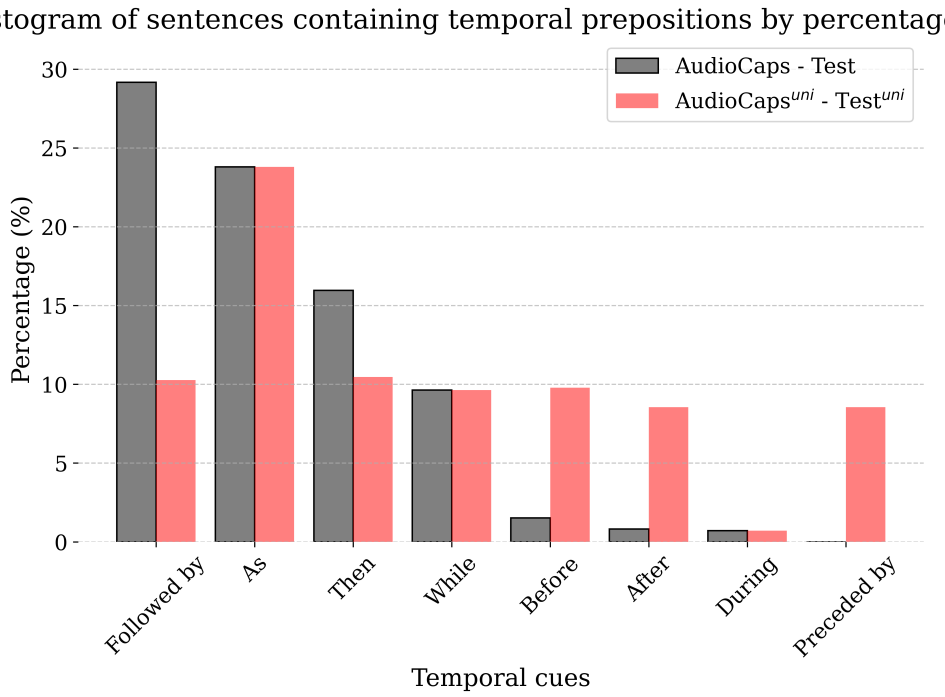


Figure 4.3: Distribution of temporal conjunctions and prepositions in the AudioCaps test dataset.

The second experiment involves reversing the ordering of sounds in the text queries of the test set. We call this $Test^{rev}$. The purpose of this experiment is to see what happens if the temporal text descriptions keep the same temporal preposition or conjunction but the sound sources are reversed, resulting in a wrongly ordered description, e.g. *Birds singing before dog barks* becomes *Dog barks before birds singing*. If the model understands the temporal ordering of events, the performance of the model should drop for the 'wrongly' ordered events as compared to the original test set performance.

For the third experiment, we replace the temporal cue in a description with its opposite, thus changing the order of events without changing their position, e.g. *Birds singing before dog barks* becomes *Birds singing after dog barks*. We refer to this as $Test^{rep}$. More concretely, we do the following replacements: ('followed by'→'preceded by'), ('after'→'before'), ('before'→'after'), ('then'→'before') and ('preceded by'→'followed by').

If the model does not understand temporal cues, we expect it to perform similarly well on *Test*, $Test^{rev}$ and $Test^{rep}$. Conversely, if it understands temporal ordering, *Test* performance should be considerably higher than $Test^{rev}$ and $Test^{rep}$. We would also expect $Test^{rev}$ and $Test^{rep}$ to be similar, as the meaning of the sentence

is the same but opposite of the *Test* sentences meaning. We additionally consider *rev* and *rep* subsets of the temporal subset *TempTest*.

In Tab. 4.3, we take the checkpoint provided by [Mei et al. 2023] which was trained on WavCaps [Mei et al. 2023] and finetune it on the original AudioCaps training dataset. We notice that on the reversed $TempTest^{rev}$ set the model performs worse, indicating that the model understands temporal ordering. However, on $TempTest^{rep}$ which contains the replacement of temporal cues, the model performs similarly to on *TempTest*. This is interesting, as the temporal ordering of both $TempTest^{rev}$ and $TempTest^{rep}$ is reversed and wrong as compared to *TempTest*. The only difference is that for the former, the actual positional text locations of the sounds are swapped, whilst for the latter the meaning is reversed by changing the temporal connector. This leads us to believe that at best, the model learns text location-based ordering rather than the ordering given by the text connector.

We then run the same experiments on the test sets of AudioCaps^{uni}. We notice that the model is unable to identify the text-based order of sound events, with results on all 'correct' (TempTest) and 'wrong' (TempTest^{rev} and TempTest^{rep}) splits being almost the same.

Next, we investigate if the model does not understand temporal ordering due to a lack of variety in the training examples. We take the same pre-trained checkpoint as before [Mei et al. 2023], and finetune it on the $AudioCaps^{uni}$ $Train^{uni}$ set. We notice that the overall performance on the $AudioCaps^{uni}$ test sets and the corresponding temporal subsets is higher when finetuning on a more uniform distribution of temporal cues (Tab. 4.4) than when finetuning on the original training data (Tab. 4.3). Thus, the lack of understanding temporal ordering is in part due to the training data not containing examples of *past* temporal cues. We also notice some signs of better temporal understanding, with a slightly bigger drop in performance on the $TempTest^{rev}$ and $TempTest^{rep}$ sets relative to *TempTest*.

Table 4.3: Performance evaluation of model pre-trained on WavCaps dataset and **fine-tuned on origAudioCaps (origTrain)** with the original audio-text loss function. Table presents retrieval accuracies R@1 on AudioCaps (Test) and AudioCaps^{uni} (Test^{uni}). Reverting order of events (generally) does not change performance.

Eval Dataset	Subset	T→A	A→T
		R@1	R@1
AudioCaps	Test	43.71	56.57
	TempTest	50.51	63.74
	TempTest^{rev}	43.90	57.71
	TempTest^{rep}	49.55	62.67
AudioCaps^{uni}	Test	41.54	53.84
	TempTest	48.61	61.37
	TempTest^{rev}	47.37	62.10
	TempTest^{rep}	47.60	60.89

Table 4.4: Performance evaluation of model pre-trained on WavCaps dataset and fine-tuned on AudioCaps^{uni} (Train^{uni}). Improved results on Test^{uni}. Slightly bigger drop in *rev* and *rep* wrt TempTest.

Eval Dataset	Subset	Loss	T→A	A→T
			R@1	R@1
AudioCaps^{uni}	Test	$\mathcal{L}_{ta}$	**43.67**	**53.88**
	TempTest	$\mathcal{L}_{ta}$	**50.67**	61.31
	TempTest^{rev}	$\mathcal{L}_{ta}$	46.82	59.31
	TempTest^{rep}	$\mathcal{L}_{ta}$	47.45	59.06

4.4 Temporal understanding in controlled setting

We analyse the text-audio model's temporal understanding in a controlled setting where we can guarantee correct alignment of text-audio pairs.

4.4.1 Data generation

We use the ESC-50 [Piczak 2015] environmental sound classification dataset to generate a synthetic dataset for text-to-audio retrieval with a focus on temporal understanding capabilities. ESC-50 is a dataset of 2000 audio samples from 50 classes. As this dataset is clean and contains clear 'atomic' sounds (i.e. 5 second audios containing only one sound), we use it for synthetic data generation.

We first ask an LLM to take the sound labels from ESC-50 and generate textual descriptions in the style of AudioCaps (e.g. 'dog'→ 'dog barking'). To generate the text-audio pairs, we take two sounds and their LLM-generated labels and concatenate them based on the temporal order we decide on. We call this dataset *SynCaps*.

To avoid any confusion, we only use *future* and *past* temporal cues. This is because synchronous temporal cues such as 'as' or 'during' can represent many things, especially in a noisily labelled dataset. They can be used for sounds that completely overlap, or for partial overlaps of sounds, ignoring the actual order in which the sounds appear. The test set contains unique sound components that are not used in the training and validation sets. This leads to 485 examples of 10 second long audio clips. For training and validation, we allow the same 5 seconds sound component to appear on average 5 times and apply 5 types of augmentation, to reduce overfitting on the training and validation sets. We use augmentations, such as time shifting, volume adjustment, pitch shift, time stretch, and added noise. We also allow an overlap between the files of up to 1 second. This results in a total of 4400 training files and 485 validation files.

Table 4.5: Performance evaluation of a model pre-trained on WavCaps (PT) and fine-tuned on SynCaps. Model uses $\mathcal{L}_{ta}$ (audio-text alignment) loss function.

Subset	T→A	A→T
	R@1	R@1
Test	67.70	65.50
Test^{rev}	66.67	64.95
Test^{rep}	67.08	63.64

4.4.2 SynCaps experiments

We analyse the temporal understanding of the text-audio model in the more controlled setting of the SynCaps dataset. For this, we take the same pre-trained model from [Mei et al. 2023] and finetune it on SynCaps.

We see that evaluating on the 'reversed'(*rev*) and 'replaced'(*rep*) datasets gives almost the same results as using the correct (original) test data (Tab. 4.5). This shows that the model indeed does not understand temporal cues even on a simple dataset.

4.4.3 Proposing new loss

We propose a loss function $\mathcal{L}_{tt}$ that enhances the understanding of temporal information. It is formulated as a text-to-text contrastive loss, which relies on pairs of positive examples (have the same temporal significance as the original sentence) and negative text examples (have the opposite temporal meaning). Concretely, given the original description *Bird sings followed by dog barks*, one positive example is *Bird sings before dog barks* and one negative example would be *Bird sings after dog barks*.

We provide the model with two positive text examples and two negative text examples for each text description containing the previously defined *future* and *past* temporal textual cues. Positive and negative text examples can be generated once, before training the model. We searched for the temporal cues we are interested in and automatically generated multiple positives and negatives by changing the temporal cues and/or the ordering of the sounds.

Table 4.6: Text-audio retrieval with model pre-trained on WavCaps and fine-tuned on SynCaps using our text-text contrastive loss $\mathcal{L}_{tt}$.

Subset	Loss	T→A	A→T
		R@1	R@1
Test	$\mathcal{L}_{ta} + \lambda\mathcal{L}_{tt}$	**69.83**	**71.13**
Test^{rev}	$\mathcal{L}_{ta} + \lambda\mathcal{L}_{tt}$	40.41	43.43
Test^{rep}	$\mathcal{L}_{ta} + \lambda\mathcal{L}_{tt}$	44.95	47.70

The contrastive loss for each query and a margin α is:

$$\mathcal{L}_{tt} = \frac{1}{2N}\sum_{n=1}^{N}\sum_{k=1}^{2}\max(0, \alpha - \text{pos\_sim}_{nk} + \text{neg\_sim}_{nk}), \tag{4.3}$$

where pos_sim_{nk} is the similarity between the n-th query and its k-th positive example, neg_sim_{nk} is the similarity between the n-th query and its k-th negative example. Our full loss then becomes:

$$\mathcal{L} = \mathcal{L}_{ta} + \lambda\mathcal{L}_{tt}. \tag{4.4}$$

In our experiments that use the text-text loss, we set $\lambda = 10$, $\alpha = 0.2$.

We now evaluate the same model pre-trained on WavCaps and finetuned on SynCaps but employing our additional loss. In Tab. 4.6, we observe that the model performs better on the original test set compared to Tab. 4.5, whilst at the same time showing a big drop in performance on the 'reversed' and 'replaced' data. This shows that employing a simple additional loss can help the model better understand time, at least in the synthetic controlled setting.

4.5 Conclusion

In this work, we dissected temporal understanding capabilities of current state-of-the-art text-audio model. We first analysed how well-suited AudioCaps is as a training and evaluation dataset for the temporal understanding of events. We then proposed a new synthetic dataset, concluding that indeed models fail to use the temporal cues even when the data is clean. Lastly, we propose a simple loss that results in better text-to-audio retrieval results on SynCaps, whilst also putting more emphasis on the temporal content of the audio and text data.

4.6 Supplementary Material

In this supplementary material, we provide additional information about AudioCaps [C. D. Kim et al. 2019] and present experimental results for our text-text contrastive loss on the AudioCaps dataset.

4.6.1 AudioCaps validation data

We consider the validation set of AudioCaps and plot the distribution of sentences containing temporal cues in Fig. 4.4. We notice that differently from the train and the test set, there are considerably more sentences containing the temporal cues 'As' and 'While'. It is important for the validation set to be representative of the training and test sets and at the same time to contain examples of different temporal cues. This is to select the best model that has the potential of performing well on the test set and at the same time, to understand temporal constraints.

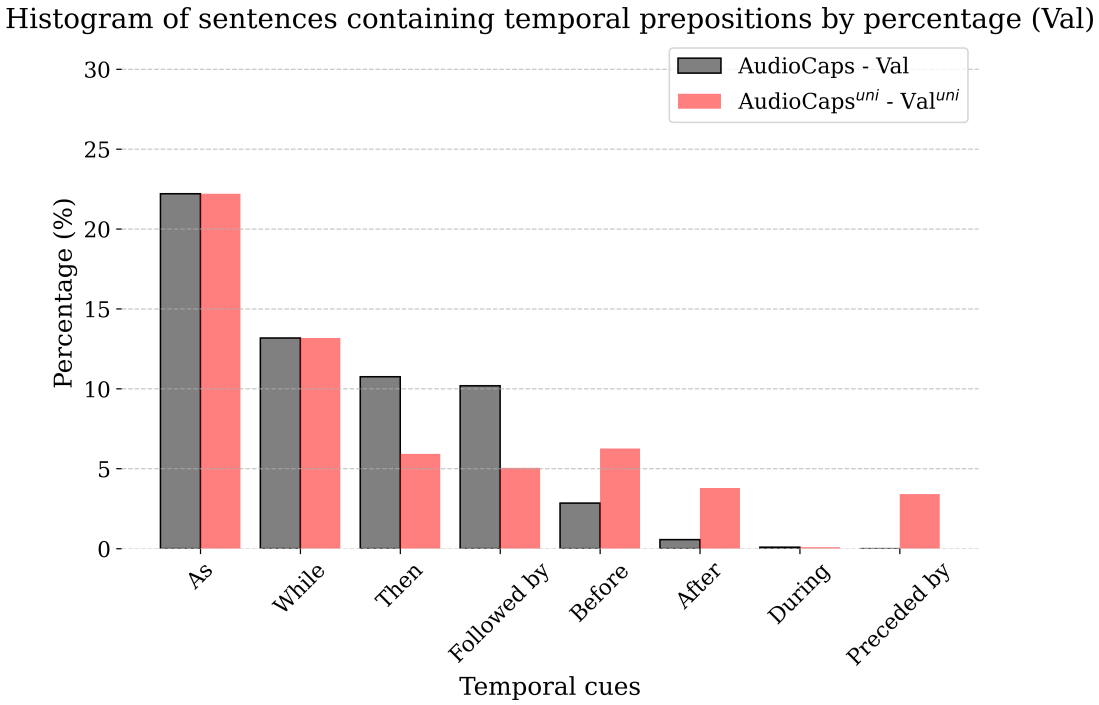


Figure 4.4: Distribution of temporal conjunctions and prepositions in the AudioCaps validation dataset.

4.6.2 Additional experiments on AudioCaps

The performance on *TempTest*, and on the modified *TempTest*^{rev} and *TempTest*^{rep} sets obtained from AudioCaps is almost the same in Tab. 4. This means that the

Table 4.7: Performance evaluation of model pre-trained on WavCaps dataset and fine-tuned on AudioCaps^{uni} (Train^{uni}). Improved results on Test^{uni}.

Eval Dataset	Subset	Loss	T→A	A→T
			R@1	R@1
AudioCaps^{uni}	Test	$\mathcal{L}_{ta} + \lambda\mathcal{L}_{tt}$	42.48	53.43
	TempTest	$\mathcal{L}_{ta} + \lambda\mathcal{L}_{tt}$	49.33	**61.86**
	TempTest^{rev}	$\mathcal{L}_{ta} + \lambda\mathcal{L}_{tt}$	39.06	50.85
	TempTest^{rep}	$\mathcal{L}_{ta} + \lambda\mathcal{L}_{tt}$	38.78	51.40

model [Mei et al. 2023] does not differentiate much between the right order of the text sound events or them being reversed. Therefore, we investigate if adding our text-text contrastive loss $\mathcal{L}_{tt}$ encourages the model to pay more attention to temporal ordering on AudioCaps, as we found this to be the case on SynCaps.

When using our additional text-to-text loss when finetuning the model, it is more sensitive to the 'reversed' and 'replaced' subsets (Tab. 4.7). However, the overall results are similar to when the loss is not used, warranting further research into improving the training objective.

4.6.3 Why AudioCaps uniform data

We propose a more uniform version of the AudioCaps dataset that is easy to use in experiments, as it only changes the original text descriptions to have the same meaning but use a more varied set of temporal cues. If models reach a point where they trully understand temporal ordering, we should expect that, the performance on the *Test* and *Test^{uni}* will be very similar. The same should apply to *TempTest* and *TempTest^{uni}*.

Authors of [H.-H. Wu et al. 2023] perform one experiment where they take 88 text-audio examples containing 'before' and 88 text-audio examples containing 'after'. They then swap 'before' with 'after' and vice versa. They use this experiment to evaluate if the model understands the temporal ordering implied by these prepositions. The reason they use such a small test set for this experiment is that there are not enough sentence examples containing 'before' and 'after'. This constraints the statistical significance of the experiments we can run. However, more sentences containing these prepositions can easily be obtained by re-writing what we already have. This allows us to run a more reliable experiment regarding the effect of re-

placing 'before' and 'after'. It also allows for a more comprehensive experiment where we not only replace 'before' with 'after', but generate other negative 'swaps'. This is the equivalent of our *rep* experiment.

4.6.4 Employing the text-text contrastive loss in other datasets

In our experiments we have used a simple approach of generating paris of positive and negative examples for the text-text contrastive loss. In this approach we searched for some pre-defined temporal cues and automatically rearranged them to suit our purpose. However, a more general approach to obtain pairs of positive and negative examples can be to employ an LLM to generate them.

4.6.5 Verify correctness of LLM evaluations of AudioCaps descriptions

In Sec. 3.1., we have proposed an empirical evaluation of the correctness and completeness of the AudioCaps descriptions. We have done so by starting from the assumption that Large Language Models act as oracles, providing reliable evaluations. To further verify the correcntess of the LLM evaluations, one approach can be to ask the LLM to not only classify the descriptions into 'correct', 'incomplete' and 'incorrect', but also ask for the correct description given the inputs. Then, for the cases where the original AudioCaps descriptions were considered 'incomplete' or 'wrong', we can repeat the process, but this time provide the LLM generated "corrected" descriptions and the sound timestamps as inputs. If now the LLM considers the generated description 'correct', then we can assume that the original evaluation of the original AudioCaps description of 'incomplete' or 'wrong' was reliable. If not, we can repeat the process from the beginning. We repeat this until the two LLM steps agree on all examples.

Chapter 5

A sound approach: Using large language models to generate audio descriptions for egocentric text-audio retrieval

The paper has been accepted for publication at International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2024

A sound approach: Using large language models to generate audio descriptions for egocentric text-audio retrieval

Andreea-Maria Oncescu[1] João F. Henriques[1]
Andrew Zisserman[1] Samuel Albanie[2] A. Sophia Koepke[3]
[1]University of Oxford [2]University of Cambridge
[3]University of Tübingen

Abstract

Video databases from the internet are a valuable source of text-audio retrieval datasets. However, given that sound and vision streams represent different "views" of the data, treating visual descriptions as audio descriptions is far from optimal. Even if audio class labels are present, they commonly are not very detailed, making them unsuited for text-audio retrieval. To exploit relevant audio information from video-text datasets, we introduce a methodology for generating audio-centric descriptions using Large Language Models (LLMs). In this work, we consider the egocentric video setting and propose three new text-audio retrieval benchmarks based on the EpicMIR and EgoMCQ tasks, and on the EpicSounds dataset. Our approach for obtaining audio-centric descriptions gives significantly higher zero-shot performance than using the original visual-centric descriptions. Furthermore, we show that using the same prompts, we can successfully employ LLMs to improve the retrieval on EpicSounds, compared to using the original audio class labels of the dataset. Finally, we confirm that LLMs can be used to determine the difficulty of identifying the action associated with a sound.

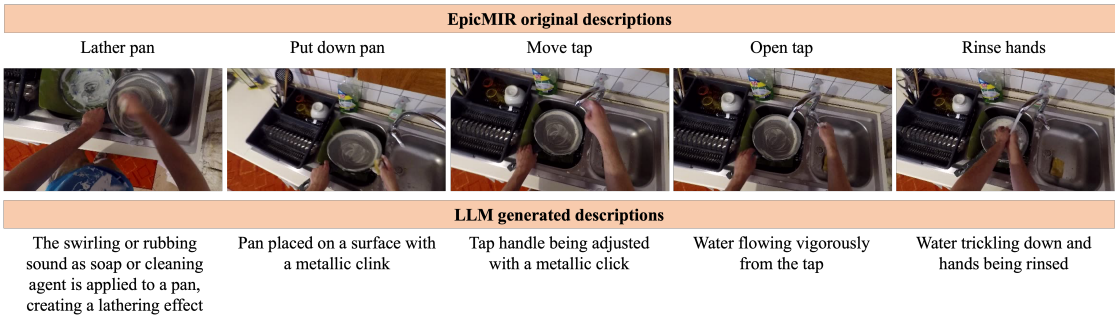


Figure 5.1: Frames from the EpicKitchens dataset together with the corresponding original visual descriptions from EpicMIR shown above and those generated with our approach (using the ChatGPT LLM [OpenAI n.d.]) below.

Index Terms— text-audio retrieval, large language models, generated audio descriptions, egocentric data

5.1 Introduction

Searching the ever-expanding supply of audio and video media hosted online has become a key technical challenge. Concurrently, LLMs have become more powerful, exhibiting early signs of commonsense reasoning and primitive world modelling [Touvron et al. 2023b]. Given their extensive text-based knowledge about the sensory world, in this work we ask whether LLMs can improve search capabilities for other modalities such as audio and video. In particular, we consider the task of egocentric audio retrieval from text queries.

Our strategy is to employ text as an intermediate medium for aligning vision and audio signals by leveraging the text-based knowledge that an LLM possesses about sight and sound. Concretely, we use LLMs to generate plausible *audio* descriptions for videos when given their *visual* descriptions. To do so, we leverage the LLM's in-context learning capability, and provide it with exemplars of the desired mapping – pairs of visual-centric and corresponding audio-centric descriptions. This few-shot shot approach is made possible by the existence of a small collection of content that has been annotated with both visual-centric and audio-centric descriptions. In this work, we source these pairs from overlapping samples between the Kinetics700-2020 [Smaira et al. 2020] and AudioCaps [C. D. Kim et al. 2019] datasets (we refer to these examples as Kinetics $\cap$ AudioCaps). Fig. 5.1 shows examples of few-shot

Acknowledgements. This work is supported by the EPSRC (VisualAI EP/T028572/1 and DTA Studentship), the Royal Academy of Engineering (RF\201819\18\163), an Isaac Newton Trust Grant and the DFG - EXC 2064/1 - project 390727645. We are grateful to Jaesung Huh, Triantafyllos Afouras and Bernie Huang for their helpful comments and suggestions.

generated audio descriptions produced with our approach.

With this "converter" in hand, we scale its application to the conversion of full text-video datasets to text-audio datasets. Specifically, we construct two text-audio datasets derived from egocentric video retrieval tasks sourced from EpicKitchens [Damen et al. 2018] and Ego4D [Grauman et al. 2022], and demonstrate the value of this data empirically. We additionally apply a similar methodology to the *audio class labels* of EpicSounds to improve retrieval results. We also demonstrate that LLMs can usefully predict when the action within a video can be reliably determined solely from its audio track, with applications for curating new text-audio datasets.

5.2 Related work

Text-audio retrieval and LLMs. Text-audio retrieval entails searching for the most appropriate audio file for a given textual query. This task was popularised by [Oncescu et al. 2021b; Koepke et al. 2022] (though related themes were studied previously [Slaney 2002b]), and has seen recent improvements through the use of transformer-based models [Y. Wu et al. 2023]. In adjacent fields, there has been a surge of efforts that harness LLMs, e.g. GPT-4 [OpenAI 2023], Vicuna [Chiang et al. 2023], and Llama [Touvron et al. 2023a; Touvron et al. 2023b], for multimodal tasks that require vision-language [D. Zhu et al. 2023; Shen et al. 2023; Kaul et al. 2023; Menon and Vondrick 2023] and audio-language understanding [Mei et al. 2023]. [Menon and Vondrick 2023] and [Kaul et al. 2023] demonstrate the benefits of using LLM-generated textual class descriptions for zero-shot image classification and open-vocabulary object detection respectively. More closely related to our approach, [Mei et al. 2023] employ ChatGPT [OpenAI n.d.] for standardizing audio descriptions across various audio-centric datasets. These refined text-audio pairs are then used to pretrain text-audio models. In contrast to prior work which focuses predominantly on audio-centric datasets (as characterized by the availability of audio descriptions or labels), we focus on leveraging datasets that generally possess only visual-centric descriptions.

Egocentric audio-visual understanding. While most video datasets are captured from a third-person perspective, recent research has shifted towards egocentric data, filmed from a first-person viewpoint. The egocentric datasets, EpicK-

itchens [Damen et al. 2018] and Ego4D [Grauman et al. 2022], have been used for various tasks, such as action recognition [Mittal et al. 2022; Planamente et al. 2022], moment localization and video retrieval [Lin et al. 2022; Zhao et al. 2023; Ashutosh et al. 2023; Pramanick et al. 2023] with a focus on the video stream. We, however, focus specifically on their audio tracks.

In a similar vein to [Kazakos et al. 2019], we analyse how to best exploit the audio information in the egocentric video setting. Also related, recent work introduced EpicSounds [Huh et al. 2023], an audio classification benchmark on the EpicKitchens [Damen et al. 2018] dataset. In contrast to these approaches, we consider the retrieval task rather than classification and employ LLMs to automatically generate additional audio descriptions for text-audio retrieval.

5.3 Datasets and approach

We first summarise existing relevant datasets in Sec. 5.3.1. Next, we detail our proposed method for employing Large Language Models (LLMs) in two key areas: firstly, to bridge the gap between visual and audio descriptions; and secondly, to create audio descriptions from audio labels. This approach is aimed at improving text-audio retrieval in egocentric datasets, as discussed in Sec. 5.3.2. We introduce our AudioEpicMIR, AudioEgoMCQ and EpicSoundsRet benchmarks in Sec. 5.3.3. Lastly, we present our method for evaluating the level of informativeness of audio samples in Sec. 5.3.4.

5.3.1 Datasets and tasks

EpicKitchens [Damen et al. 2018] & EpicMIR [Damen et al. 2022]. EpicKitchens contains 100 hours of recordings filmed from first-person perspective. The Multi-Instance Action Retrieval (EpicMIR) [Damen et al. 2022] task based on EpicKitchens consists of finding the most relevant video given a text query, and vice-versa. The test set contains 9,668 videos and corresponding visual descriptions in the form `verb/s + noun/s`. The average number of words per sentence is 2.93 with standard deviation 1.17.

Ego4D [Grauman et al. 2022] & EgoMCQ [Lin et al. 2022]. Ego4D is the largest egocentric dataset with over 3,670 video hours and accompanying narrations. The EgoMCQ [Lin et al. 2022] task based on Ego4D consists of 39,751 pairs

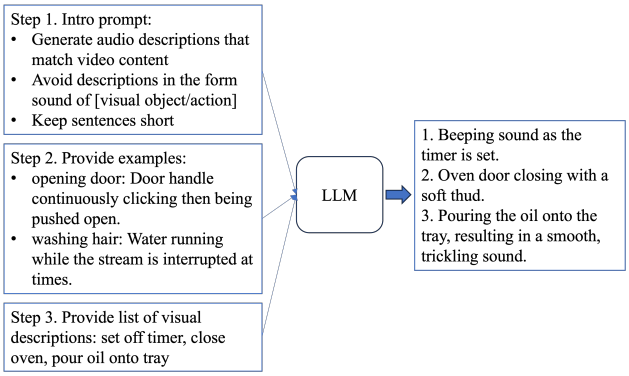


Figure 5.2: Given visual-centric descriptions, we propose to use an LLM (ChatGPT) to generate audio descriptions (*step 3*). The LLM is prompted with a task description (*step 1*) and few-shot paired examples of visual-centric and audio descriptions (*step 2*).

containing one description and 5 video clips each. It aims at finding the correct clip for a given description.

EpicSounds [Huh et al. 2023] is sourced from the EpicKitchens dataset. It contains only those audio tracks that are useful for the audio classification task as verified through manual annotations. 44 different classes are labelled for 10,276 test audio chunks.

Kinetics700-2020 [Smaira et al. 2020] contains 10s clips from YouTube and corresponding activity labels which are in the form `verb/s + noun/s`. The average word count is 2.09 with a standard deviation of 0.79.

AudioCaps [C. D. Kim et al. 2019] is curated from YouTube and contains 10s audio files and text descriptions. It serves as a common benchmark dataset for audio captioning and text-audio retrieval.

5.3.2 Audio description generation methodology

As noted in Sec. 5.1, we find that there are a number of clips in common between Kinetics700-2020 [Smaira et al. 2020] and AudioCaps [C. D. Kim et al. 2019] (283 in total). We use example correspondences to condition the LLM to generate audio descriptions in the style of AudioCaps given access to visual verb-noun descriptions. This approach is shown in Fig. 5.2, where we first give a general task description, together with heuristic constraints that were obtained by manual experimentation on a handful of samples (a form of "prompt engineering").

We combine these instructions with few-shot paired examples of visual descriptions

and audio descriptions sampled from Kinetics∩AudioCaps. In particular, we select 14 pairs to balance sufficient examples while preserving the model's focus on the task prompt (step 2). Finally, we provide the LLM with the visual descriptions or audio labels for which we want to generate new audio descriptions (step 3). We developed this approach by experimenting with samples from EpicMIR, and apply the same strategy directly to EgoMCQ and EpicSounds in Sec. 5.4.

5.3.3 Our proposed text-audio retrieval benchmarks

We apply our approach for generating audio descriptions to EpicMIR, EgoMCQ, and EpicSounds. As a result, we obtain three new benchmarks, namely AudioEpicMIR, AudioMCQ, and EpicSoundsRet. We refer to the original visual descriptions/audio class labels for all benchmark datasets as *Aud orig* and to our LLM-generated descriptions as *Aud LLM*.

AudioEpicMIR is curated from the EpicMIR task by extracting the audio tracks of the EpicMIR videos and keeping their original visual descriptions. Additionally, we generate audio-centric descriptions using ChatGPT (GPT-3.5) [OpenAI n.d.] as the LLM in our methodology, as described in Sec. 5.3.2.

AudioEgoMCQ is gathered based on the EgoMCQ task. We observe that in this dataset, some of the videos do not have an audio soundtrack at all. Therefore, to generate a text-audio dataset, we need to exclude some of the original text-video pairs. Specifically, for the **intra-video** task, we excluded all text-video pairs when the video corresponding to the text query did not contain sound. For **inter-video**, we additionally replaced clips without audio from the pairs by clips with audio. Finally, we exclude text-video pairs if any of the five videos have a silent soundtrack. This results in 23,121 text-audios pairs. The remaining original visual descriptions contain an average of 8.15 words (with a standard deviation of 3.01). We use our audio description generation approach described in Sec. 5.3.2 to obtain audio-centric descriptions. As before, we use ChatGPT (GPT-3.5) [OpenAI n.d.] as the LLM.

EpicSoundsRet differs from EpicMIR and EgoMCQ in that it originally contains *audio class labels*, not *visual descriptions*. For the retrieval task we use the audio class labels as text queries together with the corresponding audio files. We use our LLM prompts out of the box to generate *audio descriptions* for EpicSounds

starting from the *audio labels*.

5.3.4 Determining the audio relevancy using LLMs

We observe that many video clips contain audio that is too noisy or generic to inform the text-audio retrieval process. An example is 'taking clothes from basket' or 'putting clothes in basket' which both have similar associated sounds that are hard to differentiate. This prompts the question, can we identify such audio by looking only at the given visual description? To explore this, we tasked GPT-4 with splitting the original visual text descriptions into three categories according to the relevancy of the sound:

- High: audio is very informative for the visual task, e.g. 'turning on tap', 'washing dishes'.
- Moderate: audio is informative but not enough information is provided in the text description for it to be identifiable, e.g. 'putting a plate down' has a different sound based on if it is placed on a kitchen table or a sofa.
- Low: audio is not likely to be informative, e.g. 'get carrot'.

5.4 Experiments

Evaluation metrics. On AudioEpicMIR and EpicSoundsRet, we report Mean Average Precision (mAP) and normalised Discounted Cumulative Gain (nDCG) [Järvelin and Kekäläinen 2002] following [Damen et al. 2022]. nDCG measures the relevance of text descriptions in response to a video query by positively acknowledging semantically analogous descriptions in addition to the correct pair. For AudioEgoMCQ, we report the Retrieval@1 score, similar to [Lin et al. 2022]. We consider the following two tasks. For the more challenging **intra-video** task we are given a text query and aim to find the corresponding clip out of 5 clips selected from the same video. In contrast, the **inter-video** task uses a pool of 5 clips, each selected from a different video.

Models. We use two recent text-audio retrieval models, namely LAION-Clap [Y. Wu et al. 2023; K. Chen et al. 2022] and WavCaps [Mei et al. 2023]. Both models were trained and/or finetuned on the audio-centric AudioCaps [C. D. Kim et al. 2019] and Clotho [Drossos et al. 2020] datasets. We use these models to assess

Table 5.1: Zero-shot text-audio retrieval on AudioEpicMIR with LLM-generated audio descriptions compared to using visual descriptions. WavCaps-AC and WavCaps-Cl were finetuned on AudioCaps and Clotho respectively.

Pre-trained audio model	LLM-generated audio descriptions	mAP(%)			nDCG(%)		
		A->T	T->A	AVG	A->T	T->A	AVG
Random		5.6	6.4	6.0	10.7	12.3	11.5
LAION-Clap		8.9	8.4	8.6	15.3	17.0	16.2
LAION-Clap	✓	**10.0**	**9.3**	**9.6**	**17.2**	**18.2**	**17.7**
WavCaps-AC		10.3	9.0	9.7	17.5	17.5	17.5
WavCaps-AC	✓	**11.2**	**10.4**	**10.8**	**18.9**	**20.0**	**19.4**
WavCaps-Cl		10.9	9.5	10.2	18.3	18.2	18.2
WavCaps-Cl	✓	**11.5**	**10.4**	**10.9**	**19.2**	**20.2**	**19.7**

zero-shot egocentric text-audio retrieval capabilities, i.e. without any training on egocentric data.

Zero-shot egocentric text-audio retrieval. We evaluate the audio-centric pre-trained LAION-Clap and WavCaps models directly on the egocentric text-audio retrieval datasets (zero-shot setting) and provide the results on AudioEpicMIR in Tab. 5.1. We observe that the LLM-generated audio descriptions yield a consistent boost. We hypothesize that this improvement stems both from aligning the style of descriptions more closely with the training distribution for the models and from the inclusion of more audio-centric content. We additionally report results for AudioEgoMCQ in Tab. 5.2 and EpicSoundsRet in Tab. 5.3. For AudioEgoMCQ, we observe that for the intra-video task, using LLM descriptions provides consistent improvements over using the original labels. For the inter-video task we observe that the WavCaps model finetuned on Clotho yields a small decrease in performance as compared to using the original labels. We attribute the variation in task performance to the differing text queries and the more visual-centric nature of Clotho descriptions compared to AudioCaps. For EpicSoundsRet, using LLM descriptions gives a significant improvement over using the original audio class labels in most cases.

Evaluating text-audio retrieval on subsets with different audio relevancy. We employ GPT-4 to split the audio tracks in AudioEpicMIR into three subsets as described in Sec. 5.3.4. The subsets are selected by the LLM based

Table 5.2: Zero-shot text-audio retrieval results on AudioEgoMCQ with LLM-generated audio descriptions compared to using visual descriptions. The data is filtered as per Sec. 5.3.3.

Pre-trained audio model	LLM-generated audio descriptions	Intra-video(%)	Inter-video(%)
Random		20.0	20.0
LAION-Clap		24.2	27.5
LAION-Clap	✓	**25.2**	**28.9**
WavCaps-AC		24.1	30.9
WavCaps-AC	✓	**25.1**	**31.1**
WavCaps-Cl		24.6	**32.1**
WavCaps-Cl	✓	**25.6**	31.8

Table 5.3: Zero-shot text-audio retrieval on EpicSoundsRet with LLM-generated audio descriptions compared to using audio class labels.

Pre-trained audio model	LLM-generated audio descriptions	mAP(%)			nDCG(%)		
		A->T	T->A	AVG	A->T	T->A	AVG
Random		8.7	2.8	5.7	1.8	1.9	1.9
WavCaps-Cl		24.3	**12.0**	18.1	11.0	13.6	12.3
WavCaps-Cl	✓	30.2	11.3	20.8	14.8	13.7	14.3
LAION-Clap		28.5	8.2	18.3	16.1	9.5	12.8
LAION-Clap	✓	29.9	11.7	20.8	16.3	**14.3**	15.3
WavCaps-AC		24.3	11.9	18.2	12.4	13.8	13.1
WavCaps-AC	✓	**31.2**	11.9	**21.5**	**16.9**	**14.3**	**15.6**

on how difficult it is to identify the audio content solely from the sound. We compare the performance of different models to a random baseline on these newly created subsets in Tab. 5.4. The random baseline has been obtained by providing randomly sampled text descriptions as the 'correct' descriptions for a given audio. The audio retrieval model employed is WavCaps-Cl. We observe that for the subset deemed to have 'low' audio relevance, the random performance more than doubles in comparison to the random evaluation on the full test set. This is a result of the subset's pool of text descriptions being fairly repetitive. Hence, to ensure a more equitable comparison of model performance across the subsets, we consider the incremental improvement of each model from its initial random metrics (δ to rand. perf.). The accuracy increases as we go from the 'low' to the

Table 5.4: Zero-shot text-audio retrieval on different subsets of AudioEpicMIR, split according to the informativeness of audio files as judged by GPT-4. LLM-generated audio descriptions give the best results across all subsets (Aud LLM), compared to using the original visual descriptions (Aud orig). WavCaps-Cl performs best when the audio files are considered to be highly informative.

AudioEpicMIR subset	Descriptions	mAP(%)		nDCG(%)	
		AVG	δ to rand. perf.	AVG	δ to rand. perf.
Low	*Random perf.*	12.8	-	24.4	-
	Aud orig	15.5	2.7	28.2	3.8
	Aud LLM	15.6	2.8	27.9	3.5
Moderate	*Random perf.*	7.2	-	13.0	-
	Aud orig	11.5	4.3	19.1	6.1
	Aud LLM	12.4	5.2	20.2	7.2
High	*Random perf.*	5.7	-	9.7	-
	Aud orig	14.0	<u>8.3</u>	21.0	<u>11.3</u>
	Aud LLM	15.2	**9.5**	23.7	**14.0**

Table 5.5: Zero-shot text-audio retrieval on different subsets of AudioEgoMCQ, split according to the informativeness of audio files as judged by GPT-4. WavCaps-Cl performs best for highly informative audio files. Random values for both tasks are 20.0.

AudioEgoMCQ subset	Descriptions	Intra-video(%)	Inter-video(%)
Low	Aud orig	21.7	28.2
	Aud LLM	22.6	26.9
Moderate	Aud orig	25.0	32.7
	Aud LLM	26.8	34.2
High	Aud orig	30.3	39.0
	Aud LLM	**30.8**	**39.1**

'high' subset. At the same time, the incremental improvement over the random performance is most significant for the 'high' subset. We perform the same experiment on the AudioEgoMCQ subsets and observe that GPT-4 is indeed capable of selecting videos for which the audio is more likely to have informative content (Tab. 5.5).

5.5 Conclusion

In this study, we introduced three new benchmarks for egocentric text-audio retrieval. We proposed a methodology of generating audio descriptions using an LLM starting from visual-centric descriptions and audio class labels. Lastly, we have shown that we can use guidance from an LLM to filter out noisy audio content

extracted from video datasets. We believe these contributions can apply beyond the egocentric setting and hope they will improve text-audio understanding.

5.6 Supplementary Material

In this supplementary material, we will provide further experimental results and details regarding the pre-trained models that we used. In Sec. 5.6.1, we present additional information and analysis of the EpicSoundsRet benchmark based on the EpicSounds [Huh et al. 2023] dataset. For completeness, we include the prompts for generating audio descriptions in Sec. 5.6.2. In Sec. 5.6.3, we evaluate the impact of additionally using the audio modality for the text-video retrieval task. We provide more insight into the differences between the Clotho and AudioCaps finetuned checkpoints used in our work in Sec. 5.6.4. In Sec. 5.6.5 we briefly expand on the results obtained on the AudioEgoMCQ task whilst in Sec. 5.6.6 we provide more experiments on the AudioEgoMCQ benchmark. Lastly, we compare the relative performance differences of the models used for the experiments in the paper on multiple benchmarks in Sec. 5.6.7.

5.6.1 EpicSounds

In our experiments with the AudioEpicMIR and AudioEgoMCQ benchmarks, we generated *audio descriptions* using LLMs starting from *visual* class labels. For EpicSounds, we have access to *audio* class labels. As a result, our prompts are *not* necessarily optimal when using *audio* class labels as inputs. Nevertheless, despite being created from a set of prompts that might not be ideal, the *audio descriptions* still perform better in the retrieval task compared to the original audio labels as can be seen in Tab. 5.3 in the main paper.

The WavCaps model finetuned on Clotho performs slightly worse on the text-audio task when using LLM-generated descriptions on EpicSounds than when using the original audio labels. However, we do not see this behaviour when using models finetuned on AudioCaps. This discrepancy likely arises because Clotho descriptions often include substantial visual details in addition to audio content, leading to a significant style difference in our audio-focused descriptions. However, regardless of the model used, LLM-generated descriptions overall yield better results than using class labels, particularly in the audio-to-text retrieval direction.

We employed the same retrieval metrics for AudioEpicMIR and EpicSoundsRet. This required constructing a relevancy matrix (more information on this is avail-

able here [Wray et al. 2019; Damen et al. 2022]). This relevancy matrix takes into account that for one given description, multiple audio files can be equally relevant. On AudioEpicMIR, the relevancy matrix would look at all unique text descriptions, and depending on how many nouns and verbs the other descriptions would have in common, it would assign values between 0 and 1 to these similar descriptions. For EpicSoundsRet we start with audio class labels as 'descriptions' and assign a score of 1 to all audios and text descriptions where the text description belongs to the same class. We use the same relevancy matrix to calculate the retrieval scores for the *Aud orig* and *Aud LLM* descriptions. This also applies to AudioEpicMIR.

5.6.2 Prompts

We use LLMs, specifically GPT3.5, to generate audio descriptions starting from visual class labels. Tab. 5.6 shows the prompts we used.

Furthermore, we used GPT4 to decide the difficulty of identifying the action associated with a sound. The prompt we used for that is provided in Tab. 5.7.

5.6.3 Multimodal text-video retrieval

In this section, we show that text-video retrieval benefits from jointly using the audio and visual modalities. We report text-video retrieval results on the EpicMIR task in Tab. 5.8. The joint multimodal evaluation is carried out via late fusion of the outputs of the text-audio and text-video retrieval models. Boosts in performance for the multimodal evaluation compared to the unimodal models (i.e. audio only and visual only) are observed both when using the original descriptions (Joint w. Aud orig) and when using the original descriptions for the visual model together with the LLM-generated ones for the audio model (Joint w. Aud LLM).

Additionally, we investigate how text-video retrieval performs on the same audio relevancy subsets investigated in Tab. 5.4. Results are provided in Tab. 5.9 and follow a similar trend as for the audio only experiments. More specifically, the multimodal text-video retrieval performs best on the 'high' subset. We also notice that text-video retrieval when only using the visual modality also exhibits stronger performance on subsets with higher audio relevancy. We hypothesise that this is

Table 5.6: Overview of the methodology for bridging the gap between visual classes/descriptions and video soundtrack descriptions using LLMs. The process includes setting the scene, few-shot prompting with examples, and generating audio descriptions from video descriptions.

Index	Prompts
1.	You are an expert in audio and visual description of videos. You have seen the AudioCaps and Clotho datasets and know how to generate relevant audio descriptions using simple terms. You will help me generate audio descriptions that can match the audio content of a video for which the video description is provided. Avoid the use of object or actions that cannot be inferred from the audio signal alone. Try to create proper short sentences when generating the audio descriptions. When multiple audio descriptions can be possible, provide one that best generalises the sounds.Try to avoid generating audio descriptions in the form of 'sound of [visual object/action]' and instead use the actual noise or sound that object/action can make. If unsure, you can provide multiple possible sounds. To help you understand the sort of descriptions I am looking for, in the next prompt I will provide you a few pairs of video descriptions and the corresponding audio descriptions as an example. Please remember this prompt and the examples whenever generating new audio descriptions. Then I will ask you to generate audio descriptions given new video descriptions.
2.	Here are some examples in the form of (video description: audio description) and examples are separated by semicolon. ('burping', 'A man giving out a loud burp');('sneezing', 'Someone sneezes');('washing dishes', 'Metal clinking and clanging occur');('washing dishes', 'Water splashing and glasses clanging together then more clanging ending with glass squeaking sound');('opening door', 'Door handle continuously clicking then being pushed open');('spray painting', 'Powerful bursts of spraying');('crying', 'A baby cries and screams');('opening door', 'Hissing then creaking as a door is opened');('applauding', 'A large number of people clap, cheer, and shout');('closing door', 'Metal clings followed by rumbling and metal sliding');('whistling', 'A person is whistling a tune');('shearing sheep', 'Sheep bleat quietly');('washing hair', 'Water running while the stream is interrupted at times');('sawing wood', 'Rubbing and sawing of wood'). I want you to learn from them and not generate descriptions for this prompt.
3.	Generate an audio description for each of the following enumerated video descriptions separated by semicolon. Try to avoid generating audio descriptions in the form of 'sound of [visual object/action]' and instead use the actual noise or sound that object/action can make. If unsure, you can provide multiple possible sounds. Remember the original prompt. Provide your answers in the form [description index. video description: generated audio description]. <Add the list of visual labels separated by ; here>

Table 5.7: Prompts used for employing LLMs in assessing the likelihood of identifying video actions solely through audio tracks.

Index	Prompts
1.	You have a lot of experience with video and audio descriptions. You are working with videos that have an associate audio file. You only have descriptions of the visual content. Some videos are highly correlated with the audio content, such as those depicting someone cutting vegetables. Other videos have very generic audios and if only given the audio content, the video content might not be easy to figure out. I want you to tell me if a video description is relevant for the possible associated audio content. I want you to provide your answers in the form of a dictionary where the key is the description I am giving you and the value is the relevance. Relevance should be high if the sound associated is easy to assume. Relevance should be moderate if the associated sounds can be more than one. Relevance should be low if it's unlikely to hear any specific sounds. Process each entry individually. Input descriptions will be given in the form of a list. Do not provide additional comments, just the relevance.

Table 5.8: Zero-shot capabilities of text-audio (audio only) and text-video (video only) retrieval models compared to the joint evaluation with late fusion on EpicMIR. * Numbers are obtained using the text-video retrieval model from [Lin et al. 2022].

EpicMIR	mAP(%)			nDCG(%)		
	A->T	T->A	AVG	A->T	T->A	AVG
Random [Lin et al. 2022]	5.7	5.6	5.7	10.8	10.9	10.9
Audio only (WavCaps-Cl w. Aud LLM)	11.5	10.5	11.0	19.2	20.3	19.7
Video only (EgoVLP*)	24.7	18.4	21.6	27.4	24.6	26.0
Joint w. Aud orig	25.4	19.3	22.4	28.7	26.0	27.3
Joint w. Aud LLM	**25.8**	**19.9**	**22.8**	**29.1**	**26.9**	**28.0**

because the more audio significant actions tend to also be more common, or easier to identify for visual models.

For the text-video retrieval experiments, we used the EgoVLP [Lin et al. 2022] codebase. More precisely, we take the provided pre-trained EgoVLP [Show Lab 2023] model and evaluate it in a zero-shot fashion. This model processes the clips by randomly selecting a number of frames. We set the number of frames to 16 in our experiments. Additionally, we do not use the dual_softmax evaluation approach used in the codebase [Show Lab 2023] but instead leverage the standard similarity matrix at evaluation time. This is consistent with how the underlying

Table 5.9: Zero-shot multimodal text-video retrieval on different subsets of AudioEpicMIR, split according to the informativeness of audio files as judged by GPT-4. LLM-generated audio descriptions give the best results across all subsets (Aud LLM), compared to using the original visual descriptions (Aud orig). The joint model performs best when the audio files are considered to be highly informative.

AudioEpicMIR subset	Descriptions	mAP(%)		nDCG(%)	
		AVG	δ to rand. perf.	AVG	δ to rand. perf.
Low	*Random perf. vid.*	12.5	-	23.4	-
	Vid orig	23.3	10.8	30.8	7.4
	Aud orig	23.7	11.2	32.1	8.7
	Aud LLM	23.4	10.9	32.1	8.7
Moderate	*Random perf. vid.*	7.2	-	12.2	-
	Vid orig	27.3	20.1	28.0	15.8
	Aud orig	27.7	20.5	29.1	16.9
	Aud LLM	28.5	21.3	29.7	17.5
High	*Random perf. vid.*	5.8	-	9.1	-
	Vid orig	32.9	27.1	36.0	26.9
	Aud orig	33.6	27.8	37.1	28.0
	Aud LLM	34.6	**28.8**	38.3	**29.2**

EgoVLP model was trained [Show Lab 2023].

5.6.4 Leveraging models finetuned on Clotho for tasks with original visual descriptions

The Clotho [Drossos et al. 2020] dataset was collected by providing annotators with audio files and asking them to generate descriptions of the audio sound. However, by looking at descriptions, we observe that the annotators tended to provide plausible visual descriptions of the sounds rather than describe the actual sound. Some examples are “A group is in a carriage that is being drawn by a horse on a paved road." or “A man moves from the basement to upstairs, moving a heavy metal object with him.". Such visual details cannot be inferred just by listening to an audio. Therefore, this dataset matches better our setting of generating audio descriptions starting from visual descriptions since, when no obvious audio description can be generated, the output description often contains a reworded version of the provided *visual* input. In contrast, AudioCaps descriptions are shorter and more audio focused e.g. “Speech in the distance with a bleating sheep nearby." or “Food and oil sizzling followed by a woman speaking.". Furthermore, the WavCaps model has been finetuned on the Clotho data, which contains 25

Table 5.10: Zero-shot text-audio retrieval on different subsets of AudioEgoMCQ, split according to the informativeness of audio files as judged by GPT-4. WavCaps-AC performs best for highly informative audio files. Random values for both tasks are 20.0.

AudioEgoMCQ subset	Descriptions	Intra-video(%)	Inter-video(%)
Low	Aud orig	21.0	27.0
	Aud LLM	22.4	26.9
Moderate	Aud orig	26.2	32.7
	Aud LLM	26.6	33.2
High	Aud orig	28.7	37.0
	Aud LLM	**29.1**	**37.4**

times fewer training examples than AudioCaps. This could mean that the model finetuned on Clotho retains more generality which can be useful in settings with new *audio* content that the LLM might generate.

5.6.5 Intra-video vs inter-video results

When using the WavCaps-Cl model, we notice in Tab. 5.2 that the inter-video task performs slightly better when using original descriptions compared to using the LLM-generated ones. Based on our analysis, we believe that this is due to the different distribution of text queries between the two tasks (i.e. intra-video and inter-video).

5.6.6 Evaluation on different subsets of AudioEgoMCQ according to informativeness

We additionally evaluate the WavCaps model finetuned on AudioCaps on the *low*, *moderate*, and *high* subsets of AudioEgoMCQ in Tab. 5.10. We notice that when using this checkpoint, the inter-video LLM performance is higher on the *moderate* and *high* and just a bit lower on the *low* subset. This is in accordance with results in Tab. 5.5 where the checkpoint used was finetuned on Clotho. However, when using the AudioCaps finetuned model with LLM descriptions, the decrease on the *low* subset is much lower than when using the Clotho finetuned model. We believe that this is related to Clotho being more visual based as described in 5.6.4. As a result, when using AudioCaps finetuned models, the overall retrieval performance is better for the LLM descriptions than when using Clotho based models.

5.6.7 Why does WavCaps-Cl give better results on AudioEpicMIR and AudioEgoMCQ whilst the WavCaps-AC model is better on EpicSoundsRet?

We found that the WavCaps model, when finetuned with AudioCaps, performs best for EpicSounds. In contrast, for the other two datasets, the top-performing model is WavCaps finetuned on Clotho. This variation in performance is linked to the nature of the input labels provided to the LLM. Specifically, EpicSounds uses only audio inputs, while AudioEpicMIR and AudioEgoMCQ rely solely on visual inputs. LLMs tend to merge the visual and audio information in the same sentence, especially when the input leans more towards visual content. As AudioCaps descriptions primarily focus on audio, whereas Clotho provides a balanced mix of both audio and visual details. Tasks that are more audio-centric benefit significantly from models finetuned on AudioCaps. Conversely, tasks with a visual emphasis favour models finetuned on Clotho, due to their original labels being more visual-centric.

Chapter 6

Summary and extensions

In this chapter, we highlight this thesis' main contributions, provide a summary of works that have built on top of these contributions and discuss potential future research directions.

QuerYD dataset. In Chapter 2, we introduced a novel dataset, QuerYD, comprised of pairs of videos and their corresponding descriptions. QuerYD stands out for its integration of videos along with Audio Descriptions, which are derived from transcribed audio narrations tailored for users with visual impairments. This dataset serves as an important tool for assessing the performance of computational models on various tasks, including text-video retrieval, video captioning, and understanding of video content. The dataset exhibits considerable diversity, encompassing a varied vocabulary and videos from a wide range of categories, including cartoons, professional films, talk shows, and amateur recordings. We not only introduced the dataset but also presented results on the task of text-video retrieval, employing state-of-the-art models based on Convolutional Neural Networks available at that time. Nonetheless, the benchmark we proposed was constrained by the architectural designs of the models and the computational capabilities accessible during our research period.

With the emergence of foundation models and advancements in computational power, researchers are now equipped with more sophisticated tools for analysing videos, texts, and the intricate relationship between long videos and text. Consequently, QuerYD started receiving more interest, with researchers using it both as

a key dataset for training and as a relevant benchmark at evaluation time. While one recent study has used QuerYD for developing time-sensitive Large Language Models for long video understanding [Ren et al. 2024], other studies have leveraged it as a benchmark for long video understanding tasks, as evidenced by [Sun et al. 2022; Ren et al. 2023]. Although these newer models are more powerful and significantly enhance the performance beyond our original baselines, the task of text-video retrieval for long videos still needs to be solved. This ongoing challenge underscores the relevance of QuerYD as an effective benchmark for this domain.

Additionally, while numerous research studies aim to enhance video captioning performance, a limited number focus specifically on generating Audio Descriptions. QuerYD remains one of the very few datasets explicitly curated for this purpose that is also freely available to the research community. This accessibility facilitates the reproduction of results and fosters further improvements in the field. The collection of the QuerYD dataset has inspired further efforts in data collection for Audio Description tasks, exemplified by subsequent works such as those reported by [Han et al. 2023; Pitcher-Cooper et al. 2023]. This trend underscores the dataset's significant impact on advancing research in this domain.

Text-audio retrieval benchmark. In Chapter 3, we introduced the task of semantic search through an audio database using free-form text queries, extending beyond the traditional constraints of sound class-based audio retrieval. This novel approach employs state-of-the-art multimodal retrieval architectures, originally designed for text-video search, to effectively navigate the domain of text-audio search. In addition to providing benchmarks for the task of text-audio retrieval, we also collected and proposed a new dataset for the same task. Collecting new data was necessary as there were only two datasets containing free-form text descriptions of audio files at that time. These datasets were also limited in size, especially compared to their video-text counterparts. As a result, we introduced SoundDescs which is 10 times larger than AudioCaps and has a more varied distribution of audio durations and vocabulary.

Our work presented in Chapter 3 has significantly contributed to making the task of free-form text-audio retrieval an active area of research. Following our work, the popular Detection and Classification of Acoustic Scenes and Events

(DCASE) [*DCASE2022 Challenge Task 6: Automated Audio Captioning and Language-Based Audio Retrieval* 2022] community proposed a new challenge on the topic of language-based audio retrieval, underscoring the growing interest in this domain. In addition, current research focuses on two main directions. One is improving the performance of text-audio retrieval models by using foundation models [Mei et al. 2023; Y. Wu et al. 2023; Yeh et al. 2023]. The second research direction looks into collecting larger text-audio datasets that maximise the potential of these large models [Y. Wu et al. 2023; Mei et al. 2023]. Some of these works employ LLMs to generate audio captions starting from audio labels, repurposing audio classification tasks for the text-audio retrieval task [Mei et al. 2023]. Other works searched for other sources of text and audio pairs suited for this task [Y. Wu et al. 2023].

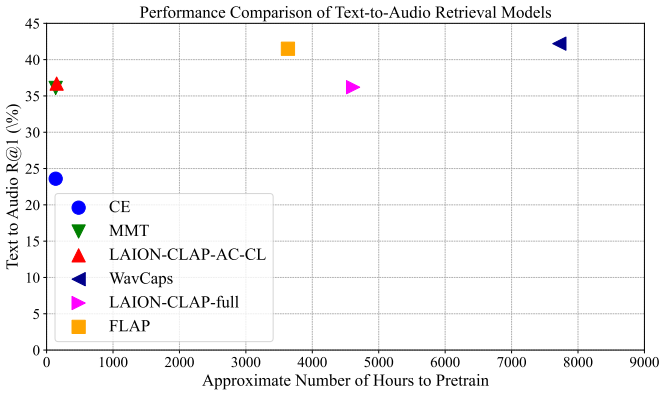
In spite of this progress in terms of the amount of data available and improved models, one interesting observation is that the text-audio retrieval performance reported in recent papers, for example on the AudioCaps dataset, has not improved that much. In Fig. 6.1, we notice how the text-audio retrieval R@1 changes with the number of hours of pre-training for different state-of-the-art models. We observe how the retrieval performance increases by only 5-7% which is equivalent to a relative increase in performance of 16%, whilst the amount of number of hours of training increases 40-70 times or by 2700%. In addition to the amount of data available, these models also suffer a considerable increase in the number of parameters. Therefore, provided we have enough data to train them, we would expect these models to perform better on existing benchmarks. Very recently, [S. Chen et al. 2024; Y. Wang et al. 2024] managed to obtain up to 55% R@1 on the text-audio AudioCaps retrieval. This is a result of training with very large multi-modal datasets containing video, audio, speech, and text data. This approach differs from the one employed in the earlier analysis that only leveraged text-audio datasets for training. However, it proves that more data could help maximise the potential of existing text-audio models.

This mismatch in how much the models and datasets have improved and the progress in the accuracy of the models is a clear indication that there is still a long way to go in model development and training. One direction of research can be to control the quality of the collected and curated data better. Manually checking audio descriptions in datasets such as Clotho [Drossos et al. 2020], we notice that

a lot of the descriptions contain an assumption of a visual action that leads to the sound (e.g. 'Outdoors, insect trapped in a spider web and then breaks loose and flies away.' [Drossos et al. 2020]), instead of a description of the actual sound (e.g. 'Buzzing sound and environmental noise'). This observation is supported by the results of our work in Chapter 5, where we compared models finetuned on Clotho with models finetuned on AudioCaps on the task of text-audio retrieval where the audios are soundtracks of video files. Here, we noticed that the Clotho-based model performs better on the task of text-audio retrieval than the AudioCaps-based finetuned models in the setting where the visual class labels are used as audio text descriptions. The behaviour is the opposite for the case where we use audio class labels as descriptions for the text-audio retrieval experiments. In this latter case, using AudioCaps finetuned models outperforms Clotho finetuned models. This is an empirical evaluation of how much the AudioCaps and Clotho descriptions rely on visual or audio-centric words. As we want to advance text-audio understanding even when the visual modality is absent, having cleaner and more uniform audio-focused descriptions could be beneficial.

An additional reason for the seemingly stagnating performance on the text-audio retrieval task, could be the metrics used. Recent works [X. Xu et al. 2024b] found that some descriptions of sounds encountered in the well-established AudioCaps dataset cannot be reasonably assumed to be different for the human ear. This can mean that we expect our models to differentiate between descriptions that actually sound the same or are very similar. Therefore, using metrics that take this into account might explain part of the problem regarding the accuracy of current models.

Text-audio retrieval and temporal ordering. In Chapter 4, we took one of the most recent transformer-based text-audio retrieval models and investigated if it understands the temporal ordering of sound events. We found that the analysed text-audio model does not have a good understanding of time. This is in line with the findings of [H.-H. Wu et al. 2023], that investigated a CNN-based audio encoder in their experiments. We further showed that the text-audio retrieval dataset AudioCaps [C. D. Kim et al. 2019] in its current form, is not a perfect dataset for the training and evaluation of a model's ability to understand temporal cues. Therefore, we introduced a synthetic dataset to evaluate if the model can



Model	Hours	Accuracy (%)
CE	136.92	23.6
MMT	136.92	36.1
LAION-CLAP-AC-CL	151.34	36.7
WavCaps	7719.26	42.2
LAION-CLAP-full	4614.44	36.2
FLAP	3633.23	41.5

Figure 6.1: Performance comparison of different models based on the training hours and accuracy. The corresponding scatter plot illustrates the R@1 performance achieved by each model against the number of hours spent in pretraining.

understand time. Lastly, we investigated using an additional contrastive text-text loss function, to guide the model to focus on temporal cues. We showed that this improves results in the more controlled, synthetic data setting.

This result underlines the need to investigate our text-audio benchmarks further, to ensure they are suited for the tasks we want to evaluate.

Repurposing large text-video datasets. In Chapter 5, we investigated how feasible it is to leverage the multitude of large-scale text-video datasets for the task of zero-shot text-audio retrieval using Large Language Models.

We successfully generated text-audio pairs starting from text-video pairs in the egocentric setting by using LLMs as a translator from visual-based descriptions to audio-based descriptions. This is possible thanks to LLMs' understanding of the visual and acoustic world.

Visual label	**Generated audio description**
bird	Continuous chirping with varied pitches.
dog	Loud barking with intermittent pauses.
Robin bird	A series of short, melodious trills and whistles.
door closing	A sharp thud followed by a faint echo.
Chaffinch bird	A repetitive, high-pitched tweet followed by a rapid trill.

Table 6.1: Examples of varied audio descriptions generated with GPT-4, starting from visual labels.

Additionally, having LLMs generate plausible audio descriptions could lead to more varied sets of audio captions, given that enough visual information is provided.

For example, asking GPT-4 to generate plausible audio descriptions starting from visual content, we obtain the captions shown in Tab. 6.1.

It can be observed that more detailed visual descriptions result in more specific audio descriptions. This is in contrast to AudioCaps and Clotho descriptions which are more repetitive and less specific, allowing for more high-level retrieval.

We hope that our approach inspires further research into finding efficient ways of using LLMs to deal with the scarcity of text-audio datasets, removing the need for expensive and time-consuming data annotation.

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan (2016). "Youtube-8m: A large-scale video classification benchmark". In: *arXiv:1609.08675*.

Explosion AI (2016). *Spacy: natural language processing library*. https://spacy.io.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell (2017). "Localizing moments in video with natural language". In: *ICCV*.

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic (2016). "NetVLAD: CNN architecture for weakly supervised place recognition". In: *Proc. CVPR*.

Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman (2023). "HierVL: Learning Hierarchical Video-Language Embeddings". In: *Proc. CVPR*.

Pradeep K Atrey, Namunu C Maddage, and Mohan S Kankanhalli (2006). "Audio based event detection for multimedia surveillance". In: *Proc. ICASSP*.

Jean-Julien Aucouturier, Boris Defreville, and Francois Pachet (2007). "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music". In: *JASA*.

Audio-Retrieval Project Page (n.d.). URL: https://www.robots.ox.ac.uk/~vgg/research/audio-retrieval/.

Pavlos Avgoustinakis, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Andreas L Symeonidis, and Ioannis Kompatsiaris (2020). "Audio-based Near-Duplicate Video Retrieval with Audio Similarity Learning". In: *arXiv:2010.08737*.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba (2017). "See, hear, and read: Deep aligned representations". In: *arXiv:1706.00932*.

Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman (2020). "Condensed Movies: Story Based Retrieval with Contextual Embeddings". In: *Proc. ACCV*.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman (2021). "Frozen in time: A joint video and image encoder for end-to-end retrieval". In: *ICCV*.

Satanjeev Banerjee and Alon Lavie (2005). "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments". In: *Proc. ACL Workshop*.

Andrei Barbu, Alexander Bridge, Zachary Burchill, Dan Coroian, Sven Dickinson, Sanja Fidler, Aaron Michaux, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, et al. (2012). "Video in sentences out". In: *arXiv:1204.2742*.

Swapnil Bhosale, Rupayan Chakraborty, and Sunil Kumar Kopparapu (2023). "A Novel Metric For Evaluating Audio Caption Similarity". In: *Proc. ICASSP*.

Pieter Broucke (July 2021). *Neo-Sumerian Record Keeping: A Cuneiform Tablet Fragment from a Mesopotamian Archive*. URL: https://sites.middlebury.edu/middartmuseum/2021/07/06/neo-sumerian-record-keeping-a-cuneiform-tablet-fragment-from-a-mesopotamian-archive/.

Lionel Casson (2001). *Libraries in the Ancient World*. [Online]. New Haven: Yale University Press. URL: https://doi.org/10.12987/9780300133134.

Gal Chechik, Eugene Ie, Martin Rehn, Samy Bengio, and Dick Lyon (2008). "Large-scale content-based audio retrieval from text queries". In: *Proc. ACM ICMIR*.

David L Chen and William B Dolan (2011). "Collecting highly parallel data for paraphrase evaluation". In: *ACL*.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman (2020). "VGGSound: A Large-scale Audio-Visual Dataset". In: *International Conference on Acoustics, Speech, and Signal Processing*.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov (2022). "HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection". In: *ICASSP*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders,

Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba (2021). *Evaluating Large Language Models Trained on Code*.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu (2024). "Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset". In: *NeurIPS*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton (2020). "A simple framework for contrastive learning of visual representations". In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing (2023). *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*. URL: https://lmsys.org/blog/2023-03-30-vicuna/.

Ioana Croitoru, Simion-Vlad Bogolin, Yang Liu, Samuel Albanie, Marius Leordeanu, Hailin Jin, and Andrew Zisserman (2021). "TEACHTEXT: CrossModal Generalized Distillation for Text-Video Retrieval". In: *ICCV*.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray (2018). "Scaling Egocentric Vision: The EPIC-KITCHENS Dataset". In: *Proc. ECCV*.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray (2022). "Rescaling Egocentric Vision". In: *IJCV*.

Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso (2013). "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching". In: *CVPR*.

DCASE2020 Challenge Task 6: Automated Audio Captioning (2020). URL: http://dcase.community/challenge2020/task-automatic-audio-captioning.

DCASE2022 Challenge Task 6: Automated Audio Captioning and Language-Based Audio Retrieval (2022). URL: https://dcase.community/challenge2022/task-language-based-audio-retrieval.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (2009). "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR*.

Soham Deshmukh, Benjamin Elizalde, Dimitra Emmanouilidou, Bhiksha Raj, Rita Singh, and Huaming Wang (2024). "Training Audio Captioning Models without Audio". In: *Proc. ICASSP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv:1810.04805*.

Jared M. Diamond (1999). "Guns, germs, and steel: the fates of human societies". In: New York: Norton. Chap. BLUEPRINTS AND BORROWED LETTERS.

Jianfeng Dong, Xirong Li, and Cees GM Snoek (2016). "Word2visualvec: Image and video to sentence matching by visual feature prediction". In: *arXiv:1604.06838*.

Jianfeng Dong, Xirong Li, and Cees GM Snoek (2018). "Predicting visual features from text for image and video caption retrieval". In: *IEEE Transactions on Multimedia*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *Proc. ICLR*.

Konstantinos Drossos, Sharath Adavanne, and Tuomas Virtanen (2017). "Automated audio captioning with recurrent neural networks". In: *IEEE WASPAA*.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen (2020). "Clotho: an Audio Captioning Dataset". In: *Proc. ICASSP*.

Benjamin Elizalde, Rohan Badlani, Ankit Shah, Anurag Kumar, and Bhiksha Raj (2018). "Nels-never-ending learner of sounds". In: *arXiv:1801.05544*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang (2023). "Clap learning audio concepts from natural language supervision". In: *Proc. ICASSP*.

Benjamin Elizalde, Shuayb Zarar, and Bhiksha Raj (2019). "Cross modal audio search and retrieval with joint embeddings based on text and audio". In: *Proc. ICASSP*.

Ayşegül Özkaya Eren and Mustafa Sert (2020). "Audio Captioning Based on Combined Audio and Semantic Embeddings". In: *IEEE ISM*.

Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell (2019). "Temporal Localization of Moments in Video Collections with Natural Language". In: *arXiv:1907.12763*.

Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel PW Ellis, and Xavier Serra (2019). "Audio tagging with noisy labels and minimal supervision". In: *arXiv:1906.02975*.

Frederic Font, Gerard Roma, and Xavier Serra (2013). "Freesound technical demo". In: *Proc. ACM Multimedia*.

Jonathan T Foote (1997). "Content-based retrieval of music and audio". In: *Multimedia Storage and Archiving Systems II*. International Society for Optics and Photonics.

Logan Ford, Hao Tang, François Grondin, and James R Glass (2019). "A Deep Residual Network for Large-Scale Acoustic Scene Analysis." In: *INTERSPEECH*.

Peter Foster, Siddharth Sigtia, Sacha Krstulovic, Jon Barker, and Mark D Plumbley (2015). "Chime-home: A dataset for sound source recognition in a domestic environment". In: *IEEE WASPAA*.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid (2020). "Multi-modal transformer for video retrieval". In: *ECCV*.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia (2017). "Tall: Temporal activity localization via language query". In: *ICCV*.

Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen (2017). "Video captioning with attention-based LSTM and semantic consistency". In: *IEEE Transactions on Multimedia*.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter (2017). "Audio set: An ontology and human-labeled dataset for audio events". In: *Proc. ICASSP*.

Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan (2019). "Large-scale weakly-supervised pre-training for video action recognition". In: *Proc. CVPR*.

Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C. Smith (1995). "Query by humming: musical information retrieval in an audio database". In: *Proc. ACMM*.

Yuan Gong, Yu-An Chung, and James R. Glass (2021). "AST: Audio Spectrogram Transformer". In: *arXiv:2104.01778*.

Google (n.d.). *Speech-to-Text*. https://cloud.google.com/speech-to-text.

Kristen Grauman, Michael Wray, Adriano Fragomeni, Jonathan P N Munro, Will Price, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, C.V. Jawahar, Kris Kitani, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Mike Zheng Shou, Antonio Torrallba, and Jitendra Malik (2022). "Ego4D: Around the World in 3,000 Hours of Egocentric Video". In: *Proc. CVPR*.

Kim Hammar, Shatha Jaradat, Nima Dokoohaki, and Mihhail Matskin (2018). "Deep Text Mining of Instagram Data without Strong Supervision". In: *IEEE/WIC/ACM International Conference on Web Intelligence (WI)*.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman (2023). "AutoAD: Movie Description in Context". In: *Proc. CVPR*.

Tengda Han, Weidi Xie, and Andrew Zisserman (2022). "Temporal Alignment Networks for Long-Term Video". In: *Proc. CVPR*.

Y.N. Harari (2014). *Sapiens: A Brief History of Humankind*. Harvill Secker. URL: https://books.google.co.uk/books?id=B4ARBAAAQBAJ.

David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass (2018). "Jointly discovering visual objects and spoken words from raw sensory input". In: *ECCV*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep Residual Learning for Image Recognition". In: *Proc. CVPR*.

Marko Helén and Tuomas Virtanen (2007). "Query by example of audio signals using Euclidean distance between Gaussian mixture models". In: *Proc. ICASSP*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt (2021a). "Aligning AI With Shared Human Values". In: *Proc. ICLR*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt (2021b). "Measuring Massive Multitask Language Understanding". In: *Proc. ICLR*.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. (2016). "CNN Architectures for Large-Scale Audio Classification". In: *Proc. ICASSP*.

Sujuan Hou and Shangbo Zhou (2013). "Audio-visual-based query by example video retrieval". In: *Mathematical Problems in Engineering*.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger (2017). "Densely connected convolutional networks". In: *CVPR*.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao (2023). "Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models". In: *Proc. ICML*.

Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman (2023). "EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound". In: *Proc. ICASSP*.

Vladimir Iashin and Esa Rahtu (2020). "Multi-Modal Dense Video Captioning". In: *CVPR Workshops*.

Shota Ikawa and Kunio Kashino (2018a). “Acoustic event search with an onomatopoeic query: measuring distance between onomatopoeic words and sounds”. In: *Proc. DCASE Workshop*.

Shota Ikawa and Kunio Kashino (2018b). “Generating sound words from audio signals of acoustic events with sequence-to-sequence model”. In: *Proc. ICASSP*.

Vanita Jain, Fadi Al-Turjman, Gopal Chaudhary, Devang Nayar, Varun Gupta, and Aayush Kumar (2022). “Video captioning: a review of theory, techniques and practices”. In: *Multimedia Tools and Applications*.

Kalervo Järvelin and Jaana Kekäläinen (2002). “Cumulated Gain-Based Evaluation of IR Techniques”. In: *ACM Trans. Inf. Syst.*

Cong Jin, Tian Zhang, Shouxun Liu, Yun Tie, Xin Lv, Jianguang Li, Wencai Yan, Ming Yan, Qian Xu, Yicong Guan, et al. (2021). “Cross-modal deep learning applications: audio-visual retrieval”. In: *Proc. ICPR*.

Qin Jin, Peter Schulam, Shourabh Rawat, Susanne Burger, Duo Ding, and Florian Metze (2012). “Event-based video retrieval using audio”. In: *Proc. ISCA*.

Michael I Jordan and Robert A Jacobs (1994). “Hierarchical mixtures of experts and the EM algorithm”. In: *Neural computation*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer (2017). “TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension”. In: *ACL*.

Prannay Kaul, Weidi Xie, and Andrew Zisserman (2023). “Multi-Modal Classifiers for Open-Vocabulary Object Detection”. In: *Proc. ICML*.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. (2017). “The kinetics human action video dataset”. In: *arXiv:1705.06950*.

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen (2019). “EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition”. In: *Proc. ICCV*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim (2019). “AudioCaps: Generating captions for audios in the wild”. In: *Proc. NACCL*.

A. Sophia Koepke, Andreea-Maria Oncescu, João F. Henriques, Zeynep Akata, and Samuel Albanie (2022). “Audio retrieval with natural language queries: A benchmark study”. In: *IEEE Transactions on Multimedia*.

Yuma Koizumi, Yasunori Ohishi, Daisuke Niizumi, Daiki Takeuchi, and Masahiro Yasuda (2020). “Audio Captioning using Pre-Trained Large-Scale

Language Model Guided by Audio-based Similar Caption Retrieval". In: *arXiv:2012.07331*.

Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga (2002). "Natural language description of human activities from video images based on concept hierarchy of actions". In: *IJCV*.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley (2020). "Panns: Large-scale pretrained audio neural networks for audio pattern recognition". In: *IEEE/ACM TASLP*.

Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D Plumbley (2018). "Audio set classification with attention model: A probabilistic perspective". In: *Proc. ICASSP*.

Qiuqiang Kong, Changsong Yu, Yong Xu, Turab Iqbal, Wenwu Wang, and Mark D Plumbley (2019). "Weakly labelled audioset tagging with attention neural networks". In: *IEEE/ACM TASLP*.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi (2023). "AudioGen: Textually Guided Audio Generation". In: *Proc. ICLR*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles (2017). "Dense-captioning events in videos". In: *Proc. ICCV*.

Hilde Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso Poggio, and Thomas Serre (2011). "HMDB: A large video database for human motion recognition". In: *Proc. ICCV*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov (2019). "Natural Questions: A Benchmark for Question Answering Research". In: *ACL*.

Ianis Lallemand, Diemo Schwarz, and Thierry Artières (2012). "Content-based retrieval of environmental sounds by multiresolution analysis". In: *Proc. SMC*.

Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld (2008). "Learning realistic human actions from movies". In: *Proc. CVPR*.

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Zhongcong Xu, Denial Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, WANG HongFa, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou (2022). "Egocentric Video-Language Pretraining". In: *NeurIPS*.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han (2019). "On the variance of the adaptive learning rate and beyond". In: *arXiv:1908.03265*.

Meng Liu, Liqiang Nie, Yunxiao Wang, Meng Wang, and Yong Rui (2023). "A Survey on Video Moment Localization". In: *ACM Comput. Surv.*

Xubo Liu, Qiushi Huang, Xinhao Mei, Tom Ko, H Lilian Tang, Mark D Plumbley, and Wenwu Wang (2021). "CL4AC: A Contrastive Loss for Audio Captioning". In: *arXiv:2107.09990*.

Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman (2019). "Use What You Have: Video Retrieval Using Representations From Collaborative Experts". In: *Proc. BMVC*.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li (2021). "Clip4clip: An empirical study of clip for end to end video clip retrieval". In: *arXiv:2104.08860*.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten (2018). "Exploring the limits of weakly supervised pretraining". In: *ECCV*.

Pranay Manocha, Rohan Badlani, Anurag Kumar, Ankit Shah, Benjamin Elizalde, and Bhiksha Raj (2018). "Content-based representations of audio using siamese neural networks". In: *Proc. ICASSP*.

Xinhao Mei, Xubo Liu, Qiushi Huang, Mark D. Plumbley, and Wenwu Wang (2021). "Audio Captioning Transformer". In: *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*.

Xinhao Mei, Xubo Liu, Jianyuan Sun, Mark D. Plumbley, and Wenwu Wang (2022). "On Metric Learning for Audio-Text Cross-Modal Retrieval". In: *INTERSPEECH*.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang (2023). "WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research". In: *arXiv:2303.17395*.

Sachit Menon and Carl Vondrick (2023). "Visual Classification via Description from Large Language Models". In: *Proc. ICLR*.

Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen (2017a). "DCASE 2017 challenge setup: Tasks, datasets and baseline system". In: *DCASE 2017*.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen (2017b). *TUT Acoustic scenes 2017, Development dataset*.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman (2019a). "End-to-End Learning of Visual Representations from Uncurated Instructional Videos". In: *arXiv:1912.06430*.

Antoine Miech, Ivan Laptev, and Josef Sivic (2018). "Learning a text-video embedding from incomplete and heterogeneous data". In: *arXiv:1804.02516*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic (2019b). "Howto100M: Learning a text-video embedding by watching hundred million narrated video clips". In: *ICCV*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient estimation of word representations in vector space". In: *arXiv:1301.3781*.

Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury (2018). "Learning joint embedding with multimodal cues for cross-modal video-text retrieval". In: *Proc. ACM ICMR*.

Himangi Mittal, Pedro Morgado, Unnat Jain, and Abhinav Gupta (2022). "Learning State-Aware Visual Representations from Audible Interactions". In: *NeurIPS*.

Manuel Molina (2016). "Archives and Bookkeeping in Southern Mesopotamia during the Ur III period". In: *Comptabilités*. Accessed: 10 April, 2024. URL: http://journals.openedition.org/comptabilites/1980.

Arsha Nagrani, Samuel Albanie, and Andrew Zisserman (2018). "Learnable PINs: Cross-Modal Embeddings for Person Identity". In: *Proc. ECCV*.

Andreea-Maria Oncescu, João F. Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie (2021a). "QuerYD: a video dataset with high-quality text and audio narrations". In: *Proc. ICASSP*.

Andreea-Maria Oncescu, João F. Henriques, Andrew Zisserman, Samuel Albanie, and A Sophia Koepke (2024). "A SOUND APPROACH: Using Large Language Models to generate audio descriptions for egocentric text-audio retrieval". In: *Proc. ICASSP*.

Andreea-Maria Oncescu, A. Sophia Koepke, João F. Henriques, Zeynep Akata, and Samuel Albanie (2021b). "Audio Retrieval with Natural Language Queries". In: *INTERSPEECH*.

OpenAI (2023). "GPT-4 Technical Report". In: *arXiv:2303.08774*.

OpenAI (n.d.). *ChatGPT*. https://openai.com/blog/chatgpt. Accessed July, August 2023.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and

Raquel Fernández (2016). "The LAMBADA dataset: Word prediction requiring a broad discourse context". In: *ACL*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay (2023). "The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Only". In: *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Karol J Piczak (2015). "ESC: Dataset for environmental sound classification". In: *Proc. ACM Multimedia*.

Charity Pitcher-Cooper, Manali Seth, Benjamin Kao, James M Coughlan, and Ilmi Yoon (2023). "You Described, We Archived: A rich audio description dataset". In: *Journal on Technology and Persons with Disabilities*.

Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo (2022). "Domain Generalization Through Audio-Visual Relative Norm Alignment in First Person Action Recognition". In: *Proc. WACV*.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang (2023). "EgoVLPv2: Egocentric Video-Language Pre-training with Fusion in the Backbone". In: *Proc. ICCV*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. (2021). "Learning transferable visual models from natural language supervision". In: *Proc. ICML*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research*.

Ranger optimiser (n.d.). URL: https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar (2018). "Do CIFAR-10 Classifiers Generalize to CIFAR-10?" In: *arXiv:1806.00451*.

Shuhuai Ren, Sishuo Chen, Shicheng Li, Xu Sun, and Lu Hou (2023). "TESTA: Temporal-Spatial Token Aggregation for Long-form Video-Language Understanding". In: *EMNLP*.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou (2024). "TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding". In: *Proc. CVPR*.

Video Description Research and Development Center (2013). *YouDescribe*. URL: http://youdescribe.ski.org.

Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele (2014). "Coherent multi-sentence video description with variable level of detail". In: *GCPR*. Springer.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele (2015). "A dataset for movie description". In: *CVPR*.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap (2017). "A simple neural network module for relational reasoning". In: *NeurIPS*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang (2023). "HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace". In: *arXiv:2303.17580*.

Show Lab (2023). *EgoVLP: A Repository for Ego-centric Vision Language Pre-training*. https://github.com/showlab/EgoVLP.

Malcolm Slaney (2002a). "Mixtures of probability experts for audio retrieval and indexing". In: *Proc. IEEE ICME*.

Malcolm Slaney (2002b). "Semantic-audio retrieval". In: *Proc. ICASSP*.

Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman (2020). "A Short Note on the Kinetics-700-2020 Human Action Dataset". In: *arXiv:2010.10864*.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng (2014). "Grounded compositional semantics for finding and describing images with sentences". In: *Transactions of the Association for Computational Linguistics*.

Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley (2015). "Detection and classification of acoustic scenes and events". In: *IEEE Transactions on Multimedia*.

Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu (2022). "Long-Form Video-Language Pre-Training with Multimodal Temporal Contrastive Learning". In: *NeurIPS*.

Didac Surís, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i-Nieto (2018). "Cross-modal embeddings for video and audio retrieval". In: *Proc. ECCVW*.

Haoyu Tang, Jihua Zhu, Qinghai Zheng, and Zhiyong Cheng (2021). "Query-graph with Cross-gating Attention Model for Text-to-Audio Grounding". In: *arXiv:2106.14136*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample (2023a). "LLaMA: Open and Efficient Foundation Language Models". In: *arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom (2023b). "Llama 2: Open Foundation and Fine-Tuned Chat Models". In: *arXiv:2307.09288*.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri (2018). "A closer look at spatiotemporal convolutions for action recognition". In: *CVPR*.

Laurens Van der Maaten and Geoffrey Hinton (2008). "Visualizing data using t-SNE." In: *Journal of machine learning research*.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *NeurIPS*.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko (2015). "Sequence to sequence-video to text". In: *Proc. ICCV*.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang (2019). "VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research". In: *Proc. ICCV*.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang (2024). "InternVideo2: Scaling Video Foundation Models for Multimodal Video Understanding". In: *ECCV*.

Erling Wold, Thom Blum, Douglas Keislar, and James Wheaten (1996). "Content-based classification, search, and retrieval of audio". In: *IEEE Multimedia*.

Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen (2019). "Fine-grained action retrieval through multiple parts-of-speech embeddings". In: *Proc. ICCV*.

Ho-Hsiang Wu, Oriol Nieto, Juan Pablo Bello, and Justin Salamon (2023). "Audio-Text Models Do Not Yet Leverage Natural Language". In: *Proc. ICASSP*.

Mengyue Wu, Heinrich Dinkel, and Kai Yu (2019). "Audio caption: Listen and tell". In: *Proc. ICASSP*.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov (2023). "Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation". In: *Proc. ICASSP*.

Huang Xie, Okko Räsänen, Konstantinos Drossos, and Tuomas Virtanen (2022). "Unsupervised Audio-Caption Aligning Learns Correspondences Between Individual Sound Events and Textual Phrases". In: *Proc. ICASSP*.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy (2018). "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification". In: *ECCV*.

Yifei Xin, Dongchao Yang, and Yuexian Zou (2023). "Improving Text-Audio Retrieval by Text-Aware Attention Pooling and Prior Matrix Revised Loss". In: *Proc. ICASSP*.

Ziyou Xiong, Regunathan Radhakrishnan, Ajay Divakaran, and Thomas S Huang (2003). "Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework". In: *Proc. IEEE ICME*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui (2016). "Msr-vtt: A large video description dataset for bridging video and language". In: *CVPR*.

Ran Xu, Caiming Xiong, Wei Chen, and Jason Corso (2015). "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework". In: *AAAI*.

Xuenan Xu, Heinrich Dinkel, Mengyue Wu, Zeyu Xie, and Kai Yu (2021a). "Investigating Local and Global Information for Automated Audio Captioning with Transfer Learning". In: *arXiv:2102.11457*.

Xuenan Xu, Heinrich Dinkel, Mengyue Wu, and Kai Yu (2021b). "Text-to-Audio Grounding: Building Correspondence Between Captions and Sound Events". In: *Proc. ICASSP*.

Xuenan Xu, Ziyang Ma, Mengyue Wu, and Kai Yu (2024a). "Towards Weakly Supervised Text-to-Audio Grounding". In: *IEEE Transactions on Multimedia*.

Xuenan Xu, Mengyue Wu, and Kai Yu (2023). "Investigating Pooling Strategies and Loss Functions for Weakly-Supervised Text-to-Audio Grounding via Contrastive Learning". In: *International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*.

Xuenan Xu, Xiaohang Xu, Zeyu Xie, Pingyue Zhang, Mengyue Wu, and Kai Yu (2024b). "A Detailed Audio-Text Data Simulation Pipeline using Single-Event Sounds". In: *Proc. ICASSP*.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu (2023). "Diffsound: Discrete Diffusion Model for Text-to-Sound Generation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville (2015). "Describing videos by exploiting temporal structure". In: *ICCV*.

Masahiro Yasuda, Yasunori Ohishi, Yuma Koizumi, and Noboru Harada (2020). "Crossmodal Sound Retrieval Based on Specific Target Co-Occurrence Denoted with Weak Labels." In: *INTERSPEECH*.

Ching-Feng Yeh, Po-Yao Huang, Vasu Sharma, Shang-Wen Li, and Gargi Gosh (2023). "Flap: Fast Language-Audio Pre-Training". In: *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Changsong Yu, Karim Said Barsim, Qiuqiang Kong, and Bin Yang (2018). "Multi-level attention model for weakly supervised audio classification". In: *arXiv:1803.02353*.

Donghuo Zeng, Yi Yu, and Keizo Oyama (2020). "Deep triplet neural networks with cluster-cca for audio-visual cross-modal retrieval". In: *ACM TOMM*.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer (2022). "Scaling Vision Transformers". In: *Proc. CVPR*.

Bowen Zhang, Hexiang Hu, and Fei Sha (2018). "Cross-modal and hierarchical modeling of video and text". In: *ECCV*.

Michael R Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba (2019). "Lookahead optimizer: k steps forward, 1 step back". In: *NeurIPS*.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar (2023). "Learning Video Representations from Large Language Models". In: *Proc. CVPR*.

Yujie Zhong, Relja Arandjelović, and Andrew Zisserman (2018). "Ghostvlad for set-based face recognition". In: *Proc. ACCV*.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba (2017). "Places: A 10 million image database for scene recognition". In: *IEEE transactions on pattern analysis and machine intelligence*.

Luowei Zhou, Chenliang Xu, and Jason J Corso (2018). "Towards Automatic Learning of Procedures From Web Instructional Videos". In: *AAAI*.

Cunjuan Zhu, Qi Jia, Wei Chen, Yanming Guo, and Yu Liu (2023). "Deep learning for video-text retrieval: a review". In: *International Journal of Multimedia Information Retrieval*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny (2023). "MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models". In: *arXiv:2304.10592*.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015). "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books". In: *Proc. ICCV*.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic (2019). "Cross-task weakly supervised learning from instructional videos". In: *ICCV*.

Appendix A

Audio Retrieval with Natural Language Queries

This paper was accepted for publication at INTERSPEECH, 2021.

Audio Retrieval with Natural Language Queries

Andreea-Maria Oncescu[1*] A. Sophia Koepke[2*]
João F. Henriques[1] Zeynep Akata[3] Samuel Albanie[1]
[1]University of Oxford [2]University of Tübingen
[*]Authors contributed equally.

Abstract

We consider the task of retrieving audio using free-form natural language queries. To study this problem, which has received limited attention in the existing literature, we introduce challenging new benchmarks for text-based audio retrieval using text annotations sourced from the AudioCaps and Clotho datasets. We then employ these benchmarks to establish baselines for cross-modal audio retrieval, where we demonstrate the benefits of pre-training on diverse audio tasks. We hope that our benchmarks will inspire further research into cross-modal text-based audio retrieval with free-form text queries.

A.1 Introduction

The large amounts of data being generated in recent years have prompted a real need to search evergrowing databases. A search query typically comprises natural language (text), which allows expressing virtually any concept. Spanning multiple modalities, different retrieval strategies were developed for content as diverse as text (including webpages and books), images [Dong et al. 2016], and videos [Miech et al. 2018; Mithun et al. 2018]. Surprisingly, while search engines currently exist for these modalities (e.g. Google, Flickr and YouTube, respectively), unstructured audio is not accessible in the same way. The aim of this paper is to address this gap.

It is important to distinguish content-based retrieval from that based on meta-

data, such as the title of a video or song or an audio tag. Meta-data retrieval is feasible for manually-curated databases such as song or movie catalogs. However, content-based retrieval is more important in user-generated data, which often has little structure or meta-data. There are methods to search for audio which matches an audio query [Manocha et al. 2018; Lallemand et al. 2012], but satisfying the requirement to input an example audio query can be difficult for a human (e.g. making convincing frog sounds is difficult). We, on the other hand, propose a framework which enables the querying of a sound database using detailed free-form natural language descriptions of the desired sound (e.g. "A man talking as music is playing followed by a frog croaking."). This enables the retrieval of audio data which matches the temporal sequence of events in the query instead of just a single class tag. Furthermore, natural language queries are a familiar user interface widely used in current search engines. Therefore, our proposed audio retrieval with free-form text queries could be a first step towards more natural and flexible audio-only search.

Text-based audio retrieval could also be beneficial for video retrieval. The majority of recent works that address the text-based video retrieval task focuses heavily on the visual and text domains [Dong et al. 2016; Miech et al. 2018; Y. Liu et al. 2019; Gabeur et al. 2020]. Since audio and visual information inherently have natural semantic alignment for a significant portion of video data, text-based audio retrieval could also be used for querying video databases by only considering the audio stream of the video data. This would allow for video retrieval in the audio domain at reduced computational cost for cases in which audio and visual information correspond, with applications for low-power IoT devices, such as microphones in natural habitats, of particular interest for conservation and biology. Historical archives with extensive sound collections, such as the BBC Sound Effects Archive, would be easier to search, facilitating historical research and public access. Furthermore, text-based retrieval could enable appealing creative applications, such as automatically finding sounds which correspond to input text. This could be especially useful given the growing popularity of audio podcasts and audiobooks which are often supplemented with (background) sounds that match their content.

We propose two new benchmarks for text-based audio retrieval, based on the AudioCaps [C. D. Kim et al. 2019] and Clotho [Drossos et al. 2020] datasets. The for-

mer consists of a subset of 10-second audio clips from the AudioSet dataset [Gemmeke et al. 2017] with additional human-written audio captions, while the latter contains audio captions for sounds sourced from the Freesound platform [Font et al. 2013], varying between 15 and 30 seconds in duration. In contrast to sound event class labels, audio captions contain detailed information about the sounds. A user searching for a particular sound would usually describe the sound using text similar to an audio caption. AudioCaps [C. D. Kim et al. 2019] and Clotho [Drossos et al. 2020] allow to leverage the matching audio-caption pairs to train text-based audio retrieval frameworks. To establish baselines for this task, we adapt existing video retrieval frameworks for audio retrieval. We employ multiple pre-trained audio expert networks and show that using an ensemble of audio experts improves audio retrieval.

In summary, we make three contributions: (1) we introduce two new benchmarks for free-form text-based audio retrieval—to the best of our knowledge, these represent the first public benchmarks for this task; (2) we provide baseline performances with existing multi-modal video retrieval models that we adapt to text-based audio retrieval; (3) we demonstrate the benefits of combining multiple pre-training datasets.

The remainder of the paper is structured as follows. We review related work in Sec. A.2 and describe the datasets, benchmarks and methodology in Sec. A.3. We then report our experimental findings in Sec. A.4 before finally concluding in Sec. A.5.

A.2 Related work

Our work relates to several themes in the literature: *sound event recognition*, *audio captioning*, *audio-based retrieval*, *text-based video retrieval* and *text-domain audio retrieval*. We discuss each of these next.

Sound event recognition. The task of sound event recognition consists of matching audio information with a text label. There is a rich literature addressing sound event recognition and detection. Examples include detecting audio events associated with sports [Xiong et al. 2003], urban sounds [Aucouturier et al. 2007], and distinguishing vocal and nonvocal events [Atrey et al. 2006]. Research in

this area has been driven by challenges, such as DCASE [Stowell et al. 2015; Mesaros et al. 2017a], and by the collection of sound event datasets. These include TUT Acoustic scenes [Mesaros et al. 2017b], the domestic sounds dataset CHIME-Home [Foster et al. 2015], the environmental sound classification dataset ESC-50 [Piczak 2015], FSDKaggle [Fonseca et al. 2019], and AudioSet [Gemmeke et al. 2017]. Of relevance to our approach, a number of prior works have employed deep learning for audio comprehension [Kong et al. 2018; Yu et al. 2018; Kong et al. 2019; Ford et al. 2019; Kong et al. 2020] on the AudioSet dataset. Our work differs from theirs in that we focus instead on the task of retrieval with natural language queries, rather than audio recognition.

Audio captioning. Audio captioning consists of generating a natural language description for a sound [Drossos et al. 2017]. This requires a more detailed understanding of the sound than simply mapping the sound to a set of labels (sound event recognition). Recently, several audio captioning datasets have been introduced, such as Clotho [Drossos et al. 2020] which was used in the DCASE automated audio captioning challenge 2020 [*DCASE2020 Challenge Task 6: Automated Audio Captioning* 2020], Audio Caption [M. Wu et al. 2019], and AudioCaps [C. D. Kim et al. 2019]. Multiple works have addressed automatic audio captioning on the AudioCaps dataset [Koizumi et al. 2020; X. Xu et al. 2021a; Eren and Sert 2020]. In this work, we use the AudioCaps and Clotho datasets for cross-modal retrieval.

Audio-based retrieval. Multiple content-based audio retrieval frameworks, in particular query by example methods, leverage the similarity of sound features that represent different aspects of sounds (e.g. pitch, or loudness) [Foote 1997; Wold et al. 1996; Helén and Virtanen 2007; Lallemand et al. 2012]. More recently, [Manocha et al. 2018] use a siamese neural network framework to learn to encode semantically similar audio close together in the embedding space. [Q. Jin et al. 2012] address multimedia event detection using only audio data, while [Avgoustinakis et al. 2020] tackle near-duplicate video retrieval by audio retrieval. These are purely audio-based methods that are applied to video datasets, but without using visual information. [Hou and S. Zhou 2013] propose a two-step approach for video retrieval which uses audio (coarse) and visual (fine) information together.

Text-based video retrieval. More closely related to our work, a number of methods showed that embedding video and text jointly into a shared space (such that their similarity can be computed efficiently) is an effective approach [Dong et al. 2016; Mithun et al. 2018; Miech et al. 2018; Y. Liu et al. 2019; Wray et al. 2019; Gabeur et al. 2020]. One particular trend has been to combine cues from several "experts"—pre-trained models that specialise in different tasks (such as object recognition, action classification etc.) to inform the joint embedding. In this work we propose to adapt two such methods: the Mixture of Embedded Experts method of [Miech et al. 2018] and the Collaborative Experts model of [Y. Liu et al. 2019], by repurposing them for the task of audio retrieval (described in more detail in Sec. A.3).

Text-domain audio retrieval. Some methods retrieve audio by matching associated text, instead of matching the audio content directly. This has the implicit assumption that the associated text, such as meta-data or sound event labels, is relevant [Elizalde et al. 2018]. [Slaney 2002b] is an early work which links audio and text representations in hierarchical semantic and acoustic spaces. [Chechik et al. 2008] propose a text-based sound retrieval framework which uses single-word audio tags as queries rather than caption-like natural language. The creative approach of [Aytar et al. 2017] learns to align visual, audio, and text representations to enable cross-modal retrieval. Their framework is trained with captioned images and paired image-sound data (sourced from videos) and evaluated using the soundtrack of captioned videos. More recently, [Elizalde et al. 2019] use a siamese network to learn a shared latent text and sound space for cross-modal retrieval. While they use one or two class labels as text labels, we study unconstrained natural language descriptions as queries.

A.3 Methods, datasets and benchmarks

In this section, we first formulate the problem of audio retrieval with natural language queries. Next, we describe two cross-modal embedding methods, originally developed in the context of video retrieval, that we adapt for the task of audio retrieval. Finally, we describe the four datasets used in our experimental study, and the two benchmarks that we propose to evaluate performance on the audio

retrieval task.

Problem formulation. Given a natural language query (i.e. a written description of an audio event to be retrieved) and a pool of audio samples, the objective of text-to-audio (abbreviated to `t2a`) retrieval is to rank the audio samples according to their similarity to the query. We also consider the converse `a2t` task, viz. retrieving text with audio queries.

Methods. To tackle the problem of audio retrieval, we propose to learn cross-modal embeddings. Specifically, given a collection of N audio samples with corresponding text descriptions, $\{(a_i, t_i) : i \in \{1, \ldots, N\}\}$ we aim to learn a pair of embedding functions, ψ_a and ψ_t, that project each audio sample a_i and text sample t_i into a shared space, such that $\psi_a(a_i)$ and $\psi_t(t_i)$ are close when the text describes the audio, and far apart otherwise. Writing s_{ij} for the cosine similarity of the audio embedding $\psi_a(a_i)$ and the text embedding $\psi_t(t_j)$, we learn the embedding functions by minimising a batch-wise contrastive ranking loss [Socher et al. 2014]:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1, j \neq i}^{B} [m + s_{ij} - s_{ii}]_+ + [m + s_{ji} - s_{ii}]_+ \tag{A.1}$$

where B denotes the batch size, m the *margin* (set as a hyperparameter), and $[\cdot]_+ = max(\cdot, 0)$ the hinge function.

We consider two frameworks for learning such embedding functions ψ_a and ψ_t: Mixture-of-Embedded Experts (MoEE) [Miech et al. 2018] and Collaborative-Experts (CE) [Y. Liu et al. 2019]. MoEE, originally designed for video-text retrieval, constructs its video encoder from a collection of "experts" (features extracted from networks pre-trained for object recognition, action classification, sound classification, etc.) which are computed for each video, aggregated along their temporal dimension with NetVLAD [Arandjelovic et al. 2016], projected to a lower dimension via a self-gated linear map and then L2-normalised. The text encoder first embeds each word token with word2vec [Mikolov et al. 2013] and aggregates the results with NetVLAD [Arandjelovic et al. 2016]. The result is projected by a sequence of self-gated linear maps (one for each expert) into a shared embedding space with the outputs of the video encoder. Finally, a scalar-weighted sum of the embedded experts in each joint space is used to compute the

overall cosine similarity between the video and text (see [Miech et al. 2018] for more details). CE adopts the same text encoder as MoEE and similarly makes use of multiple video experts. However, rather than projecting them directly into independent spaces against the embedded text, CE first applies a "collaborative gating" mechanism, which filters each expert with an element-wise attention mask that is generated with a small MLP that ingests all pairwise combinations of experts (see [Y. Liu et al. 2019] for further details).

To adapt these frameworks for audio retrieval, we use the same text encoder structure ψ_t that they share, and build audio encoders ψ_a by mimicking the structure of their video encoders, replacing the "video experts" with "audio experts" (described in Sec. A.4).

Datasets. As the primary focus of our work, we study two *audio-centric datasets*—these are datasets which comprise audio streams (sometimes with accompanying visual streams) paired with natural language descriptions that focus explicitly on the content of the audio track. To explore differences between audio retrieval and video retrieval, we also consider two *visual-centric datasets*, that comprise audio and video streams paired with natural language which focus primarily (though not always exclusively) on the content of the video stream. Details of the four datasets we employ are given next.

1. AudioCaps [C. D. Kim et al. 2019] *(audio-centric)* is a dataset of sounds with event descriptions that was introduced for the task of audio captioning, with sounds sourced from the AudioSet dataset [Gemmeke et al. 2017]. Annotators were provided the audio tracks together with category hints (and with additional video hints if needed). We use a subset of the data, excluding a small number of samples for which either: (i) the YouTube-hosted source video is no longer available, (ii) the source video overlaps with the training partition of the VGGSound dataset [H. Chen et al. 2020]. Filtering to exclude samples affected by either issue leads to a dataset with 49,291 training, 428 validation and 816 test samples.[1]

2. Clotho [Drossos et al. 2020] *(audio-centric)* is a dataset of described sounds that was also introduced for the task of audio captioning, with sounds sourced from the Freesound platform [Font et al. 2013]. During labelling, annotators only

[1] We make the sample list publicly available at the project page [*Audio-Retrieval Project Page* n.d.].

Dataset	Anno.	Focus	Pool	Text ⟹ Audio R@1 ↑	Text ⟹ Audio R@10 ↑	Audio ⟹ Text R@1 ↑	Audio ⟹ Text R@10 ↑
AudioCaps [C. D. Kim et al. 2019]	audio		816	$18.0_{\pm 0.2}$	$62.0_{\pm 0.5}$	$21.0_{\pm 0.8}$	$62.7_{\pm 1.6}$
Clotho [Drossos et al. 2020]	audio		1045	$4.0_{\pm 0.2}$	$25.4_{\pm 0.5}$	$4.7_{\pm 0.3}$	$25.8_{\pm 1.7}$
ActNetCaps [Krishna et al. 2017]	visual		4917	$1.4_{\pm 0.1}$	$8.5_{\pm 0.2}$	$1.1_{\pm 0.1}$	$7.9_{\pm 0.0}$
QuerYD [Oncescu et al. 2021a]	visual		1954	$3.7_{\pm 0.2}$	$17.3_{\pm 0.6}$	$3.8_{\pm 0.2}$	$16.8_{\pm 0.2}$

Table A.1: **Audio retrieval on audio-centric and visual-centric datasets.** Performance is strongest on the audio-centric AudioCaps dataset and weakest on the visual-centric ActivityNet-Captions (ActNetCaps) dataset.

had access to the audio stream (i.e. no visual stream or meta tags) to avoid their reliance on contextual information for removing ambiguity that could not be resolved from the audio stream alone. The descriptions are filtered to exclude transcribed speech. The publicly available version of the dataset includes a `dev` set of 2893 audio samples and an `evaluation` set of 1045 audio samples, each with a duration of between 15 and 30 seconds. Every audio sample is accompanied by 5 written descriptions. We used a random split of the `dev` set into a training and validation set with 2,314 and 579 samples, respectively.

3. ActivityNet-Captions [Krishna et al. 2017] (visual-centric) consists of videos sourced from YouTube and annotated with dense event descriptions. It allocates 10,009 videos for training and 4,917 videos for testing (we use the public `val_1` split provided by [Krishna et al. 2017]). For this dataset, descriptions also tend to focus on the visual stream.

4. QuerYD [Oncescu et al. 2021a] (visual-centric) is a dataset of described videos sourced from YouTube and the YouDescribe [Research and Center 2013] platform. It is accompanied by *audio descriptions* that are provided with the explicit aim of conveying the video content to visually impaired users. Therefore, the provided descriptions focus heavily on the visual modality. We use the version of the dataset comprising trimmed videos with 9,113 training, 1,952 validation, and 1,954 test samples.

Benchmark. To facilitate the study of the text-based audio retrieval task, we propose to re-purpose the two *audio-centric* datasets described above, AudioCaps and Clotho, to provide benchmarks for text-based audio retrieval. This approach is inspired by precedents in the vision and language communities, where datasets, such as [J. Xu et al. 2016], that were originally introduced for the task of video captioning, have become popular benchmarks for text-based video retrieval [Miech et al. 2018; Wray et al. 2019; Y. Liu et al. 2019; Gabeur et al. 2020].

	Text ⟹ Audio/Video		Audio/Video ⟹ Text	
Expert	R@1 ↑	R@10 ↑	R@1 ↑	R@10 ↑
Visual experts only				
Scene	$6.1_{\pm 0.4}$	$35.8_{\pm 0.6}$	$6.5_{\pm 0.8}$	$31.3_{\pm 1.6}$
R2P1D	$8.2_{\pm 0.5}$	$44.7_{\pm 0.9}$	$10.2_{\pm 0.4}$	$41.8_{\pm 3.1}$
Inst	$7.7_{\pm 0.4}$	$46.7_{\pm 1.3}$	$9.8_{\pm 0.9}$	$40.6_{\pm 0.7}$
Scene + R2P1D	$8.8_{\pm 0.1}$	$46.8_{\pm 0.1}$	$11.1_{\pm 0.6}$	$45.1_{\pm 1.7}$
Scene + Inst	$8.7_{\pm 0.5}$	$47.4_{\pm 0.5}$	$10.6_{\pm 0.6}$	$41.4_{\pm 1.5}$
R2P1D + Inst (*CE-Visual*)	$\mathbf{10.1_{\pm 0.2}}$	$\mathbf{49.6_{\pm 1.1}}$	$\mathbf{12.2_{\pm 0.3}}$	$\mathbf{46.1_{\pm 1.3}}$
Audio experts only				
VGGish	$18.0_{\pm 0.2}$	$62.0_{\pm 0.5}$	$21.0_{\pm 0.8}$	$62.7_{\pm 1.6}$
VGGSound	$20.2_{\pm 0.5}$	$67.2_{\pm 1.2}$	$24.0_{\pm 0.6}$	$69.8_{\pm 0.4}$
VGGish + VGGSound (*CE-Audio*)	$\mathbf{23.1_{\pm 0.8}}$	$\mathbf{70.7_{\pm 0.7}}$	$\mathbf{25.0_{\pm 0.9}}$	$\mathbf{73.2_{\pm 1.6}}$
Audio and visual experts				
CE-Visual + VGGish	$23.9_{\pm 0.7}$	$74.4_{\pm 0.2}$	$29.0_{\pm 2.0}$	$77.2_{\pm 1.9}$
CE-Visual + VGGSound	$27.4_{\pm 0.7}$	$78.2_{\pm 0.3}$	$34.0_{\pm 1.5}$	$82.5_{\pm 1.2}$
CE-Visual + *CE-Audio*	$\mathbf{28.1_{\pm 0.6}}$	$\mathbf{79.0_{\pm 0.5}}$	$\mathbf{33.9_{\pm 1.6}}$	$\mathbf{83.7_{\pm 0.4}}$

Table A.2: **The influence of different experts on AudioCaps**. A comparison of audio and visual experts (applied to the video from which the audio was sourced) using CE [Y. Liu et al. 2019]. We observe that audio features are significantly more effective than visual features (which nevertheless provide some complementary signal as can be seen when jointly using audio and visual features).

A.4 Experiments

In this section, we first compare the text-based audio retrieval and audio-based text retrieval performances on audio-centric and visual-centric datasets. Next, we perform an ablation study on the contributions of different experts and present our baselines for the proposed AudioCaps and Clotho benchmarks. Finally we perform experiments to assess the influence of pre-training and training dataset size, and give qualitative examples of retrieval results. Throughout the section, we use the standard retrieval metrics R@k (recall at rank k, higher is better) and report the mean and standard deviation of three different randomly seeded runs.

Implementation details. To encode the audio signal, we use two pre-trained audio feature extractors. For each audio clip, we extract features using VGGish pre-trained for audio classification on YouTube8M [Hershey et al. 2016] and ResNet18 pre-trained for audio classification on the VGGSound dataset [H. Chen et al. 2020].[2] All audio retrieval models were trained with a batch size of 128 using the contrastive ranking loss (Eqn.(A.1.)), with the margin hyperparameter, m, set to 0.2. The learning rate was set to 0.01 with a weight decay of 0.001. Further

[2]Since the AudioCaps test set is a subset of the AudioSet training set (unbalanced), we do not use audio experts pre-trained on AudioSet.

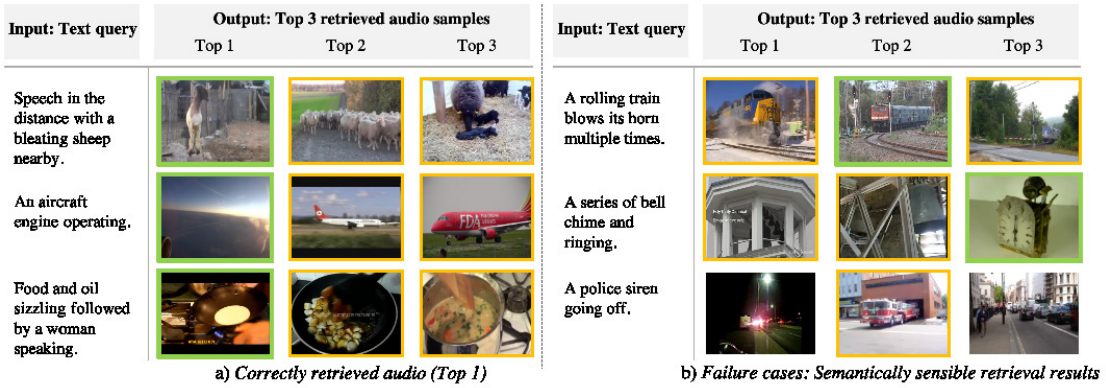


Figure A.1: **Qualitative results.** Text-based audio retrieval results on AudioCaps using CE with VGGish and VGGSound features. For an input text query, we visualise the top 3 retrieved audio samples using a video frame from the corresponding videos (the audio can be heard at the project webpage [*Audio-Retrieval Project Page* n.d.]). We mark the audio samples which correspond to the query with green boxes. Successful retrievals are shown in a), failures in b). Note, in particular, the examples in b), where the model's top 1 retrieved audio is not the correct one, but the retrieved results nevertheless sound reasonable (visually convincing results are marked with yellow boxes).

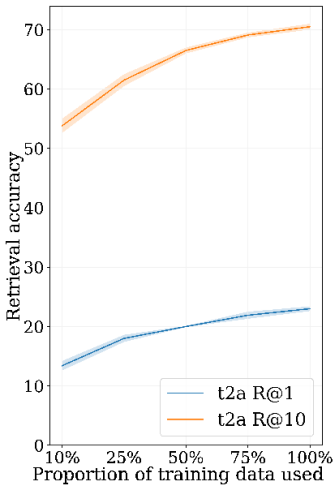


Figure A.2: **The influence of training data scale.** CE retrieval performance on AudioCaps with different proportions of the available AudioCaps training data.

details on optimisation, pre-trained experts (for audio and visual modalities) and hyperparameter choices are provided in the supplementary material.

Audio-centric vs. visual-centric queries. We first investigate audio retrieval with audio-centric and visual-centric queries. For this experiment, we use CE with a single expert (VGGish audio features). In Table **??**, we observe that performance is strongest on the audio-centric AudioCaps dataset and weakest overall on the ActivityNet-Captions dataset. We note that the Clotho dataset is particularly challenging, with performance weaker (accounting for pool size) than the visual-centric QuerYD dataset. We hypothesise that this is for two reasons: (1) the significantly smaller training set size of Clotho, compared to all other datasets, (2) Clotho was constructed by selecting audio samples that were as diverse as possible, making it a potentially more difficult benchmark. We note, however, that the QuerYD experiments suggest that computationally efficient video retrieval using only the audio stream can still be obtained, although at a lower accuracy.

Ablation study. We next conduct an ablation study to investigate the effectiveness of different experts for audio retrieval on the AudioCaps dataset. We present the results in Tab. A.2, where we observe that audio experts significantly outperform visual experts (pre-trained models for visual tasks like scene classification, which we compute from the video from which the audio was sourced). We note that the combination of audio and visual experts performs strongest overall, suggesting that the audio-centric queries contain information that is more accessible from the visual modality. We also observe that the strongest audio-only retrieval is achieved by combining VGGish and VGGSound features—we therefore adopt this combination for the remaining experiments.

Additionally, we experimented with using speech features (word2vec [Mikolov et al. 2013] encodings of speech-to-text transcriptions [Google n.d.] of the audio stream). However, this did not improve the retrieval performance. Upon further investigation, we found that the audio captions and the spoken words in the AudioCaps dataset do not have a significant overlap (corresponding to a METEOR [Banerjee and Lavie 2005] score of only 0.03).

Benchmark results. Incorporating the strongest combination of experts from the ablation study, we report our final baselines for text-audio and audio-text retrieval

		Text $\Longrightarrow$ Audio		Audio $\Longrightarrow$ Text	
Benchmark	Pre-training	R@1 ↑	R@10 ↑	R@1 ↑	R@10 ↑
AudioCaps					
MoEE	None	$22.5_{\pm 0.3}$	$69.5_{\pm 0.9}$	$25.1_{\pm 0.9}$	$72.9_{\pm 1.2}$
CE	None	$23.1_{\pm 0.8}$	$70.7_{\pm 0.7}$	$25.0_{\pm 0.9}$	$73.2_{\pm 1.6}$
Clotho					
MoEE	None	$6.0_{\pm 0.1}$	$32.3_{\pm 0.3}$	$7.3_{\pm 0.5}$	$33.2_{\pm 1.1}$
CE	None	$6.7_{\pm 0.4}$	$33.2_{\pm 0.3}$	$7.1_{\pm 0.3}$	$34.6_{\pm 0.5}$
Clotho					
MoEE	AudioCaps	$8.6_{\pm 0.4}$	$39.3_{\pm 0.7}$	$10.0_{\pm 0.3}$	$40.1_{\pm 1.3}$
CE	AudioCaps	$9.6_{\pm 0.3}$	$40.1_{\pm 0.7}$	$10.8_{\pm 0.7}$	$40.8_{\pm 1.4}$

Table A.3: **Audio retrieval benchmarks.** Text-audio and audio-text retrieval results for MoEE [Miech et al. 2018] and CE [Y. Liu et al. 2019] with VGGish and VGGSound features on the proposed AudioCaps and Clotho retrieval benchmarks. Pre-training on AudioCaps improves the performance on Clotho.

for two methods on the AudioCaps and Clotho datasets in Tab. A.3, where we observe that CE slightly outperforms MoEE. We also report the performance after pre-training models for retrieval on AudioCaps and then fine-tuning on Clotho (lower part of Tab. A.3). Here, we observe that pre-training on the audio-centric AudioCaps brings a boost to both, CE and MoEE. Furthermore, we explored pre-training on Clotho and fine-tuning on the AudioCaps dataset, but found negligible change in performance (likely due to the fact that the AudioCaps training set is significantly larger than that of Clotho).

Qualitative results. The qualitative results in Figure A.1 show examples in which the CE model with VGGish and VGGSound expert modalities (CE-Audio) is used to retrieve audio with natural language queries. The retrieved results mostly contain audio that is semantically similar to the input text queries. Observed failure cases arise from audio samples sounding very similar to one another despite being semantically distinct (e.g. the siren of a fire engine sounds very similar to a police siren).

Influence of scale. Finally, we present experiments using different proportions of the AudioCaps dataset for training CE-Audio in Fig. A.2. We observe that as more training data becomes available, retrieval performance increases monotonically. We also observe that there is still clear room for improvement in terms of retrieval results simply by collecting additional training data, motivating further dataset

construction work to support future research on this important task.

A.5 Conclusion

We introduced a new benchmark for natural language based audio retrieval, and provided baseline results by adapting strong multi-modal video retrieval methods. Our results show that these methods are relatively well-suited for the audio retrieval task, however there is room for improvement, as expected for an under-explored problem. We hope that our proposed benchmarks will facilitate the development of future audio search engines, and make this large fraction of the world's produced media available for public use.

Acknowledgements. AMO was supported by an EPSRC DTA Studentship. ASK and ZA were supported by the ERC (853489 - DEXIM) and by the DFG (2064/1 – Project number 390727645). JFH is supported by the Royal Academy of Engineering (RF\201819\18\163). SA was supported by EPSRC EP/T028572/1 Visual AI.

A.6 Supplementary material

We give details on the pre-trained experts used, and on optimisation and hyper-parameter choices below.

A.6.1 Pre-trained experts

This section contains information about the audio and visual expert features that were obtained from pre-trained networks.

Audio experts

We use two different audio experts, which we abbreviate as VGGish and VGGSound.

VGGish. Theses audio features are obtained with a VGGish model [Hershey et al. 2016], trained for audio classification on the YouTube-8M dataset [Abu-El-Haija et al. 2016]. To produce the input for this model, the audio stream of each video is re-sampled to a 16kHz mono signal, converted to an STFT with a window size of 25ms and a hop size of 10ms with a Hann window, then mapped to a 64 bin log mel spectrogram. Finally, the features are parsed into non-overlapping 0.96s collections of frames (each collection comprises 96 frames, each of 10ms duration), which is mapped to a 128-dimensional feature vector.

VGGSound. These features are extracted using a ResNet-18 model [He et al. 2016] that has been pre-trained on the VGGSound dataset (model H) [H. Chen et al. 2020]. We modify the last average pooling layer to aggregate along the frequency dimension, but keep the full temporal dimension. This results in features of dimension t x 512, where t denotes the number of time steps.

Visual experts

We use three different visual experts, abbreviated as Inst, Scene, and R2P1D.

Inst. These features are extracted using a ResNeXt-101 model [R. Xu et al. 2015] that has been pre-trained on Instagram hashtags [Mahajan et al. 2018] and fine-tuned on ImageNet [Deng et al. 2009] for the task of image classification. Features are extracted from frames extracted at 25 fps, where each frame is resized to 224 $\times$ 224 pixels. Embeddings are 2048-dimensional.

			Text ⟹ Audio		Audio ⟹ Text	
Model	Zero-pad	Pre-training	R@1 ↑	R@10 ↑	R@1 ↑	R@10 ↑
MoEE	z^0	None	$5.1_{\pm 0.6}$	$30.1_{\pm 1.6}$	$6.3_{\pm 0.7}$	$29.9_{\pm 2.1}$
CE	z^0	None	$5.8_{\pm 0.2}$	$31.3_{\pm 0.8}$	$7.3_{\pm 0.5}$	$32.8_{\pm 1.0}$
MoEE	z^1	None	$6.0_{\pm 0.1}$	$32.3_{\pm 0.2}$	$7.2_{\pm 0.5}$	$33.3_{\pm 1.0}$
CE	z^1	None	$6.4_{\pm 0.2}$	$33.1_{\pm 0.4}$	$7.2_{\pm 0.1}$	$34.6_{\pm 0.7}$
MoEE	z^0	AudioCaps	$8.4_{\pm 0.2}$	$37.3_{\pm 0.9}$	$9.7_{\pm 0.9}$	$37.7_{\pm 0.5}$
CE	z^0	AudioCaps	$7.9_{\pm 0.4}$	$35.9_{\pm 0.7}$	$9.1_{\pm 0.1}$	$36.1_{\pm 0.5}$
MoEE	z^1	AudioCaps	$9.0_{\pm 0.4}$	$39.4_{\pm 1.1}$	$10.2_{\pm 0.9}$	$40.2_{\pm 1.6}$
CE	z^1	AudioCaps	$9.5_{\pm 0.4}$	$40.3_{\pm 1.3}$	$10.6_{\pm 0.7}$	$41.7_{\pm 0.9}$

Table A.4: **Varying zero-padding of expert features.** Text-audio and audio-text retrieval results for different choices of zero-padding hyperparameters z^i for $i \in \{0, 1\}$ for MoEE [Miech et al. 2018] and CE [Y. Liu et al. 2019] with VGGish and VGGSound features on the Clotho retrieval benchmark.

Scene. Scene features are extracted from 224×224 pixel centre crops with a DenseNet-161 model [G. Huang et al. 2017] pre-trained on Places365 [B. Zhou et al. 2017]. Embeddings are 2208-dimensional.

R2P1D. Features are extracted with a 34-layer R(2+1)D model [Tran et al. 2018] trained on IG-65M [Ghadiyaram et al. 2019] which processes clips of 8 consecutive 112 × 112 pixel frames, extracted at 30 fps. Embeddings are 512-dimensional.

A.6.2 Optimisation details and hyperparameter choices

Optimiser

We used the Lookahead solver [M. R. Zhang et al. 2019] in combination with RAdam [L. Liu et al. 2019] (implementation by [*Ranger optimiser* n.d.]) with an intial learning rate of 0.01 and weight decay of 0.001. We use a learning rate decay for each parameter group with a factor of 0.95 every epoch. All models were trained for a maximum of 20 epochs. We chose the models that gave the best performance on the geometric mean of R@1, R@5, and R@10.

NetVLAD hyperparameters

For text, we used 20 VLAD clusters and one ghost cluster. For the VGGish and VGGSound features we used 16 VLAD clusters.

Zero-padding hyperparameters

In the main paper, we used zero-padding of variable length experts to a common shape $z^0 = (z_t^0, z_a^0, z_v^0)$, with $z_t^0 = 20$ for text tokens, $z_a^0 = 29$ for VGGish feature tokens, and $z_v^0 = 29$ for the VGGSound features.

We include an ablation experiment for the Clotho dataset using zero-padding parameters $z^1 = (z_t^1, z_a^1, z_v^1)$ with $z_t^1 = 21$, $z_a^1 = 31$, and $z_v^1 = 95$ for text, VGGish features, and VGGSound features, respectively. This choice of hyperparameters corresponds more closely to the distribution of word and audio feature shapes for the Clotho dataset. With this configuration the results improve when training with both VGGish and VGGSound features as presented in Table A.4.

Appendix B

Statement of Authorship

A statement of authorship is provided for each multi-authored paper included in this thesis. The statements describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication, there exists a complete statement that is filled out and signed by the candidate and supervisor.

Statement of Authorship for the paper "QuerYD: A Video Dataset with High-Quality Text and Audio Narrations" in Chapter 2.

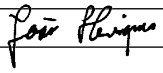
Paper title	QuerYD: A Video Dataset with High-Quality Text and Audio Narrations
Authors	Andreea-Maria Oncescu, João F. Henriques, Yang Liu, Andrew Zisserman, Samuel Albanie
Publication status	Published
Publication details	International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2021.

Student Confirmation

Student name	Andreea-Maria Oncescu	
Contribution to the paper	First-author contribution: • implementation of models • data collection and curation • writing and presentation of the paper	
Signature and Date		Apr. 8th 2024

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	Dr. João F. Henriques	
Supervisor comments		
Signature and Date		Apr. 8th 2024

Statement of Authorship for the paper "Audio Retrieval with Natural Language Queries: A Benchmark Study" in Chapter 3.

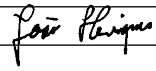
Paper title	Audio Retrieval with Natural Language Queries: A Benchmark Study
Authors	A. Sophia Koepke∗, Andreea-Maria Oncescu∗, João F. Henriques, Zeynep Akata, Samuel Albanie
Publication status	Published
Publication details	IEEE Transactions on Multimedia, 2022.

Student Confirmation

Student name	Andreea-Maria Oncescu	
Contribution to the paper	First-author contribution: • implementation of models • data collection and curation • writing of the paper	
Signature and Date		Apr. 8th 2024

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	Dr. João F. Henriques	
Supervisor comments		
Signature and Date		Apr. 8th 2024

Statement of Authorship for the paper "Dissecting Temporal Understanding in Text-Audio Retrieval" in Chapter 4.

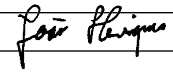
Paper title	Dissecting Temporal Understanding in Text-Audio Retrieval
Authors	A. Sophia Koepke∗, Andreea-Maria Oncescu∗, João F. Henriques, Zeynep Akata, Samuel Albanie
Publication status	Under submission
Publication details	Association for Computing Machinery's International Conference on Multimedia, 2024

Student Confirmation

Student name	Andreea-Maria Oncescu	
Contribution to the paper	First-author contribution: • conception of research ideas • data generation • running experiments • writing of the paper	
Signature and Date		Apr. 8th 2024

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	Dr. João F. Henriques	
Supervisor comments		
Signature and Date		Apr. 8th 2024

Statement of Authorship for the paper "A sound approach: Using large language models to generate audio descriptions for egocentric text-audio retrieval" in Chapter 5.

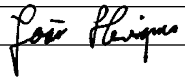
Paper title	A sound approach: Using large language models to generate audio descriptions for egocentric text-audio retrieval
Authors	Andreea-Maria Oncescu, João F. Henriques, Andrew Zisserman, Samuel Albanie, A. Sophia Koepke
Publication status	Published
Publication details	International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2024.

Student Confirmation

Student name	Andreea-Maria Oncescu	
Contribution to the paper	First-author contribution: • conception of research ideas • implementation of models • data collection and curation • writing and presentation of the paper	
Signature and Date		Apr. 8th 2024

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	Dr. João F. Henriques	
Supervisor comments		
Signature and Date		Apr. 8th 2024