Towards Better Video Understanding through Language Guidance

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Thomas Hummel
aus Landshut

Tübingen
2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	23.06.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatterin:	Prof. Dr. Zeynep Akata
2. Berichterstatter:	Prof. Dr. Hendrik Lensch

imprs-is

machine learning
new perspectives
for science

Department of
Computer Science

Towards Better Video Understanding through Language Guidance

Thomas Hummel

Master of Science in Intelligent Adaptive Systems

Adviser: Zeynep Akata
Full Professor, Technical University of Munich

Co-adviser: Hendrik Lensch
Full Professor, University of Tübingen

Examination Committee

Chair: Martin Butz
Full Professor, University of Tübingen

Members: Zeynep Akata
Full Professor, Technical University of Munich

Hendrik Lensch
Full Professor, University of Tübingen

Seong Joon Oh
Full Professor, University of Tübingen

Dissertation for the award of the degree "*Doctor rerum naturalium*" in Computer Science.

University of Tübingen and International Max Planck Research School for Intelligent Systems
2025

Acknowledgements

Climbing a mountain, especially one as daunting as the pursuit of a doctoral degree is never a solitary endeavour. While this thesis represents the summit of that ascent, the journey would have been impossible without the support, guidance, and encouragement of many.

Foremost, I want to thank my PhD supervisor Zeynep Akata for her unwavering support throughout this doctoral journey and the freedom she granted me to pursue my research interests. I am deeply grateful for her mentorship and the opportunities she provided. In addition, I want to give my gratitude to Hendrik Lensch for his invaluable support during my PhD, as well as to my other committee members, Seong Joon Oh, Martin Butz, and Samira Samadi. I also want to thank the University of Tübingen and the International Max Planck Research School for Intelligent Systems for providing such a stimulating, open, and genuinely supportive research environment.

I am also grateful to the past and current EML group members. You provided the essential fuel for my journey. The EML group has been such a stimulating environment for research, and every group member was ready to take a short hike and discuss anything from technical details to philosophizing about the next breakthrough. From the scientific discussions over coffee to mandatory foosball battles, cherished Thursday traditions, and our collective travels. Thank you for creating this amazing environment that went beyond academics. To my co-authors Stephan, Almut Sophia, Otniel, Shyam, and Lili. Your collaboration, insights, and dedication were instrumental in shaping this research. I learned so much from each of you, and I am truly grateful for having the opportunity to work with you.

A climbing expedition would not be possible without a support network. The list is endless, but let me mention a few highlights. Thank you, Almut Sophia, for your mentorship during my PhD, for helping me to become a better researcher, and for guiding me to follow the correct trail. Thank you, Stephan and Massi, for motivating me with your enthusiasm for science and for sharing the passion for Mate, Mario Kart, and random competitions. Thank you, Otniel and Shyam, for always cheering me up with your humour and for discussing every random scientific and non-scientific thought. Thank you, Leonard,

for our downhill walks, and thank you, Jae Myung, for enriching the summer school and for your excellent tennis teaching skills.

Finally, to my friends at home, thank you for always grounding me whenever I had the time to come home. To my family, thank you for your endless support. Particularly to my parents for your incredible patience and for being there for me throughout my studies and PhD – without you, this journey would not have been possible and I cannot thank you enough. To my brother, grandparents, aunt and cousins – Danke! Finally, to Claudia, thank you for your infinite encouragements, for always turning me around when I doubted myself, and for always believing in me.

Abstract

Video understanding is a crucial area of computer vision, with applications ranging from autonomous driving and robotics to multimedia interaction. Despite significant progress in image analysis, video understanding remains a complex problem due to the temporal nature of videos, requiring models to analyse both individual frames and their relationships over time. This work explores how various forms of language, ranging from class labels to natural language instructions, can be leveraged to overcome these challenges and to improve model capabilities and generalisation. Moreover, through novel settings, benchmarks, and frameworks, this work explores how integrating language with visual information can address key challenges in video understanding.

First, we explore audio-visual video classification in low-data regimes to address the limitations of traditional supervised learning. In audio-visual generalised zero-shot learning, class labels represented as pre-trained word embeddings serve as a semantic bridge, enabling models to classify unseen video classes by aligning audio-visual features with textual representations in a shared embedding space. Our Temporal and Cross-Attention Framework (TCaF) improves the alignment and, consequently, generalisation by better modelling temporal relationships and cross-modal interactions.

Next, this setting is extended to audio-visual generalised few-shot learning, where models must learn to classify new video classes with only a few labelled examples. In addition to protocols and benchmarks, we propose AV-Diff, which uses class-label text representations to guide a diffusion model to generate synthetic training samples, thereby enhancing model generalisation for novel video classes.

Beyond classification, this thesis explores fine-grained action understanding through video-adverb retrieval. This task extends traditional action recognition by incorporating adverbs to provide richer information about actions. By learning compositional embeddings that combine actions and adverbs, our proposed model achieves a more nuanced understanding of video content.

Finally, this thesis tackles composed video retrieval (CVR), a task where natural language instructions modify a reference video query to retrieve semantically altered videos. To successfully solve this task, a model requires compositional reasoning to interpret

both video content and the transformative effect of textual instructions. We propose the egocentric evaluation benchmark EgoCVR, which tests the fine-grained temporal video understanding capabilities of vision-language models. Furthermore, we present TFR-CVR, a modular and training-free framework that achieves improved temporal reasoning by strategically utilising the reasoning abilities of large language models.

By integrating language at different levels – from class labels to fine-grained action modifications and natural language instructions – the work presented pushes beyond traditional video classification towards more robust, flexible, and fine-grained video understanding capabilities.

Zusammenfassung

Videoverständnis ist ein wichtiger Bereich in Computer Vision, mit Anwendungen, die von autonomem Fahren und Robotik bis hin zur Multimedia-Interaktion reichen. Trotz bedeutender Fortschritte in der Bildanalyse bleibt das Videoverständnis aufgrund der zeitlichen Natur von Videos ein komplexes Problem, das von Modellen erfordert, sowohl einzelne Frames als auch ihre Beziehungen über die Zeit zu analysieren. Diese Arbeit untersucht, wie verschiedene Formen von Sprache, von Klassenbezeichnungen bis hin zu Instruktionen in natürlicher Sprache, genutzt werden können, um diese Herausforderungen zu überwinden und die Modellfähigkeiten und Generalisierung zu verbessern. Durch neuartige Aufgaben, Benchmarks und Frameworks untersucht diese Arbeit, wie die Integration von Sprache mit visuellen Informationen wichtige Herausforderungen in der Videoanalyse angehen kann.

Zuerst untersuchen wir die audio-visuelle Videoklassifizierung in Low-Data-Regimen und gehen auf die Einschränkungen des traditionellen überwachten Lernens ein. Im audio-visuellen generalisierten Zero-Shot-Lernen dienen als vortrainierte Wortvektoren dargestellte Klassenbezeichnungen als semantische Brücke. Diese ermöglichen es Modellen, ungesehene Videoklassen zu klassifizieren, indem audio-visuelle Merkmale mit textuellen Darstellungen in einem geteiltem Projektsraum ausgerichtet werden. Unser Temporal and Cross-Attention Framework (TCAF) verbessert die Ausrichtung und folglich die Generalisierung, indem es zeitliche Beziehungen und zwischenmodale Interaktionen besser modelliert.

Als nächstes wird dieser Ansatz auf das audio-visuelle, generalisierte Few-Shot-Lernen erweitert. Bei diesem Ansatz müssen Modelle lernen, neue Videoklassen mit nur wenigen gelabelten Beispielen zu klassifizieren. Zusätzlich zu Protokollen und Benchmarks schlagen wir AV-Diff vor, das Textdarstellungen von Klassenbezeichnungen verwendet, um ein Diffusionsmodell zur Generierung synthetischer Trainingsbeispiele zu leiten und so die Modellgeneralisierung für neuartige Videoklassen zu verbessern.

Über die Klassifizierung hinaus untersucht diese Dissertation das feinkörnige Aktionsverständnis durch Video-Adverb-Retrieval. Diese Aufgabe erweitert die traditionelle

Aktionserkennung durch die Einbeziehung von Adverbien, um genauere Details über Aktionen zu erlangen. Durch das Erlernen kompositioneller Einbettungen, die Aktionen und Adverbien explizit kombinieren, erreicht unser vorgeschlagenes Modell ein nuancierteres Verständnis von Videoinhalten.

Schließlich befasst sich diese Dissertation mit Composed Video-Retrieval (CVR), einer Aufgabe, bei der Anweisungen aus natürlicher Sprache eine Referenzvideoabfrage modifizieren, um semantisch veränderte Videos abzurufen. Um diese Aufgabe erfolgreich zu lösen, ist kompositionelles Denken erforderlich, um sowohl Videoinhalte als auch die transformative Wirkung textueller Anweisungen zu interpretieren. Wir schlagen den egozentrischen Evaluierungs-Benchmark `EgoCVR` vor, der die feinkörnigen zeitlichen Videoverständnisfähigkeiten von Bild-Sprach-Modellen herausfordert. Darüber hinaus präsentieren wir `TFR-CVR`, ein modulares und trainingsfreies Framework, das durch die strategische Nutzung der Fähigkeiten großer Sprachmodelle eine verbesserte zeitliche Argumentation erreicht.

Durch die Integration von Sprache auf verschiedenen Ebenen – von Klassenbezeichnungen über feinkörnige Aktionsmodifikationen bis hin zu natürlichsprachlichen Anweisungen – geht die vorgestellte Arbeit über die traditionelle Videoklassifizierung hinaus zu einem robusteren, flexibleren und feinkörnigeren Videoverständnis.

Contents

List of Figures

List of Tables

1 Introduction

With the rapid rise of digital media, videos have become one of the most dominant forms of content across various domains. Online platforms receive uploads of vast amounts of videos daily. Unlike static images, video data is rich in visual, audio, and metadata, offering deeper insights but also presenting challenges in processing and interpretation. The applications of video understanding are vast, spanning multiple important areas such as action recognition [Cab+15; CZ17], video captioning [Seo+22; Yan+23], human-computer interaction [Gra+22; Gra+24], and autonomous driving [Xu+17]. Given the growing volume and importance of videos, developing intelligent systems that effectively and efficiently analyse and interpret videos is crucial.

This thesis aims to improve video understanding by leveraging language in various forms as guidance. In the introductory chapter, Section 1.1 highlights the importance of video understanding. Section 1.2 introduces the concept of language guidance and its benefits for video understanding and highlights the forms of language guidance present in each task tackled in this thesis. Section 1.3 highlights the main contributions of this thesis, which cover three different directions: advancing multi-modal learning, advancing compositional understanding, and establishing benchmarks for improving video understanding. Section 1.4 provides an outline for each chapter in this thesis.

1.1 Video Understanding: From Images to Videos

Artificial intelligence (AI) has experienced remarkable progress in recent years. A key driver of this progress has been the advancement in computer vision, enabling machine learning systems to perceive and interpret the visual world. Video understanding is a fundamental challenge in computer vision. The ability to automatically analyse and interpret video content is crucial for building systems that can interact with and understand the dynamic visual world around them. For example, video understanding is essential for developing safe and reliable autonomous driving systems [Xu+17], enhancing the capabilities of robotics in human-robot interaction [Ser+24], and powering more natural multimedia interactions by summarizing video content [Seo+22], searching for moments

in a video [Ann+17], or reasoning about its content [Li+24].

Despite its importance, progress in video understanding often lags behind the relative progress in image understanding. This is primarily due to the complexities associated with video data. While images are static snapshots, videos are inherently temporal. This temporal dimension requires models to analyse the content of individual frames and understand the relationships between frames, including actions, events, and the evolution of entities and scenes over time. Furthermore, the large volume of video data leads to substantially higher computational costs and training data requirements.

The evolution of video understanding research [Ngu+24; Tan+23] mirrors the broader development of image-based computer vision, often with a delay due to these inherent complexities. Early efforts in video analysis were heavily inspired by image processing techniques, focusing on frame-by-frame analysis without explicitly modelling temporal dynamics [Lap05]. Later, traditional machine learning methods like support vector machines were combined with more sophisticated methods developed for capturing temporal information like optical flow [DG07; Sid04] and hand-crafted features like dense trajectories [WS13] for action recognition or event detection. Progress during this period was significantly driven by the availability of benchmark datasets like KTH [SLC04], HMDB51 [Kue+11], and UCF101 [SZS12].

The development of deep learning in the 2010s, and in particular Convolutional Neural Networks (CNNs), signalled a significant shift [Ngu+24; Tan+23]. Directly applying two-dimensional (2D) CNNs to videos proved insufficient, as they could not model temporal dependencies effectively. Advancements in video understanding were limited mainly by architectural shortcomings. To tackle this, spatiotemporal architectures were developed, such as two-stream networks [SZ14] that separately process spatial and motion information, recurrent neural networks on 2D CNN features [Don+15], and three-dimensional (3D) convolutional networks like C3D [Tra+15], I3D [CZ17], and S3D [Xie+18] that extend traditional CNNs to capture temporal patterns. These approaches were, however, still limited in their ability to efficiently capture long-time dependencies in videos. The introduction of attention mechanisms and transformers, such as the vision transformer (ViT) [Dos+21], and video transformer models like ViViT [Arn+21] and TimeSformer [BWT21] significantly improved the modelling of long-range interactions. These models pushed the state-of-the-art in video recognition, often benefiting significantly from training on large-scale video datasets like Sports-1M [Kar+14] and later Kinetics [Kay+17].

Recently, the field has shifted towards large-scale pre-training on often weakly labelled or unlabelled video datasets using self-supervised and multi-modal learning approaches [Ngu+24; Tan+23]. This shift leverages readily available, unstructured text data like descriptions and captions alongside videos for large-scale pre-training. Consequently, language plays an increasingly important role, enabling models to perform complex tasks like video question answering [Yan+23] and video-text retrieval [Bai+21], moving beyond simple action recognition towards deeper comprehension.

This thesis aims to enhance video understanding with a particular focus on video

classification and fine-grained retrieval tasks. The presented research investigates widely available web and egocentric videos, which offer unique perspectives and challenges. A key aspect of this work involves exploring the relationship between visual information and language, building upon the recent advancements in multi-modal learning. Towards this end, this thesis introduces novel settings, benchmarks, and frameworks designed to advance video understanding.

1.2 Language as a Powerful Paradigm for Video Understanding

One of the most significant advancements in recent years has been the integration of language with vision models [Ngu+24; Tan+23]. Prior to the shift to language-centric, large-scale pre-training, computer vision research has largely focused on distinct tasks such as classification or detection, as described above. Each task typically required specialised model architectures and training procedures and often struggled to generalise beyond their intended purpose. Instead of treating computer vision tasks as isolated problems, language offers a universal interface for representing and reasoning about visual content and other modalities.

For instance, the output of image classification can be represented as the category name, while for object detection as bounding box coordinates paired with class labels. This language-centric approach allows for a more flexible and generalisable way of modelling visual understanding. By learning an alignment of visual and textual representations, vision-language models (VLMs) like CLIP [Rad+21], BLIP [Li+22], or SigLIP [Zha+23a] enable open-world recognition and retrieval, showing the potential of language to guide visual understanding.

This paradigm shift is particularly valuable for video understanding, where semantic ambiguity and the need for temporal reasoning present significant challenges. The success of image-language pre-training motivated efforts to extend these alignment techniques to the temporal domain of video. In recent years, the research focused on developing architectures and pre-training strategies specifically designed for joint video-language understanding. Models like VideoCLIP [Xu+21] and InternVideo [Wan+22] are pre-trained on a mixture of video datasets with associated text, like HowTo100M [Mie+19] and WebVid [Bai+21], and can be fine-tuned for more specific video understanding tasks. These video-language models are often combined with pre-trained large language models (LLMs) to leverage their advanced reasoning and generation capabilities across various modalities [Che+23; Li+23; Wan+24].

The development of these dedicated video-language models advanced the state-of-the-art across a wide range of tasks. Language now serves as a powerful mechanism for guiding video understanding, offering a structured and expressive way to describe visual content. It can capture complex relationships and semantics, resolve visual ambiguities, and enable intuitive interaction with video data. Moving beyond simple classification with fixed categories, this paradigm enabled or improved many open-world tasks such as video-text

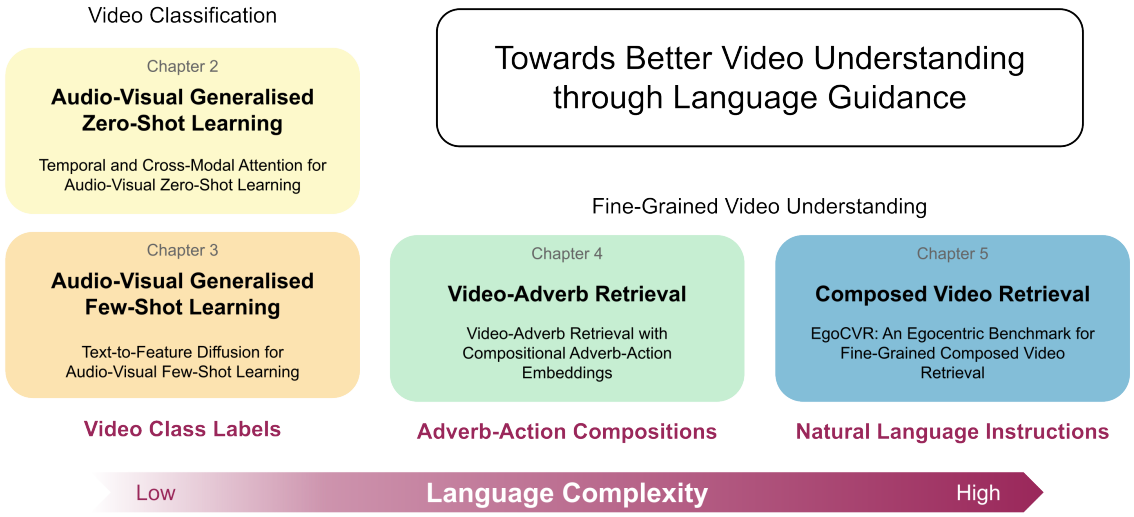

Figure 1.1: Overview of the thesis progression, the tasks tackled in each chapter, the title of the published work, and their level of language complexity. Chapter 2 has been published at ECCV 2022 [Mer+22a], Chapter 3 at DAGM GCPR 2023 [Mer+23], Chapter 4 at BMVC 2023 [Hum+23], and Chapter 5 at ECCV 2024 [Hum+24].

retrieval [Bai+21], video question answering [Lei+21; SNS21], video captioning [Seo+22; Yan+23], temporal action localization [Nag+22], and video grounding [Lin+23b]. Language provides the flexibility to move beyond classification schemes towards deeper comprehension, allowing models to generalise more effectively to unseen concepts and tasks.

1.2.1 The Role of Language in This Thesis

Building upon existing advancements and training paradigms, this thesis explores the role of language in enhancing video understanding. More specifically, it explores the use of: i) semantic class labels for video classification in low-shot learning scenarios, ii) adverb-action compositions for video-adverb retrieval, and iii) natural language instructions for composed video retrieval. The works presented follow a natural progression, employing language guidance of increasing complexity. Figure 1.1 highlights this progression. Initially, Chapters 2 and 3 focus on video classification, leveraging semantic information derived from class labels to enable audio-visual generalised zero-shot learning (Chapter 2) and to augment training for audio-visual generalised few-shot learning (Chapter 3). In Chapters 4 and 5, the focus shifts from coarse towards more fine-grained video understanding. Chapter 4 investigates fine-grained video-adverb retrieval using textual adverb-action compositions to guide visual representations, while Chapter 5 tackles composed video retrieval, utilising natural language instructions to steer the fine-grained video retrieval processes. These tasks, their motivation, and the specific uses of language designed to tackle them are described in more detail in the following sections.

1.2.2 Semantic Class Labels for Audio-Visual Low-Shot Learning

Learning from videos often benefits significantly from a multi-modal approach, as videos accompanied by audio are inherently multi-modal. Relying solely on visual information can lead to incomplete or ambiguous interpretations. The information present in the audio modality complements and extends the information present in the visual modality. For example, consider a video clip showing a close-up of a violinist's hands moving across the strings of a violin. Without audio, it might be challenging to assess if the violinist is the sole performer, is accompanied by a full orchestra, or is even just miming. Ambiguities in one modality can often be resolved by considering information from other available modalities, underscoring the importance of a multi-modal approach.

Additionally, traditional supervised learning approaches are severely limited by the scope of their labelled training data. A crucial capability of practical video classification systems, however, is the ability to generalise to novel scenarios not encountered during training. To address this limitation and to leverage the multi-modal nature of video, this thesis explores the use of class labels as a form of semantic guidance for zero- and few-shot learning. Chapter 2 explores the task of generalised audio-visual zero-shot learning, first introduced by Mercea et al. [Mer+22b], while Chapter 3 extends this setting to generalised audio-visual few-shot learning. Both tasks and their use of language guidance are highlighted in the following.

Audio-Visual Generalised Zero-Shot Learning. This setting addresses the video classification problem for both seen and unseen classes. In this setting, a model is trained on a set of seen classes (e.g. *playing the violin*). At inference time, however, the model is tasked to classify both seen and unseen classes, i.e. classes for which no video examples were given during training. To enable this generalisation to unseen classes, text representations in the form of pre-trained class-label word embeddings are used as a semantic bridge between seen and unseen domains. Instead of classification, this problem is approached as a retrieval task. During training, the audio-visual information is projected into a shared embedding space together with the class label embeddings. By learning the relationship between audio-visual features and their corresponding semantic text representations for the seen classes, the model implicitly learns a mapping that can generalise to unseen classes. During inference, we can classify videos from unseen classes based on their distance to class labels of seen and unseen classes in the learned shared embedding space.

Audio-Visual Generalised Few-Shot Learning. This setting relaxes the assumption of having no video examples available for novel classes to a more realistic scenario of having only a few labelled video examples available, usually ranging from 1 (*1-shot*) to 20 (*20-shot*) examples per novel class. One of the main challenges of this problem lies in the imbalance in the number of labelled training examples for regular classes and these novel classes. When training the model on novel classes, the model has to retain the knowledge of the seen classes while accurately classifying new classes from only a handful of video examples. This problem is solved as a classification problem and traditionally does not

require any form of language guidance. In this work, however, we use the semantic information in class labels to augment the training. More specifically, a generative model conditioned on the textual class labels generates synthetic features to augment and balance the classifier's training for novel classes. Language is not used here to align audio-visual embeddings but to contribute to creating new synthetic data points that capture some of the characteristics of novel classes. This allows the model to overcome the lack of real data to learn a robust classifier effectively.

1.2.3 Adverb-Action Compositions for Video-Adverb Retrieval

While class labels provide a coarse form of language guidance for video understanding, many real-world applications require a more fine-grained understanding of video content. For instance, the class label *driving* does not specify how the driving action is performed. Is the depicted person driving *slowly*, *quickly*, or *erratically*? This level of detail could be vital for applications like autonomous driving or human-robot interaction. In these contexts, such a system needs to understand and react to the actions of pedestrians and other road users or for a robot to interpret fine distinctions in human movement. Chapter 4 tackles the task of video-adverb retrieval and moves beyond simple action recognition to understanding the interplay between actions and adverbs, improving fine-grained video understanding.

Video-Adverb Retrieval. This task focuses on retrieving the adverb that best describes a given action within a video, and vice versa, retrieving videos that show an action being performed in a particular way. This is achieved by aligning video embeddings with text representations that combine the meaning of actions and adverbs. Compared to the previously discussed settings, the proposed model extends the employed language guidance beyond simple class labels. First, by conditioning the video representation on a given textual representation of an action using an attention mechanism, the model is guided to focus on the temporal segments of an action. Second, by learning compositional text embeddings that explicitly combine action and adverb information, the model is guided to distinguish between modifications of the action introduced by various adverbs. This moves the focus from coarse action classification to a fine-grained understanding of actions.

1.2.4 Natural Language Instructions for Composed Video Retrieval

Adverbs that modify actions provide a finer level of detail than class labels, but they still operate within a limited vocabulary. Real-world video understanding often requires the processing of more complex language or instructions beyond single-word modifiers. Natural language provides a far richer and more flexible mechanism for specifying the desired video content. Instructions when interacting with video content may incorporate changes in object interactions (e.g. *search a similar video, in which a child kicks the ball instead of an adult*), temporal relationships (e.g. *show the part before the car stops*), or spatial

configurations (e.g. *focus on the object to the left of the table*). Consider a robot tasked with following instructions in a home environment or a content-based video search system allowing users to refine their search with natural language. Such a system must understand the complex interplay between video content and the modifying effects of natural language instructions. Chapter 5 addresses these challenges using the task of composed video retrieval.

Composed Video Retrieval. This task extends the concept of text-to-video retrieval. Instead of providing only a text query, composed video retrieval provides both a reference video and a textual description that modifies the content of that reference video. The goal is to retrieve a different video that matches the modified description. Composed video retrieval requires compositional reasoning about both the video content and the transformative effect of the text instruction, bridging modalities. The language guidance in this setting is the most complex, requiring the models to interpret complex relationships between the video and the text.

1.3 Contributions

This thesis aims to advance video understanding by leveraging language guidance. By integrating language at various levels, this work aims to push video understanding beyond conventional classification toward more flexible and robust solutions. The thesis tackles different tasks, including audio-visual zero-shot and few-shot video classification, video-adverb retrieval, and composed video retrieval. In each task, language guidance plays a different but crucial role. The primary goal is to improve video understanding by introducing novel representation learning approaches and evaluation protocols while advancing the state-of-the-art in each task. The main contributions of this thesis can be categorized into three main areas:

- Advancing multi-modal learning for video understanding
- Enhancing compositional understanding in video and language
- Establishing new benchmarks and protocols

Advancing Multi-modal Learning for Video Understanding. A core focus of this thesis is on improving how models learn from and integrate information from multiple modalities in the context of videos. Regarding modalities, a focus is set in particular on visual and audio modalities, but also text information is used as a modality. The results presented in this thesis contribute to cross-modal learning and temporal reasoning.

Chapters 2 and 3 tackle audio-visual video classification from a zero-shot and few-shot perspective. For this, it is essential to fuse information across modalities and across their temporal span effectively. The goal hereby is to create robust audio-visual representations that generalise well in data-constrained settings. The possibilities to fuse information from multiple modalities are vast, but transformer-based [Vas+17] architectures have proven

powerful. Their flexibility stems from utilising an attention mechanism, which aggregates and processes information from learned tokens in a modality-agnostic manner.

Specifically, Chapter 2 proposes the transformer-based Temporal and Cross-Attention Framework (TCAF) for generalised audio-visual zero-shot learning to learn multi-modal embeddings that encode information about the audio and visual modalities. Unlike traditional transformer architectures, TCAF constrains the attention mechanism to cross-attention, allowing modalities to not attend to themselves but only to other modalities. This constraint enforces a more effective aggregation across modalities and time, resulting in more generalisable representations.

In Chapter 3, we propose the text-to-feature diffusion framework AV-DIFF for generalised audio-visual few-shot learning. Building upon the constrained audio-visual fusion of TCAF, AV-DIFF proposes a hybrid attention mechanism that first employs intra-modality attention, followed by full attention. This improves learning in the few-shot setting compared to a cross- or full-attention mechanism.

In addition to architectural improvements, this thesis also contributes to multi-modal learning by proposing novel training objectives. In Chapter 4, we propose the REGADA framework for video-adverb retrieval. In addition to existing triplet losses for this task [Dou+20], REGADA also utilises a direct regression objective between learned visual and textual representations, improving the alignment of representations in the shared embedding space. TCAF (Chapter 2) simplifies the loss functions compared to previous work [Mer+22b] for this task, while AV-DIFF is, to our knowledge, the first method in few-shot learning that employs a diffusion model for multi-modal feature generation.

Enhancing Compositional Understanding in Video and Language. Besides multi-modal learning, the works from this thesis also contribute to compositional understanding. This thesis explores how models can learn to compose information within a single modality and from different modalities to achieve a more fine-grained video understanding.

Chapter 4 addresses compositionality between adverbs and verbs. As adverbs can describe multiple actions (e.g. *driving quickly* and *drinking quickly*), the proposed REGADA framework uses a residual gating mechanism to compose textual adverb-action representations before projecting them to a shared embedding space. Explicitly modelling the compositional relationship within language proves beneficial for improving fine-grained understanding of actions in videos.

While REGADA focuses on compositionality within language, Chapter 5 tackles compositionality between videos and natural language in the context of composed video retrieval. Here, a model is tasked to retrieve videos from a video database based on a multi-modal input consisting of a reference video and textual instruction that modifies the given video. The modular re-ranking framework `TFR-CVR` using off-the-shelf foundation models is proposed to improve compositional reasoning across modalities. This training-free framework leverages an LLM to handle the core compositional video-language reasoning. By transforming the vision-language problem into a natural language reasoning task, `TFR-CVR` bridges the gap between visual and linguistic understanding of current VLMs.

Establishing New Benchmarks and Protocols. Besides advancing methodologies, this thesis also contributes to the field by establishing new benchmarks and protocols. This is essential for driving progress in video understanding to address limitations in existing evaluations, establish novel research settings, and allow for more rigorous and generalisable assessment of model capabilities.

Chapter 5 identifies and addresses shortcomings regarding temporal video understanding in benchmarks for composed video retrieval. By analysing the type of modifications described in the textual queries, we find that the existing benchmark [Ven+24] mainly focuses on object-level modifications that do not require a temporal understanding of video events. To address this, we propose the evaluation benchmark `EgoCVR` to allow for a more truthful evaluation regarding the temporal video understanding capabilities of existing methods.

Additionally, Chapter 4 contributes to fine-grained video-adverb retrieval by proposing additional evaluation zero-shot evaluation splits. These splits are designed to benchmark the performance of models on unseen adverb-action compositions to provide a more robust assessment of the generalisation capabilities of existing methods for their understanding of the compositional relationship between adverbs and actions.

Finally, Chapter 3 establishes a new few-shot learning setting for audio-visual video classification, including three new benchmarks, a training and evaluation protocol, and multiple baseline methods. Existing audio-visual video classification benchmarks focused on scenarios with abundant labelled data [Che+20a; Gem+17], and the multi-modal aspect of audio-visual videos was not adequately considered in existing few-shot learning settings [Cao+20; Zha+20]. The developed benchmarks allow for a more realistic evaluation of audio-visual methods regarding their generalisation capabilities.

1.4 Outline

An outline of the thesis chapters is provided below. A summary of the contributions of the individual authors to each publication chapter (Chapters 2 to 5) is provided in Appendix E.2.

Chapter 1: Introduction provides an overview of video understanding and language guidance, introduces the tasks considered, and highlights the main contributions of this thesis.

Chapter 2: Temporal and Cross-Modal Attention for Audio-Visual Zero-Shot Learning introduces our work on audio-visual generalised zero-shot learning, where a model is tasked to recognise video events that have not been observed during training from synchronized audio-video data. Unlike previous works, the developed TCAF framework directly utilises the temporal information naturally present in video data while at the same time simplifying the used loss formulation. In addition, this work proposes a cross-modal attention mechanism that effectively fuses information from audio and visual modalities

to create more generalisable representations. The content of this chapter was published at ECCV 2022 [Mer+22a].

Chapter 3: Text-to-Feature Diffusion for Audio-Visual Few-Shot Learning discusses a continuation of the audio-visual zero-shot learning discussed in Chapter 2 to few-shot learning. This work is the first to explore this specific problem within audio-visual video classification and introduces new few-shot benchmarks with training and evaluation protocols. It employs a diffusion model to generate additional synthetic audio-visual samples for few-shot learning to augment the training. Furthermore, a novel transformer architecture leverages a hybrid attention mechanism, combining intra-modality attention in the initial layers with full attention in subsequent layers. The work from this chapter was published at DAGM GCPR 2023 [Mer+23].

Chapter 4: Video-Adverb Retrieval with Compositional Adverb-Action Embeddings proposes a novel framework for the bidirectional video-adverb retrieval task. Unlike the previous chapter, this work focuses not only on recognising actions but also on understanding the subtle differences in their execution. To address this, this work proposes the REGADA framework, which uses a residual gating mechanism to explicitly exploit the textual compositionality between verbs and adverbs. Furthermore, this work proposes additional zero-shot dataset splits to evaluate the generalisability of learned compositional embeddings more comprehensively. The research of this chapter was published as an oral presentation at BMVC 2023 [Hum+23].

Chapter 5: EgoCVR: An Egocentric Benchmark for Fine-Grained Composed Video Retrieval addresses the task of fine-grained composed video retrieval.This work proposes a new egocentric benchmark, `EgoCVR`, for evaluating composed video retrieval, explicitly focusing on temporal video understanding. This study finds that existing composed video retrieval frameworks often lack the necessary temporal understanding. Our developed training-free re-ranking framework improves performance on `EgoCVR` by modularly leveraging existing VLMs and LLMs to address compositional reasoning. The work from this chapter was published at ECCV 2024 [Hum+24].

Chapter 6: Discussion and Conclusion discusses the contributions of this thesis both individually and collectively. Furthermore, it discusses their limitations, highlights directions for potential future research, and offers concluding remarks.

2

Temporal and Cross-Modal Attention for Audio-Visual Zero-Shot Learning

Audio-visual generalised zero-shot learning for video classification requires understanding the relations between the audio and visual information in order to be able to recognise samples from novel, previously unseen classes at test time. The natural semantic and temporal alignment between audio and visual data in video data can be exploited to learn powerful representations that generalise to unseen classes at test time. We propose a multi-modal and Temporal Cross-attention Framework (TCAF) for audio-visual generalised zero-shot learning. Its inputs are temporally aligned audio and visual features that are obtained from pre-trained networks. Encouraging the framework to focus on cross-modal correspondence across time instead of self-attention within the modalities boosts the performance significantly. We show that our proposed framework that ingests temporal features yields state-of-the-art performance on the UCF-GZSL^{cls}, VGGSound-GZSL^{cls}, and ActivityNet-GZSL^{cls} benchmarks for (generalised) zero-shot learning. Code for reproducing all results is available at https://github.com/ExplainableML/TCAF-GZSL.

2.1 Introduction

Learning task-specific audio-visual representations commonly requires a great number of annotated data samples. However, annotated datasets are limited in size and in the labelled classes that they contain. If a model which was trained with supervision on such a dataset is applied in the real world, it encounters classes that it has never seen. To recognise those novel classes, it would not be feasible to train a new model from scratch. Therefore, it is essential to analyse the behaviour of a trained model in new settings. Ideally, a model should be able to transfer knowledge obtained from classes seen during training to previously unseen categories. This ability is probed in the zero-shot learning (ZSL) task. In addition to zero-shot capabilities, a model should retain the class-specific information from seen training classes. This is challenging and is investigated in the so-called generalised ZSL (GZSL) setting which considers the performance on both, seen

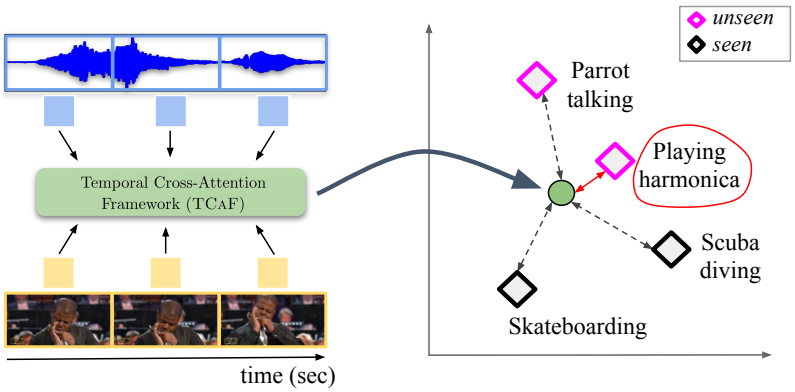

Figure 2.1: Our temporal cross-attention framework for audio-visual (G)ZSL learns a multi-modal embedding (green circle) by exploiting the temporal alignment between audio and visual data in videos. Textual label embeddings (grey squares) are used to transfer information from seen training classes (black) to unseen test classes (pink). The correct class is playing harmonica (red).

and unseen classes.

Prior works [Maz+21; Mer+22b; Par+20] have proposed frameworks that address the (G)ZSL task for video classification using audio-visual inputs. Those methods learn a mapping from the audio-visual input data to textual label embeddings, enabling the classification of samples from unseen classes. At test time, the class whose word embedding is closest to the predicted audio-visual output embedding is selected. Similar to this, we use the textual label embedding space to allow for information transfer from training classes to previously unseen classes. However, [Maz+21; Mer+22b; Par+20] used temporally averaged features as inputs that were extracted from networks pre-trained on video data. The averaging disregarded the temporal dynamics in videos. We propose a Temporal Cross-attention Framework (TCAF) which builds on [Mer+22b] and additionally exploits temporal information by using temporal audio and visual data as inputs. This gives a significant boost in performance for the audio-visual (G)ZSL task compared to using temporally averaged input features. Different from computationally expensive methods that operate directly on raw visual inputs [Bra+20; Ker+21; Lin+22a], our TCAF uses features extracted from networks pre-trained for audio and video classification as inputs. This leads to an efficient setup that uses temporal information instead of averaging across time.

The natural alignment between audio and visual information in videos, e.g. a frog being visible in a frame while the sound of a frog croaking is audible, provides a rich training signal for learning video representations. This can be attributed to the semantic and temporal correlation between the audio and visual information when comparing the two modalities. We encourage our TCAF to put special emphasis on the correlation across the two modalities by employing repeated cross-attention. This attention mechanism only allows attention to tokens from the other modality. This effectively acts as a bottleneck which results in cheaper computations and gives a boost in performance over using full

self-attention across all tokens from both modalities.

We perform a detailed model ablation study to show the benefits of using temporal inputs and our proposed cross-attention. Furthermore, we confirm that our training objective is well-suited to the task at hand. We also analyse the learnt audio-visual embeddings with t-SNE visualisations which confirm that training our TCAF improves the class separation for both seen and unseen classes.

To summarise, our contributions are as follows: (1) We propose a temporal cross-attention framework TCAF for audio-visual (G)ZSL. (2) Our proposed model achieves state-of-the-art results on the UCF-GZSL^{cls}, VGGSound-GZSL^{cls}, and ActivityNet-GZSL^{cls} datasets, demonstrating that using temporal information is extremely beneficial for improving the (generalised) zero-shot classification accuracy compared to using temporally averaged features as model inputs. (3) We perform a detailed analysis of the use of enhanced cross-attention across modalities and time, demonstrating the benefits of our proposed model architecture and training setup.

2.2 Related Work

Our work relates to several themes in the literature: audio-visual learning, ZSL with side information, audio-visual ZSL with side information, and multi-modal transformer architectures. We discuss those in more detail in the following.

Audio-visual learning. The temporal alignment between audio and visual data in videos is a strong learning signal which can be exploited for learning audio-visual representations [Alw+20; AVT16; KTT18; Owe+16; Owe+18; Pat+20]. In addition to audio and video classification, numerous other tasks benefit from audio-visual inputs, such as the separation and localisation of sounds in video data [Afo+22; Afo+20; AZ18; Che+21a; GG19; OE18; Tia+18], audio-driven synthesis of images [JCZ19; WKZ18], audio synthesis driven by visual information [Gan+20; GM18; Koe+20; KWZ19; Nar+21; SLS20; Zho+19], and lip reading [Afo+18; ACZ20]. Some approaches use class-label supervision between modalities [Che+21b; FK20] which does not require the temporal alignment between the input modalities. In contrast to full class-label supervision, we train our model only on the subset of seen training classes.

ZSL with side information. Visual ZSL methods commonly map the visual inputs to class side information [Aka+15a; Aka+15b; Fro+13], e.g. word2vec [Mik+13] class label embeddings. This allows to determine the class with the side information that is closest at test time as the class prediction. Furthermore, attribute annotations have been used as side information [Far+09; Wah+11; Xia+18a; Xia+10]. Recent non-generative methods identify key visual attributes [Xu+20], use attention to find discriminative regions [Xie+19], or disambiguate class embeddings [Liu+19b]. In contrast, feature generation methods train a classifier on generated and real features [Nar+20; Xia+18b; Xia+19; Zhu+19a]. Unlike methods for ZSL with side information with unimodal (visual) inputs, our proposed framework uses multi-modal audio-visual inputs.

Audio-visual ZSL with side information. The task of GZSL from audio-visual data was introduced by [Maz+21; Par+20] on the AudioSetZSL dataset [Par+20] using class label word embeddings as side information. Recently, [Mer+22b] proposed the AVCA framework which uses cross-attention to fuse information from the averaged audio and visual input features for audio-visual GZSL. Our proposed framework builds on [Mer+22b], but instead of using temporally averaged features as inputs [Maz+21; Mer+22b; Par+20], we explore the benefits of using temporal cross-attention information. Unlike [Mer+22b]'s two-stream architecture, we propose the fusion into a single output branch with a classification token that aggregates multi-modal information. Furthermore, we simplify the training objective, and show that the combination of using temporal inputs, our architecture, and training setup leads to superior zero-shot classification performance.

Multi-modal transformers. The success of transformer models in the language domain [Dev+19; Rad+19; Vas+17] has been translated to visual recognition tasks with the Vision Transformer [Dos+21]. Multi-modal vision-language representations have been obtained with a masked language modelling objective, and achieved state-of-the-art performance on several text-vision tasks [Li+20a; Li+19a; Lu+19; Su+19; Sun+19a; Sun+19b; TB19]. In this work, we consider audio-visual multi-modality. Transformer-based models that operate on audio and visual inputs have recently been proposed for text-based video retrieval [Gab+20; Liu+21a; WZY21], dense video captioning [IR20], audio-visual event localization [LW20], and audio classification [BV19]. Different to vanilla transformer-based attention, our TCAF puts special emphasis on cross-attention between the audio and visual modalities in order to learn powerful representations for the (G)ZSL task.

2.3 TCAF Model

In this section, we describe the problem setting (Section 2.3.1), our proposed model architecture (Section 2.3.2), and the loss functions used to train TCAF (Section 2.3.3).

2.3.1 Problem Setting

We address the task of (G)ZSL using audio-visual inputs. The aim of ZSL is to be able to generalise to previously unseen test classes at test time. For GZSL, the model should additionally preserve knowledge about seen training classes, since the GZSL test set contains samples from both, seen and unseen classes.

We denote an audio-visual dataset with N samples and K (seen and unseen) classes by $\mathcal{V} = \{\mathcal{X}_{a[i]}, \mathcal{X}_{v[i]}, y_{[i]}\}_{i=1}^{N}$, consisting of audio data $\mathcal{X}_{a[i]}$, visual data $\mathcal{X}_{v[i]}$, and ground-truth class labels $y_{[i]} \in \mathbb{R}^K$. Naturally, video data contains temporal information. In the following, we use T_a and T_v to denote the number of audio and visual segments in a video clip.

A pre-trained audio classification CNN is used to extract a sequence of audio features $\boldsymbol{a}_{[i]} = \{a_1, \ldots, a_t, \ldots, a_{T_a}\}_i$ to encode the audio information $\mathcal{X}_{a[i]}$. The visual data $\mathcal{X}_{v[i]}$ is

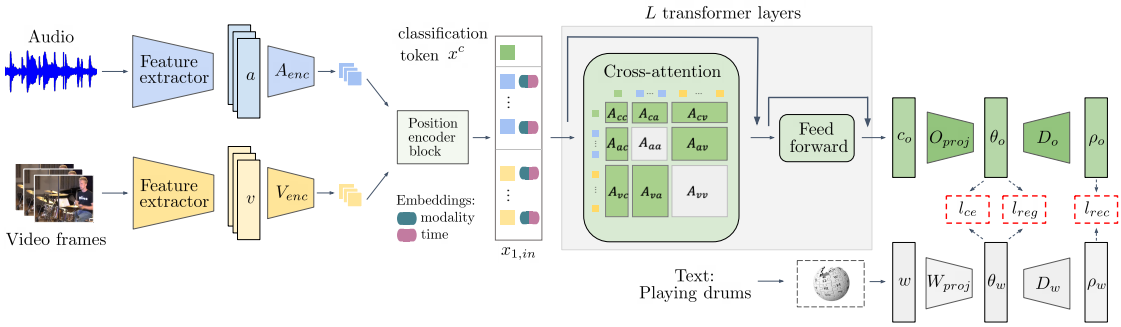

Figure 2.2: TCAF takes audio and visual features extracted from video data as inputs. Those are embedded and equipped with modality and time embeddings before passing through a sequence of L transformer layers with cross-attention. The output classification token c_o is then projected to embedding spaces that are shared with the textual information. The loss functions operate on the joint embedding spaces. At test time, the class prediction c is obtained by determining the word label embedding θ_w^j that is closest to θ_o.

encoded into a temporal sequence of features $\boldsymbol{v}_{[i]} = \{v_1, ..., v_t, ..., v_{T_v}\}_i$ by representing visual segments with features extracted from a pre-trained video classification network.

2.3.2 Model Architecture

In the following, we describe the architecture of our proposed TCAF (see Figure 2.2).

Embedding the inputs and position encoder block. TCAF takes pre-extracted audio and visual features $\boldsymbol{a}_{[i]}$ and $\boldsymbol{v}_{[i]}$ as inputs. For readability, we will drop the subscript i in the following which denotes the $i-$th sample. In order to project audio and visual features to the same feature dimension, $\boldsymbol{a}$ and $\boldsymbol{v}$ are passed through two modality-specific embedding blocks, giving embeddings:

$$\phi_a = A_{enc}(\boldsymbol{a}) \text{ and } \phi_v = V_{enc}(\boldsymbol{v}), \tag{2.1}$$

with $\phi_a \in \mathbb{R}^{T_a * d_{dim}}$ and $\phi_v \in \mathbb{R}^{T_v * d_{dim}}$. The embedding blocks are composed of two linear layers f_1^m, f_2^m for $m \in \{\boldsymbol{a}, \boldsymbol{v}\}$, where $f_1^m : \mathbb{R}^{T_m * d_{in_m}} \rightarrow \mathbb{R}^{T_m * d_{fhidd}}$ and $f_2^m : \mathbb{R}^{T_m * d_{fhidd}} \rightarrow \mathbb{R}^{T_m * d_{dim}}$. f_1^m, f_2^m are each followed by batch normalisation [IS15], a ReLU [NH10], and dropout [Sri+14] with dropout rate $drop_{enc}$.

The position encoder block adds learnt modality and temporal positional embeddings to the outputs of the modality-specific embedding blocks. We explain this in detail below. To handle different frame rates in the audio and visual modalities, we use Fourier features [Tan+20b] $pos_t \in \mathbb{R}^{d_{pos}}$ for the temporal embeddings that encode the actual point in time in the video which corresponds to an audio or visual representation. This allows to capture the relative temporal position of the audio and visual features across the modalities.

For an audio embedding ϕ_{a_t} at time t, a linear map $g_a : \mathbb{R}^{d_{pos}+d_{dim}} \rightarrow \mathbb{R}^{d_{dim}}$, and a dropout layer g^D with dropout probability $drop_{prob,pos}$, we obtain position-aware audio

feature tokens

$$a_t^p = g^D(g_a(concat(\phi_{a_t}, pos_{at}))) \qquad \text{with} \qquad pos_{at} = pos_a + pos_t, \tag{2.2}$$

with modality and temporal embeddings $pos_a, pos_t \in \mathbb{R}^{d_{pos}}$ respectively. Position-aware visual tokens v_t^p are obtained analogously.

Furthermore, we prepend a learnt classification token $x^c \in \mathbb{R}^{d_{dim}}$ to the sequence of feature tokens. The corresponding output classification token c_o is used by our output projection O_{proj} to obtain the final prediction.

Audio-visual transformer layers. TCAF contains L stacked audio-visual transformer layers that allow for enhanced cross-attention. Each of our transformer layers consists of an attention function $f_{l,Att}$, followed by a feed forward function $g_{l,FF}$. The output of the l-th transformer layer is given as

$$x_{l,out} = x_{l,ff} + x_{l,att} = g_{l,FF}(x_{l,att}) + x_{l,att}, \tag{2.3}$$

with

$$x_{l,att} = f_{l,Att}(x_{l,in}) + x_{l,in}, \tag{2.4}$$

where

$$x_{l,in} = \begin{cases} [x^c, a_1^p, \cdots, a_{T_a}^p, v_1^p, \cdots, v_{T_v}^p] & \text{if } l = 1, \\ x_{l-1,out} & \text{if } 2 \geq l \leq L. \end{cases}$$

We explain the cross-attention used in our transformer layers in the following.

Transformer cross-attention. TCAF primarily exploits cross-modal audio-visual attention to combine the information across the audio and visual modalities. All attention mechanisms in TCAF consist of multi-head attention [Vas+17] with H heads and a dimension of d_{head} per head.

We describe the first transformer layer $\mathcal{M}_1$, the transformer layer $\mathcal{M}_l$ operates analogously. We project the position-aware input features x^c, $\{a_t^p\}_{t\in[1,T_a]}$, $\{v_t^p\}_{t\in[1,T_v]}$ to queries, keys, and values with linear maps $g_s : \mathbb{R}^{d_{dim}} \to \mathbb{R}^{d_{head}H}$ for $s \in \{q, k, v\}$. We can then write the outputs of the projection as zero-padded query, key, and value features. We write those out for the queries below, the keys and values are padded in the same way:

$$\mathbf{q}_c = [g_q(x^c), 0, \cdots, 0], \tag{2.5}$$

$$\mathbf{q}_a = [0, \cdots, 0, g_q(a_1^p), \cdots, g_q(a_{T_a}^p), 0, \cdots, 0], \tag{2.6}$$

$$\mathbf{q}_v = [0, \cdots, 0, g_q(v_1^p), \cdots, g_q(v_{T_v}^p)]. \tag{2.7}$$

The full query, key, and value representations, $\mathbf{q}$, $\mathbf{k}$, and $\mathbf{v}$, are the sums of their modality-specific components

$$\mathbf{q} = \mathbf{q}_c + \mathbf{q}_a + \mathbf{q}_v, \qquad \mathbf{k} = \mathbf{k}_c + \mathbf{k}_a + \mathbf{k}_v, \qquad \text{and } \mathbf{v} = \mathbf{v}_c + \mathbf{v}_a + \mathbf{v}_v. \tag{2.8}$$

The output of the first attention block $x_{1,att}$ is the aggregation of the per-head attention with a linear mapping $g_h : \mathbb{R}^{d_{head}H} \rightarrow \mathbb{R}^{d_{dim}}$, g^{DL} dropout with dropout probability $drop_{prob}$ and layer normalisation g^{LN} [BKH16], such that

$$x_{1,att} = f_{1,Att}(x_{l,in}) = g^{DL}(g_h(f^1_{1,att}(g^{LN}(x_{1,in})), \cdots, f^H_{1,att}(g^{LN}(x_{1,in})))), \tag{2.9}$$

with the attention f^h_{att} for the attention head h. We can write the attention for the head h as

$$f^h_{att}(x_{1,in}) = softmax\left(\frac{\mathbf{A}}{\sqrt{d_{head}}}\right)\mathbf{v}, \tag{2.10}$$

where $\mathbf{A}$ can be split into its cross-attention and self-attention components:

$$\mathbf{A}_c = \mathbf{q}_c\,\mathbf{k}^T + \mathbf{k}\,\mathbf{q}_c^T, \qquad \mathbf{A}_x = \mathbf{q}_a\,\mathbf{k}_v^T + \mathbf{q}_v\,\mathbf{k}_a^T, \tag{2.11}$$

$$\mathbf{A}_{self} = \mathbf{q}_a\,\mathbf{k}_a^T + \mathbf{q}_v\,\mathbf{k}_v^T.$$

We then get

$$\mathbf{A} = \mathbf{A}_c + \mathbf{A}_x + \mathbf{A}_{self} = \begin{pmatrix} A_{cc} & A_{ca} & A_{cv} \\ A_{ac} & \ddots & \vdots \\ A_{vc} & \dots & 0 \end{pmatrix} + \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & A_{av} \\ 0 & A_{va} & 0 \end{pmatrix} + \begin{pmatrix} 0 & \dots & 0 \\ \vdots & A_{aa} & \vdots \\ 0 & \dots & A_{vv} \end{pmatrix}, \tag{2.12}$$

where the A_{mn} with $m, n \in \{c, a, v\}$ describe the attention contributions from the classification token, the audio and the visual modalities respectively.

Our TCAF uses the cross-attention $\mathbf{A}_c + \mathbf{A}_x$ to put special emphasis on the attention across modalities. Results for different model variants that use only the within-modality self-attention ($\mathbf{A}_c + \mathbf{A}_{self}$) or the full attention which combines self-attention and cross-attention are presented in Section 2.4.3.

Feed forward function. The feed forward function $g_{l,FF} : \mathbb{R}^{d_{dim}} \rightarrow \mathbb{R}^{d_{dim}}$ is applied to the output of the attention function

$$x_{l,ff} = g_{l,FF}(x_{l,att}) = g^{DL}(g_{l,F2}(g^{DL}(g^{GD}(g_{l,F1}(g^{LN}(x_{l,att})))))) \tag{2.13}$$

where $g_{l,F1} : \mathbb{R}^{d_{dim}} \rightarrow \mathbb{R}^{d_{ff}}$ and $g_{l,F2} : \mathbb{R}^{d_{ff}} \rightarrow \mathbb{R}^{d_{dim}}$ are linear mappings, g^{GD} is a GELU layer [HG16] and a dropout layer with dropout probability $drop_{prob}$, g^{DL} is dropout with $drop_{prob}$ and g^{LN} is layer normalisation.

Output prediction. To determine the final class prediction, the audio-visual embedding is projected to the same embedding space as the textual class label representations. We project the output classification token c_o of the temporal cross-attention to $\theta_o = O_{proj}(c_o)$ where $\theta_o \in \mathbb{R}^{d_{out}}$. The projection block is composed of a sequence of two linear layers f_3 and f_4, where $f_3 : \mathbb{R}^{d_{dim}} \rightarrow \mathbb{R}^{d_{fhidd}}$ and $f_4 : \mathbb{R}^{d_{fhidd}} \rightarrow \mathbb{R}^{d_{out}}$. f_3, f_4 are each followed by batch normalisation, a ReLU, and dropout with rate $drop_{proj_o}$. We project the word2vec class label embedding w^j for class j using the projection block $W_{proj}(w^j) = \theta^j_w$, where $\theta^j_w \in \mathbb{R}^{d_{out}}$. W_{proj} consists of a linear projection followed by batch normalisation, ReLU, and

dropout with dropout rate $drop_{proj_w}$. The class prediction c is obtained by determining the projected word2vec embedding which is closest to the output embedding:

$$c = \underset{j}{\operatorname{argmin}}(\| \theta_w^j - \theta_o \|_2). \tag{2.14}$$

2.3.3 Loss Functions

Our training objective l combines a cross-entropy loss l_{ce}, a reconstruction loss l_{rec}, and a regression loss l_{reg}:

$$l = l_{ce} + l_{rec} + l_{reg}. \tag{2.15}$$

Cross-entropy loss. For the ground-truth label y_i with corresponding class index $k_{gt} \in \mathbb{R}^{K_{seen}}$, the output of our temporal cross-attention θ_{o_i}, and a matrix containing the textual label embeddings for the K_{seen} seen classes $\theta_{w_{seen}}$, we define the cross-entropy loss for n training samples as

$$l_{ce} = -\frac{1}{n} \sum_{i}^{n} y_i \log \left(\frac{\exp\left(\theta_{w_{seen}, k_{gt}} \theta_{o_i}\right)}{\sum_{k_j}^{K_{seen}} \exp\left(\theta_{w_{seen}, k_j} \theta_{o_i}\right)} \right). \tag{2.16}$$

Regression loss. While the cross-entropy loss updates the probabilities for both the correct and incorrect classes, our regression loss directly focuses on reducing the distance between the output embedding for a sample and the corresponding projected word2vec embedding. The regression loss is based on the mean squared error metric with the following formulation:

$$l_{reg} = \frac{1}{n} \sum_{i=1}^{n} (\theta o_i - \theta_{w_i})^2, \tag{2.17}$$

where θ_{o_i} is the audio-visual embedding, and θ_{w_i} is the projection of the word2vec embedding corresponding to the i-th sample.

Reconstruction loss. The goal of the reconstruction loss is to ensure that the embeddings θ_o and θ_w contain semantic information from the word2vec embedding w. We use $D_u : \mathbb{R}^{d_{out}} \mapsto \mathbb{R}^{d_{dim}}$ with $\rho_u = D_u(\theta_u)$ for $u \in \{o, w\}$. D_w is a sequence of one linear layer, batch normalisation, a ReLU, and dropout with rate $drop_{proj_w}$. D_o is composed of a sequence of two linear layers each followed by batch normalisation, a ReLU, and dropout with dropout rate $drop_{proj_o}$. Our reconstruction loss encourages the reconstruction of the output embedding, ρ_{o_i}, and the reconstruction of the word2vec projection, ρ_{w_i}, to be close to the original word2vec embedding w_i:

$$l_{rec} = \frac{1}{n} \sum_{i=1}^{n} (\rho_{o_i} - w_i)^2 + \frac{1}{n} \sum_{i=1}^{n} (\rho_{w_i} - w_i)^2. \tag{2.18}$$

2.4 Experiments

In this section, we detail our experimental setup (Section 2.4.1), and compare to state-of-the-art methods for audio-visual GZSL (Section 2.4.2). Furthermore, we present an

ablation study in Section 2.4.3 which shows the benefits of using our proposed attention scheme and training objective. Finally, we present t-SNE visualisations of our learnt audio-visual embeddings in Section 2.4.4.

2.4.1 Experimental Setup

Here, we describe the datasets used, the evaluation metrics, and the implementation details for all models.

Datasets. We use the UCF-GZSL^{cls}, VGGSound-GZSL^{cls}, and ActivityNet-GZSL^{cls} datasets [Mer+22b] for audio-visual (G)ZSL for training and testing all models. [Mer+22b] introduced benchmarks for two sets of features, the first uses a model pre-trained using self-supervision on the VGGSound dataset from [Asa+20], the second takes features extracted from pre-trained VGGish [Her+17] and C3D [Tra+15] audio and video classification networks. Since the VGGSound dataset is also used for the zero-shot learning task (VGGSound-GZSL), we selected the second option (using VGGish and C3D) and use the corresponding dataset splits proposed in [Mer+22b]. We additionally provide results on the UCF-GZSL, VGGSound-GZSL, and ActivityNet-GZSL datasets in the supplementary material.

In particular, the audio features are extracted using VGGish [Her+17] to obtain one 128-dimensional feature vector for each 0.96 s snippet. The visual features are obtained using C3D [Tra+15] pre-trained on Sports-1M [Kar+14]. For this, all videos are resampled to 25 fps. A 4096-dimensional feature vector is then extracted for 16 consecutive video frames.

Evaluation metrics. We follow [Mer+22b; Xia+18a] and use the mean class accuracy to evaluate all models. The ZSL performance is obtained by considering only the subset of test samples from the unseen test classes. For the GZSL performance, the models are evaluated on the full test set which includes seen and unseen classes. We then report the performance on the subsets of seen (S) and unseen (U) classes, and also report their harmonic mean (HM).

Implementation details. For TCAF, we use $d_{in_a} = 128$, $d_{in_v} = 4096$, $d_{fhidd} = 512$, $d_{dim} = 300$ and $d_{out} = 64$. Furthermore, TCAF has $L = 6$ transformer layers layers for UCF-GZSL^{cls} and ActivityNet-GZSL^{cls}, and $L = 8$ for VGGSound-GZSL^{cls}. We set $d_{pos} = 64$, $d_{ff} = 128$. For ActivityNet-GZSL^{cls} / UCF-GZSL^{cls} / VGGSound-GZSL^{cls} we use dropout rates $drop_{enc} = 0.1/0.3/0.2$, $drop_{prob,pos} = 0.2/0.2/0.1$, $drop_{prob} = 0.4/0.3/0.5$, $drop_{proj_w} = 0.1/0.1/0.1$, and $drop_{proj_o} = 0.1/0.1/0.2$. All attention blocks use $H = 8$ heads with a dimension of $d_{head} = 64$ per head. We train all models using the Adam optimizer [KB14] with running average coefficients $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 0.00001. We use a batch size of 64 for all datasets. In order to efficiently train on ActivityNet-GZSL^{cls}, we randomly trim the features to a maximum sequence length of 60 during training, and we evaluate on features that have a maximum sequence length of 300 and which are centered in the middle of the video. We note, that TCAF can be efficiently trained on a single

Nvidia 2080-Ti GPU. All models are trained for 50 epochs. We use a base learning rate of 0.00007 for UCF-GZSL^{cls} and ActivityNet-GZSL^{cls}, and 0.00006 for VGGSound-GZSL^{cls}. For UCF-GZSL^{cls} and ActivityNet-GZSL^{cls} we use a scheduler that reduces the learning rate by a factor of 0.1 when the HM on the validation set has not improved for 3 epochs. To eliminate the bias that the ZSL methods have towards seen classes, we used calibrated stacking [Cha+16] on the search space composed of the interval $[0, 3]$ with a step size of 0.2.

We train all models with a two-stage training protocol [Mer+22b]. In the first stage, we determine the calibrated stacking [Cha+16] and the epoch with the best HM performance on the validation set. In the second stage, using the hyperparameters from the first stage, we re-train the models on the union of the training and validation sets. We evaluate the final models on the test set.

2.4.2 Quantitative Results

We compare our proposed TCAF to state-of-the-art audio-visual ZSL frameworks and to audio-visual frameworks that we adapted to the ZSL task.

Audio-visual ZSL baselines. We compare our TCAF to three audio-visual ZSL frameworks. **CJME** [Par+20] consists of a relatively simple architecture which maps both input modalities to a shared embedding space. The modality-specific embeddings in the shared embedding space are input to an attention predictor module that determines the dominant modality which is used for the output prediction. **AVGZSLNet** [Maz+21] builds on CJME by adding a shared decoder and introducing additional loss functions to improve the performance. AVGZSLNet removes the attention predictor network and replaces it with a simple average between the output from the head of each modality. **AVCA** [Mer+22b] is a recent state-of-the-art method for audio-visual G(ZSL). It uses a simple cross-attention mechanism on the temporally averaged audio and visual input features to combine the information from the two modalities. Our proposed TCAF improves upon the closely related AVCA framework by additionally ingesting temporal information in the audio and visual inputs with an enhanced cross-attention mechanism that gathers information across time and modalities.

Audio-visual baselines adapted to ZSL. We adapt two attention-based audio-visual frameworks to the ZSL setting. **Attention Fusion** [FK20] is a method for audio-visual classification which is trained to classify unimodal information. It then fuses the unimodal predictions with learnt attention weights. The **Perceiver** [Jae+21] is a scalable multi-modal transformer framework for flexible learning with arbitrary modality information. It uses a latent bottleneck to encode input information by repeatedly attending to the input with transformer-style attention. The Perceiver allows for a comparison to another transformer-based architecture with focus on multi-modality. We adapt the Perceiver to use the same positional encodings and model capacity as TCAF. We use 64 latent tokens and the same number of layers and dimensions as TCAF. Both Attention Fusion and Perceiver use the

same input features, input embedding functions A_{enc} and V_{enc}, learning rate and loss functions as TCAF. For Attention Fusion, we temporally average the input features after A_{enc} and V_{enc} to deal with non-synchronous modality sequences due to different feature extraction rates.

All baselines, except for the Perceiver, operate on temporally averaged audio and visual features. This decreases the amount of information contained in the inputs, in particular regarding the dynamics in a video. In contrast to methods that use temporally averaged inputs, TCAF exploits the temporal dimension which boosts the (G)ZSL performance.

Results. We compare the results obtained with our TCAF to state-of-the-art baselines for audio-visual (G)ZSL and for audio-visual learning in Table 2.1. TCAF outperforms all previous methods on the VGGSound-GZSL^{cls}, UCF-GZSL^{cls}, and ActivityNet-GZSL^{cls} datasets for both, GZSL performance (HM) and ZSL performance. For ActivityNet-GZSL^{cls}, our proposed model is significantly better than its strongest competitor AVCA, with a HM of 12.20% compared to 9.92% and a ZSL performance of 7.96% compared to 7.58%. The CJME and AVGZSLNet frameworks are weaker than the AVCA model. Similar patterns are exhibited for the VGGSound-GZSL^{cls} and UCF-GZSL^{cls} datasets. Interestingly, the GZSL performance for TCAF is improved by a more significant margin than the ZSL performance compared to AVCA across all three datasets. This shows that using temporal information and allowing our model to attend across time and modalities is especially beneficial for the GZSL task.

Furthermore, we observe that the audio-visual Attention Fusion framework and the Perceiver give worse results than AVGZSLNet and AVCA on all three datasets. In particular, our TCAF yields stronger ZSL and GZSL performances than the Perceiver which also takes temporal audio and visual features as inputs, with a HM of 8.77% on VGGSound-GZSL^{cls} for TCAF compared to 4.93% for the Perceiver. Attention Fusion and the Perceiver architecture were not designed for the (G)ZSL setting that uses text as side information. Our proposed training objective, used to also train the Perceiver, aims to regress textual embeddings which might be challenging for the Perceiver given its tight latent bottlenecks.

2.4.3 Ablation Study on the Training Loss and Attention Variants

Here, we analyse different components of our proposed TCAF. We first compare the performance of our model when trained using different loss functions. We then investigate the influence of the attention mechanisms used in the model architecture on the (G)ZSL performance. Finally, we show that using multi-modal inputs is beneficial and results in outperforming unimodal baselines.

Comparing different training losses. We show the contributions of the different components in our training loss function to the (G)ZSL performance in Table 2.2. Using only the regression loss l_{reg} to train our model results in the weakest performance across all datasets, with HM/ZSL performances of 16.25%/30.17% on UCF-GZSL^{cls} compared to 50.78%/44.64% for our full TCAF. Interestingly, the seen performance (S) when using only

Model	VGGSound-GZSL^{cls}				UCF-GZSL^{cls}				ActivityNet-GZSL^{cls}			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
Attention Fusion	14.13	3.00	4.95	3.37	39.34	18.29	24.97	20.21	11.15	3.37	5.18	4.88
Perceiver	13.25	3.03	4.93	3.44	46.85	26.82	34.11	28.12	18.25	4.27	6.92	4.47
CJME	10.86	2.22	3.68	3.72	33.89	24.82	28.65	29.01	10.75	5.55	7.32	6.29
AVGZSLNet	**15.02**	3.19	5.26	4.81	**74.79**	24.15	36.51	31.51	13.70	5.96	8.30	6.39
AVCA	12.63	6.19	8.31	6.91	63.15	30.72	41.34	37.72	16.77	7.04	9.92	7.58
TCAF	12.63	**6.72**	**8.77**	**7.41**	67.14	**40.83**	**50.78**	**44.64**	**30.12**	**7.65**	**12.20**	**7.96**

Table 2.1: Performance of our TCAF and of state-of-the-art methods for audio-visual (G)ZSL on the VGGSound-GZSL^{cls}, UCF-GZSL^{cls}, and ActivityNet-GZSL^{cls} datasets. The mean class accuracy for GZSL is reported on the seen (S) and unseen (U) test classes, and their harmonic mean (HM). For the ZSL performance, only the test subset of unseen classes is considered.

Loss	VGGSound-GZSL^{cls}				UCF-GZSL^{cls}				ActivityNet-GZSL^{cls}			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
l_{reg}	0.10	2.41	0.19	2.50	14.30	18.82	16.25	30.17	1.09	0.27	0.43	2.11
$l_{reg} + l_{ce}$	**13.67**	4.06	6.26	4.31	**75.31**	37.15	49.76	41.75	11.36	5.28	7.21	5.31
$l = l_{reg} + l_{ce} + l_{rec}$	12.63	**6.72**	**8.77**	**7.41**	67.14	**40.83**	**50.78**	**44.64**	**30.12**	**7.65**	**12.20**	**7.96**

Table 2.2: Influence of using different components of our proposed training objective for training TCAF on the (G)ZSL performance on the VGGSound-GZSL^{cls}, UCF-GZSL^{cls}, and ActivityNet-GZSL^{cls} datasets.

l_{reg} is relatively weak, likely caused by the calibrated stacking. Similarly, on ActivityNet-GZSL^{cls}, using only l_{reg} yields a low test performance of 0.43% HM. Jointly training with the regression and cross-entropy loss functions ($l_{reg} + l_{ce}$) improves the GZSL and ZSL performance significantly, giving a ZSL performance of 4.31% compared to 2.50% for l_{reg} on VGGSound-GZSL^{cls}. The best results are obtained when training with our full training objective l which includes a reconstruction loss term, giving the best performance on all three datasets.

Comparing different attention variants. We study the use of different attention patterns in Table 2.3. In particular, we analyse the effect of using within-modality ($\mathbf{A}_{self}$) and cross-modal ($\mathbf{A}_x$) attention (cf. Equation (2.11)), on the GZSL and ZSL performance. Additionally, we investigate models that use a classification token x^c with corresponding output token c_o (*with class. token*) and models for which we simply average the output of the transformer layers which is then used as input to O_{proj} (*w/o class. token*).

Interestingly, we observe that with no global token, using the full attention $\mathbf{A}_{self} + \mathbf{A}_x$ gives better results than using only cross-attention on UCF-GZSL^{cls} and ActivityNet-GZSL^{cls} for ZSL and GZSL, but is slightly worse on VGGSound-GZSL^{cls}. This suggests that the bottleneck introduced by limiting the information flow in the attention when using only cross-attention is beneficial for (G)ZSL on VGGSound-GZSL^{cls}. When not using the classification token and only self-attention $\mathbf{A}_{self}$, representations inside the

Model	VGGSound-GZSL^{cls}				UCF-GZSL^{cls}				ActivityNet-GZSL^{cls}			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
w/o class. token												
$\mathbf{A}_{self} + \mathbf{A}_x$	**18.40**	3.78	6.27	4.25	31.70	32.57	32.13	33.26	11.87	3.80	5.75	3.90
$\mathbf{A}_{self}$	16.08	3.56	5.83	4.00	42.59	24.04	30.73	27.49	9.51	4.33	5.95	4.39
$\mathbf{A}_x$	14.62	4.22	6.55	4.59	19.52	29.80	23.62	31.35	1.85	3.50	2.42	3.50
with class. token												
$\mathbf{A}_c + \mathbf{A}_{self} + \mathbf{A}_x$	11.36	5.50	7.41	5.97	36.73	**41.99**	39.18	42.56	17.75	6.79	9.83	6.89
$\mathbf{A}_c + \mathbf{A}_{self}$	12.23	4.63	6.71	5.25	40.14	34.95	37.37	35.74	4.24	3.23	3.67	3.25
$\mathbf{A}_c + \mathbf{A}_x$ (TCAF)	12.63	**6.72**	**8.77**	**7.41**	**67.14**	40.83	**50.78**	**44.64**	**30.12**	**7.65**	**12.20**	**7.96**

Table 2.3: Ablation of different attention variants with and without a classification token on the VGGSound-GZSL^{cls}, UCF-GZSL^{cls}, and ActivityNet-GZSL^{cls} datasets.

Model	VGGSound-GZSL^{cls}				UCF-GZSL^{cls}				ActivityNet-GZSL^{cls}			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
TCAF- audio	5.11	4.06	4.53	4.28	35.51	19.75	25.38	24.24	9.28	4.26	5.84	4.65
TCAF- visual	3.97	3.12	3.50	3.19	38.10	26.84	31.49	27.25	2.75	3.11	2.92	3.11
TCAF	**12.63**	**6.72**	**8.77**	**7.41**	**67.14**	**40.83**	**50.78**	**44.64**	**30.12**	**7.65**	**12.20**	**7.96**

Table 2.4: Influence of using multiple modalities for training and evaluating our proposed model on the (G)ZSL performance on the VGGSound-GZSL^{cls}, UCF-GZSL^{cls}, and ActivityNet-GZSL^{cls} datasets.

transformer are created solely within their respective modalities.

Using a classification token (*with class. token*) and the cross-attention variant ($\mathbf{A}_c + \mathbf{A}_x$) yields the strongest ZSL and GZSL results across all three datasets. The most drastic improvements over full attention can be observed on the UCF-GZSL^{cls} dataset, with a HM of 50.78% for the cross-attention with classification token ($\mathbf{A}_c + \mathbf{A}_x$) compared to 39.18% for the full attention ($\mathbf{A}_c + \mathbf{A}_{self} + \mathbf{A}_x$). Furthermore, when using x_c, cross-attention $\mathbf{A}_x$ instead of self-attention $\mathbf{A}_{self}$ leads to a better performance on all three datasets. For $\mathbf{A}_x$ and x_c, we obtain HM scores of 8.77% and 50.78 % on VGGSound-GZSL^{cls} and UCF-GZSL^{cls} compared to 6.71% and 37.37% with $\mathbf{A}_{self}$ and x_c. This shows that using information from both modalities is important for creating strong and transferable video representations for (G)ZSL. Using the global token relaxes the pure cross-attention setting to a certain extent, since $\mathbf{A}_c$ allows for attention between all tokens from both modalities and the global token. The results in Table 2.3 have demonstrated the clear benefits of our cross-attention variant used in TCAF.

The influence of multi-modality. We compare using only a single input modality for training TCAF to using multiple input modalities in Table 2.4. For the unimodal baselines TCAF- audio and TCAF- visual, we train TCAF only with the corresponding input modality. Using only audio inputs gives stronger GZSL and ZSL results than using only visual inputs on VGGSound-GZSL^{cls} and ActivityNet-GZSL^{cls}. We obtain a HM of 5.84% for audio compared to 2.92% for visual inputs on ActivityNet-GZSL^{cls}. Interestingly

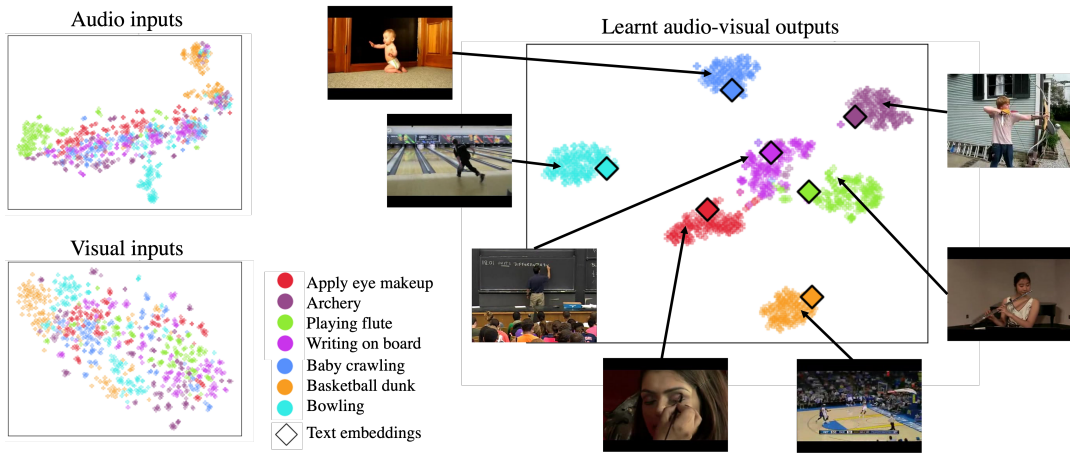

Figure 2.3: t-SNE visualisation for five seen (*apply eye makeup, archery, baby crawling, basketball dunk, bowling*) and two unseen (*playing flute, writing on board*) test classes from the UCF-GZSL^{cls} dataset, showing audio and visual input embeddings extracted with C3D and VGGish, and audio-visual output embeddings learned with TCAF. Textual class label embeddings are visualised with a square.

this pattern is reversed for the UCF-GZSL^{cls} dataset where using visual inputs only results in a slightly higher performance than using the audio inputs with HM scores of 31.49% compared to 25.38%, and ZSL scores of 27.25% and 24.24%. However, using both modalities (TCAF) increases the HM to 50.78% and ZSL to 44.64% on UCF-GZSL^{cls}. Similar trends can be observed for VGGSound-GZSL^{cls} and ActivityNet-GZSL^{cls} which highlights the importance of the tight multi-modal coupling in our TCAF.

2.4.4 Qualitative Results

We present a qualitative analysis of the learnt audio-visual embeddings in Figure 2.3. For this, we show t-SNE [VH08] visualisations for the audio and visual input features and for the learnt multi-modal embeddings from 7 classes in the UCF-GZSL^{cls} test set. We averaged the input features for both modalities across time. We observe that the audio and visual input features are poorly clustered. In contrast, the audio-visual embeddings (θ_o) are clearly clustered for both, seen and unseen classes. This suggests that our network is actually learning useful representations for unseen classes, too. Furthermore, the word2vec class label embeddings (θ_w^j) lie inside the corresponding audio-visual clusters. This confirms that the learnt audio-visual embeddings are mapped to locations that are close to the corresponding word2vec embeddings, showing that our embeddings capture semantic information from the word2vec representations.

2.5 Conclusion

We presented a cross-attention transformer framework that addresses (G)ZSL for video classification using audio-visual input data with temporal information. Our proposed

model achieves state-of-the-art performance on the three audio-visual (G)ZSL datasets UCF-GZSL^{cls}, VGGSound-GZSL^{cls}, and ActivityNet-GZSL^{cls}. The use of pre-extracted audio and visual features as inputs results in a computationally efficient framework compared to using raw data. We demonstrated that using cross-modal attention on temporal audio and visual input features and suppressing the contributions from the within-modality self-attention is beneficial for obtaining strong audio-visual embeddings that can transfer information from classes seen during training to novel, unseen classes at test time.

3

Text-to-Feature Diffusion for Audio-Visual Few-Shot Learning

Training deep learning models for video classification from audio-visual data commonly requires vast amounts of labeled training data collected via a costly process. A challenging and underexplored, yet much cheaper, setup is few-shot learning from video data. In particular, the inherently multi-modal nature of video data with sound and visual information has not been leveraged extensively for the few-shot video classification task. Therefore, we introduce a unified audio-visual few-shot video classification benchmark on three datasets, i.e. the VGGSound-FSL, UCF-FSL, ActivityNet-FSL datasets, where we adapt and compare ten methods. In addition, we propose AV-Diff, a text-to-feature diffusion framework, which first fuses the temporal and audio-visual features via cross-modal attention and then generates multi-modal features for the novel classes. We show that AV-Diff obtains state-of-the-art performance on our proposed benchmark for audio-visual (generalised) few-shot learning. Our benchmark paves the way for effective audio-visual classification when only limited labeled data is available. Code and data are available at https://github.com/ExplainableML/AVDIFF-GFSL.

3.1 Introduction

The use of audio-visual data can yield impressive results for video classification [Nag+21; Pat+20; Xia+20]. The complementary knowledge contained in the two modalities results in a richer learning signal than using unimodal data. However, video classification frameworks commonly rely on significant amounts of costly training data and computational resources. To mitigate the need for large amounts of labeled data, we consider the few-shot learning (FSL) setting where a model is tasked to recognise new classes with only few labeled examples. Moreover, the need for vast computational resources can be alleviated by operating on the feature level, using features extracted from pre-trained visual and sound classification networks.

In this work, we tackle the task of few-shot action recognition in videos from audio and visual data which is an understudied problem in computer vision. In the few-shot setting,

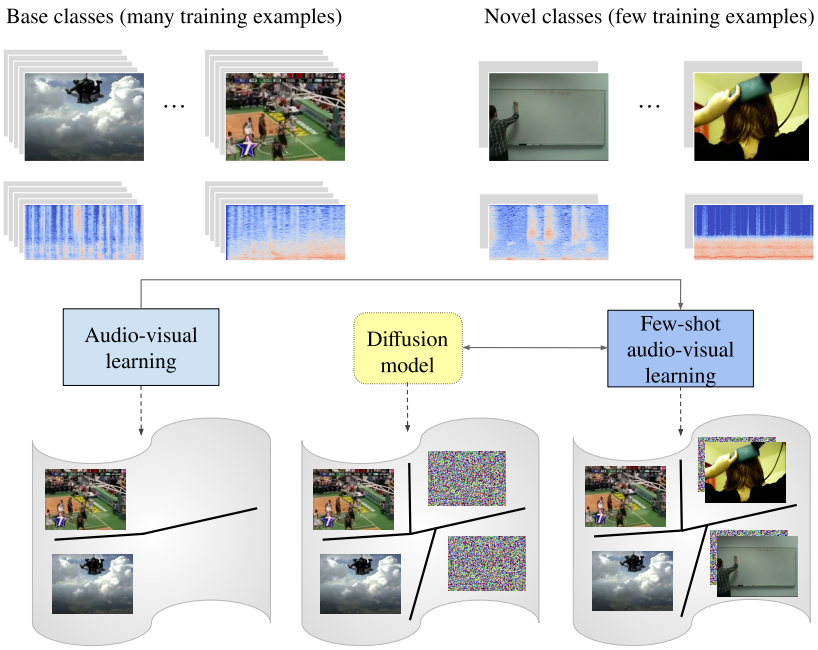

Figure 3.1: AV-DIFF learns to fuse the audio-visual inputs into multi-modal representations in the audio-visual learning stage (left). In the few-shot learning stage (right), the multi-modal representations from the previous stage are used to concurrently train (double arrow line) a text-conditioned diffusion model on all the classes (middle) and a classifier. The classifier is trained on real features from base classes and real and synthetic features from novel classes.

a model has to learn a transferable audio-visual representation which can be adapted to new classes with few annotated data samples. In particular, we focus on the more practical generalised FSL (GFSL) setting, where the aim is to recognise samples from both the base classes, i.e. classes with many training samples, and from novel classes which contain only few examples. Additional modalities, such as text and audio, are especially useful for learning transferable and robust representations from few samples.

To the best of our knowledge, the FSL setting with audio-visual data has only been considered for speech recognition [Zha+22], and for learning an acoustic model of 3D scenes [Maj+22]. Moreover, existing video FSL benchmarks are not suitable for the audio-visual setting. In particular, the SomethingV2 and HMDB51 benchmarks proposed in [Cao+20] and [Zha+20] do not contain audio and about 50% of the classes in the UCF101 benchmark from [Xia+21] have no sound either. The Kinetics split in [ZY18] suffers from an overlap with the classes used to pre-train the feature extractors [Xia+21], and [Nag+21; Xia+20] show that the audio modality in Kinetics is less class-relevant than the visual modality. Existing audio-visual zero-shot learning benchmarks [Mer+22a; Mer+22b] cannot directly be used for few-shot learning due to their distinct training and testing protocols. Moreover, the baselines in both settings differ significantly as state-of-the-art few-shot learning methods usually necessitate knowledge of novel classes through classification objectives and generative models, a condition that is not possible in zero-shot

learning. Thus, we introduce a new benchmark for generalised audio-visual FSL for video classification that is comprised of three audio-visual datasets and ten methods carefully adapted to this challenging, yet practical task.

To tackle our new benchmark, we propose AV-DIFF which uses a novel hybrid cross-modal attention for fusing audio-visual information. Different to various attention fusion techniques in the audio-visual domain [Mer+22a; Mer+22b; Nag+21] which use a single attention type or different transformers for each modality, our model makes use of a novel combination of within-modality and cross-modal attention in a multi-modal transformer. This allows the effective fusion of information from both modalities and across the temporal dimension of the inputs. Furthermore, we introduce a novel text-conditioned diffusion model for generating audio-visual features to augment the few samples in the novel classes. In the image and video domain, generative adversarial networks (GANs) have been used to generate uni-modal features for data augmentation in the FSL setting [HG17; Kum+19; Nar+20; Xia+21; Xia+19]. However, we are not aware of prior works that have used diffusion models for multi-modal (audio-visual) feature generation in FSL. Both, cross-modal fusion and the text-to-feature diffusion contribute to significant boosts in performance on our proposed benchmark.

To summarise, our contributions are: 1) We introduce the audio-visual generalised few-shot learning task for video classification and a benchmark on three audio-visual datasets. We additionally adapt and compare ten methods for this task. 2) We propose a hybrid attention mechanism to fuse multi-modal information, and a diffusion model for multi-modal feature generation to augment the training dataset with additional novel-class samples. 3) We obtain state-of-the-art performance across all three datasets, outperforming the adapted multi-modal zero-shot learning and video FSL models.

3.2 Related Work

We discuss prior works in learning from audio-visual data, FSL, and feature generation in low-shot learning.

Audio-visual learning. Multi-modal inputs, such as audio and visual data, provide significantly more information than unimodal data, resulting in improved overall performance for video classification and acoustic scene classification [Alw+20; AVT16; KTT18; Owe+16; Owe+18; Pat+20]. Approaches, such as [Che+21b; FK20], use class-label supervision between modalities without requiring temporal alignment between the input modalities. Besides audio and video classification, other domains also benefit from multi-modal data, such as lip reading [Afo+18; ACZ20], audio synthesis based on visual information [Gan+20; GM18; Koe+20; KWZ19; Nar+21; SLS20; Zho+19], and localisation and separation of sounds in videos [Afo+22; Afo+20; AZ18; Che+21a; GG19; OE18; Tia+18]. Recently, transformer models have gained popularity in audio-visual learning, e.g. for classification [BV19], event localization [LW20], dense video captioning [IR20], and text-based video retrieval [Gab+20; WZY21]. As shown in these works, transformers can effectively process

multi-modal input. Thus, our proposed framework fuses audio-visual information using a transformer-based mechanism.

FSL has been explored in the image domain [Che+19b; Dou+18; HG17; Li+19b; Liu+18; QBL18; RL17; Roy+22; SSZ17; Sun+18; Vin+16; Wan+18; Wan+19b; Ye+20] and in the video domain [BZP19; Cao+20; KC22; Xia+21; ZY18]. The popular meta-learning paradigm in FSL [BZP19; Cao+20; Li+19b; Liu+18; RL17; Sun+18; Vin+16; Wan+19b; Ye+20; ZY18] has been criticised by recent works [Che+19b; Kan+20; Wan+19b; Xia+21]. In the video domain, commonly a query and support set is used and each query sample is compared to all the support samples [BZP19; Cao+20; Per+21; ZY18]. The number of comparisons grows exponentially with the number of ways and shots. These methods become prohibitively expensive for GFSL, where models are evaluated on both the base and the novel classes. Hence, we focus on the non-meta learning approach in this work. Some non-meta learning approaches have addressed the more challenging and practical GFSL setting for videos [Kum+19; Xia+21] using unimodal visual data. In contrast, we propose to use multi-modal data in our novel (G)FSL benchmark for audio-visual video classification which provides the possibility to test a model in both scenarios (FSL and GFSL).

Feature generation. Due to the progress of generative models, such as GANs [AL18; GEB15; Goo+20; Iso+17; MO14] and diffusion models [Bla+22; Ess+21; Rom+22], different works have tried to adapt these systems to generate features as a data augmentation mechanism. GANs have been used in zero-shot learning (ZSL) and FSL [Kum+19; Nar+20; Xia+21; Xia+19] to increase the number and diversity of samples especially for unseen or novel classes. Diffusion models have also been applied to image generation in the feature space, such as [Rom+22; VKK21], but not in the ZSL or FSL setting. It is known that GANs are hard to optimize [SC21] while diffusion models appear to be more stable, leading to better results [DN21]. Therefore, our proposed framework uses a text-conditioned diffusion model to generate features for the novel classes in the FSL setting.

3.3 Audio-Visual (G)FSL Benchmark

We describe the audio-visual (G)FSL setting, present our proposed benchmark that we construct from audio-visual datasets, and explain the methods that we used to establish baselines for this task.

3.3.1 Audio-Visual (G)FSL Setting

We address the tasks of (G)FSL using audio-visual inputs. The aim of FSL is to recognise samples from classes that contain very few training samples, so-called *novel classes*. In addition, the goal of GFSL is to recognise both *base classes*, which contain a significant amount of samples, and novel classes.

Given an audio-visual dataset $\mathcal{V}$ with M samples and C classes, containing base and novel classes, we have $\mathcal{V} = \{\mathcal{X}_{a[i]}, \mathcal{X}_{v[i]}, y_{[i]}\}_{i=1}^{M}$, where $\mathcal{X}_{a[i]}$ represents the audio

input, $\mathcal{X}_{v[i]}$ the video input and $y_{[i]} \in \mathbb{R}^C$ the ground-truth class label. Both the audio and the video inputs contain temporal information. Two frozen, pretrained networks are used to extract features from the inputs, VGGish [Her+17] for the audio features $a_{[i]} = \{a_1, \ldots, a_t, \ldots, a_{F_a}\}_i$ and C3D [Tra+15] for video features $v_{[i]} = \{v_1, ..., v_t, ..., v_{F_v}\}_i$. We use these specific feature extractors to ensure that there is no leakage to the novel classes from classes seen when training the feature extractors (Sports1M [Kar+14] for the visual and Youtube-8M [Abu+16] for the audio modality), similar to [Mer+22b]. A potential leakage is harmful as it would artificially increase the performance and will not reflect the true performance.

All models are evaluated in the FSL and GFSL settings for k samples in the novel classes (called shots), with $k \in \{1, 5, 10, 20\}$. During inference, in the FSL setting, the class search space is composed only of the novel class labels and the samples belonging to these classes. In the GFSL setting, the search space contains both the novel and base class labels and their corresponding samples.

Meta-learning approaches commonly use the notion of episodes, where each episode only uses P novel classes randomly sampled from the total number of novel classes in a dataset, usually $P \in \{1, 5\}$ (coined P-way). However, similar to [Xia+21], we suggest to use higher values for P (e.g. all the classes in the dataset), so that the evaluation is closer to the real-world setting, as argued in [HG17; Xia+21]. In our proposed FSL setting, P corresponds to the total number of novel classes $P = N$, while for GFSL $P = C$. Our evaluation protocol is in line with [HG17].

3.3.2 Dataset Splits and Training Protocol

We provide training and evaluation protocols for audio-visual (G)FSL along with splits for UCF-FSL, ActivityNet-FSL and VGGSound-FSL. These are based on the UCF-101 [SZS12], ActivityNet [Cab+15] and VGGSound [Che+20a] datasets.

Our proposed training and evaluation protocol is similar to [HG17; Mer+22a; Mer+22b]. The training protocol is composed of two stages, indicated by subscripts $_{1,2}$. In the first stage, a model is trained on the training set $Train_1 = \mathcal{V}_{B_1} \cup \mathcal{V}_{N_1}$ where $\mathcal{V}_{B_1}$ consists of dataset samples from base classes, and $\mathcal{V}_{N_1}$ contains k samples for each of the classes N_1. The trained model is then evaluated on $Val = Val_B \cup Val_N$, where Val is the validation dataset which contains the same classes as $Train_1$. In the first stage, the hyperparameters for the network are determined, such as the number of training epochs and the learning rate scheduler parameters.

In the second stage, the model is retrained on the training set $Train_2$, using the hyperparameters determined in the first stage. Here, $Train_2 = \mathcal{V}_{B_2} \cup \mathcal{V}_{N_2}$ with $\mathcal{V}_{B_2} = Train_1 \cup Val$, and $\mathcal{V}_{N_2}$ contains k samples for the novel classes in the $Test$ set. The final model is evaluated on $Test = Test_B \cup Test_N$ with $Train_2 \cap Test = \emptyset$. With a small number of shots, e.g. $k = 1$, models risk a bias towards the novel samples in $Train_2$. To obtain robust evaluation results, the second stage is repeated three times with k randomly selected, but

	# **classes**				# **videos** *stage 1*				# **videos** *stage 2*			
	all	$\mathcal{V}_{B_1}$	$\mathcal{V}_{N_1}$	$\mathcal{V}_{N_2}$	$\mathcal{V}_{B_1}$	$\mathcal{V}_{N_1}$	Val_B	Val_N	$\mathcal{V}_{B_2}$	$\mathcal{V}_{N_2}$	$Test_B$	$Test_N$
(1)	271	138	69	64	70351	345	7817	2757	81270	320	9032	2880
(2)	48	30	12	6	3174	60	353	1407	4994	30	555	815
(3)	198	99	51	48	9204	255	1023	4052	14534	240	1615	3812

Table 3.1: Statistics for our VGGSound-FSL **(1)**, UCF-FSL **(2)**, and ActivityNet-FSL **(3)** benchmark datasets, showing the number of classes and videos in our proposed splits in the 5-shot setting. $\mathcal{V}_{B_1} \cup \mathcal{V}_{N_1}$ are used for training, Val_B and Val_N for validation in the first training stage. $\mathcal{V}_{B_2} \cup \mathcal{V}_{N_2}$ serves as training set in the second stage, and evaluation is done on $Test_B$ and $Test_N$.

fixed samples from $\mathcal{V}_{N_2}$. We provide dataset statistics in Table 3.1.

3.3.3 Benchmark Comparisons

To establish benchmark performances for audio-visual GFSL task, we adapt ten recent state-of-the-art methods for video FSL from visual information only, from audio-visual representation learning, and from audio-visual ZSL.

We provide results with several few-shot video recognition frameworks adapted to the multimodal audio-visual setting.

ProtoGan [Kum+19] uses GANs conditioned on the visual prototypes of classes that are obtained by averaging the features of all videos in that class. We adapt it to audio-visual inputs by concatenating the visual and audio features before passing them into the model.

SLDG [BLH20] is a multi-modal video FSL that uses video frames and optical flow as input. It weighs the frame features according to normal distributions. We replace the optical flow in [BLH20] with audio features.

TSL [Xia+21] is the current state-of-the-art video FSL which uses a GAN to generate synthetic samples for novel classes. It does not fully use temporal information, as the final score is the average of scores obtained on multiple short segments. We adapt it to the multi-modal setting by concatenating input features from the audio and visual modalities.

Moreover, we have adapted audio-visual representation learning methods to the few-shot task as can be seen below.

Perceiver[Jae+21], **Hierarchical Perceiver (HiP)** [Car+22], and **Attention Fusion** [FK20] are versatile video classification methods and we provide comparisons with them. We use the implementations of the adapted Perceiver and Attention Fusion frameworks provided by [Mer+22a] and we implement HiP in a similar way.

MBT [Nag+21] learns audio-visual representations for video recognition. It uses a transformer for each modality and these transformers can only exchange information using bottleneck attention.

Zorro[Rec+23], in contrast to MBT, uses two transformers that do not have access to the bottleneck attention. We adapt it by using a classifier on top of the averaged bottleneck

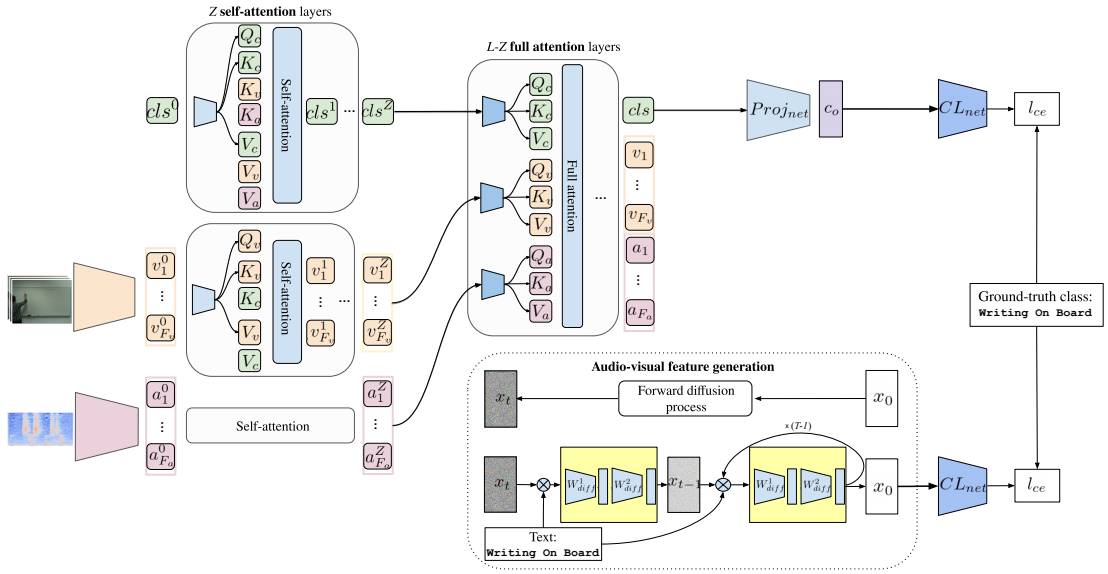

Figure 3.2: Our AV-DIFF model for audio-visual (G)FSL takes audio and visual features extracted from pre-trained audio and video classification models as inputs. During training, the features from both modalities are fused into a classification token, denoted by cls. At the same time, our diffusion model (bottom) generates additional synthetic features for the novel classes (denoted by x_0). Finally, we train our classifier CL_{net} (right) on fused real features c_o of both novel and base classes and synthetic features of novel classes. $\otimes$ is the concatenation operator.

attention tokens.

Finally, we have adapted the state-of-the-art methods in the audio-visual zero-shot learning domain, as shown below.

AVCA [Mer+22b] is an audio-visual ZSL method which uses temporally averaged features for the audio and visual modalities. We adapt it by using a classifier on the video output, which is the strongest of the two outputs in [Mer+22b].

TCAF [Mer+22a] is the state-of-the-art audio-visual ZSL method. It utilizes a transformer architecture with only cross-modal attention, leveraging temporal information in both modalities. As it does not use a classifier, TCAF outputs embeddings, and we determine the class by computing the distance to the semantic descriptors and selecting the closest one.

3.4 AV-DIFF Framework

In this section, we provide details for our proposed cross-modal AV-DIFF framework which employs cross-modal fusion (Section 3.4.1) and a diffusion model to generate audio-visual features (Section 3.4.2). Then, we describe the training curriculum in Section 3.4.3. Figure 3.2 illustrates AV-DIFF's full architecture.

3.4.1 Audio-Visual Fusion with Cross-Modal Attention

Audio-visual fusion. We project the audio $a_{[i]}$ and visual features $v_{[i]}$ to a shared embedding space. Then we use Fourier features [Tan+20b] as temporal positional embeddings and modality embeddings respectively and obtain positional aware video v_t^E and audio a_t^E tokens for timestep t. We prepend a classification token $cls^0 \in \mathbb{R}^{d_{dim}}$ to the audio and visual tokens. The output token cls corresponding to cls^0 is the final fused audio-visual representation which is input to $Proj_{net}$. Our audio-visual fusion mechanism contains L layers, which are based on multi-head attention [Vas+17] Att^l, followed by a feed forward function $\text{FF}^l : \mathbb{R}^{d_{dim}} \rightarrow \mathbb{R}^{d_{dim}}$. The input to the first layer is $x_{in}^1 = [cls^0, a_1^E, \cdots, a_{T_a}^E, v_1^E, \cdots, v_{T_v}^E]$. The output of a layer is:

$$x_{out}^l = \text{FF}^l(\text{Att}^l(x_{in}^l) + x_{in}^l) + \text{Att}^l(x_{in}^l) + x_{in}^l. \tag{3.1}$$

In the following, we describe the first layer of the audio-visual fusion. The other layers work similarly. Our input x_{in}^1 is projected to queries, keys and values with linear maps $s : \mathbb{R}^{d_{dim}} \rightarrow \mathbb{R}^{d_{dim}}$ for $s \in \{q, k, v\}$. The outputs of the projection are written as zero-padded query, key and value features. For the keys we get:

$$\mathbf{K}_c = [k(cls^0), 0, \cdots, 0], \tag{3.2}$$

$$\mathbf{K}_a = [0, \cdots, 0, k(a_1^E), \cdots, k(a_{F_a}^E), 0, \cdots, 0], \tag{3.3}$$

$$\mathbf{K}_v = [0, \cdots, 0, k(v_1^E), \cdots, k(v_{F_v}^E)]. \tag{3.4}$$

The final keys are obtained as $\mathbf{K} = \mathbf{K}_c + \mathbf{K}_a + \mathbf{K}_v$. The queries and values are obtained in a similar way. We define full attention as $\mathbf{A} = \mathbf{A}_c + \mathbf{A}_{cross} + \mathbf{A}_{self}$:

$$\begin{gathered} \mathbf{A}_c = \mathbf{Q}_c \mathbf{K}^T + \mathbf{K} \mathbf{Q}_c^T, \qquad \mathbf{A}_{cross} = \mathbf{Q}_a \mathbf{K}_v^T + \mathbf{Q}_v \mathbf{K}_a^T, \\ \mathbf{A}_{self} = \mathbf{Q}_a \mathbf{K}_a^T + \mathbf{Q}_v \mathbf{K}_v^T. \end{gathered} \tag{3.5}$$

The novelty in the attention mechanism in AV-DIFF is that it exploits a hybrid attention mechanism composed of two types of attention: within-modality self-attention and full-attention. The first Z layers use self-attention $\mathbf{A}_{self} + \mathbf{A}_c$, the subsequent $L - Z$ layers leverage full attention $\mathbf{A}$.

Audio-visual classification. We project cls to $\mathbb{R}^{d_{out}}$ by using a projection network, $c_o = Proj_{net}(cls)$. Then, we apply a classification layer to c_o, $logits = CL_{net}(c_o)$. Given the ground-truth labels gt, we use a cross-entropy loss, $L_{ce} = CE(logits, gt)$ to train the full architecture.

3.4.2 Text-Conditioned Feature Generation

AV-DIFF uses a diffusion process to generate audio-visual features which is based on the Denoising Diffusion Probabilistic Models (DDPM) [HJA20]. In particular, we condition the generation of features for novel classes on a conditioning signal, such as the word

embedding (e.g. word2vec [Mik+13]) of a class name. The diffusion framework consists of a forward process and a reverse process.

The forward process adds noise to the data sample x_0 for T timesteps:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) = \prod_{t=1}^{T} \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}\right), \tag{3.6}$$

where $\beta_1, \ldots, \beta_T$ is the variance schedule.

As the **reverse process** $q(x_{t-1}|x_t)$ is intractable, we approximate it with a parameterised model p_θ:

$$p_\theta(x_{0:T}) = p_\theta(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) = p_\theta(x_T) \prod_{t=1}^{T} \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{3.7}$$

We condition the model on the timestep t and the class label embedding w,

$$L_{\text{diff},w} = E_{x_0,t,w,\epsilon}[||\epsilon - \epsilon_\theta(\sqrt{\bar{a}_t}x_0 + \sqrt{1-\bar{a}_t}\epsilon, w, t)||^2], \tag{3.8}$$

where ϵ is the noise added at each timestep and ϵ_θ is a model that predicts this noise. The sample at timestep $t-1$ is obtained from timestep t as:

$$p_\theta(x_{t-1}|x_t, w) = \mathcal{N}(x_{t-1}; \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, w, t)), \sigma_t^2 \mathcal{I}). \tag{3.9}$$

The input to ϵ_θ at timestep t is obtained by concatenating x_t, w, and t. We optimize $L_{\text{diff},w}$ to learn p_θ.

3.4.3 Training Curriculum and Evaluation

Each training stage (explained in Section 3.3.2) is split into two substages. In the first substage, we train the full architecture (the fusion mechanism, the diffusion model, $Proj_{net}$ and the classifier CL_{net}) on base classes $\mathcal{V}_{B_1}$ (or $\mathcal{V}_{B_2}$ in the second stage) by minimizing $L_{ce} + L_{\text{diff},w}$. The classifier CL_{net} is trained only on real features for the base classes in $\mathcal{V}_{B_1}$ (or $\mathcal{V}_{B_2}$ for the second stage) in the first substage.

During the second substage, we freeze the fusion mechanism and continue to train the diffusion model, $Proj_{net}$ and CL_{net} with the same training objective $L_{ce} + L_{\text{diff},w}$. Here we consider both base and novel classes $\mathcal{V}_{B_1}$ and $\mathcal{V}_{N_1}$ classes (or $\mathcal{V}_{B_2}$ and $\mathcal{V}_{N_2}$ in the second stage), unlike in the first substage where we only used base classes. For each batch composed of real samples from novel classes, we generate a corresponding batch of the same size with synthetic samples using our diffusion model. CL_{net} is then trained on real features from $\mathcal{V}_{B_1}$ (or $\mathcal{V}_{B_2}$ in the second stage) and on real and synthetic features for the classes in $\mathcal{V}_{N_1}$ (or $\mathcal{V}_{N_2}$ in the second stage). Freezing the audio-visual transformer ensures that its fusion mechanism does not overfit to the few samples from the novel classes.

The diffusion model is not used for inference, and the output of the classifier CL_{net} for c_0 provides the predicted score for each class (including the novel classes). The class with the highest score is selected as the predicted class.

Model ↓	VGGSound-FSL						UCF-FSL						ActivityNet-FSL					
	1-shot		*5-shot*		*10-shot*		*1-shot*		*5-shot*		*10-shot*		*1-shot*		*5-shot*		*10-shot*	
	HM	FSL	HM	FSL	HM	FSL	HM	FSL	HM	FSL	HM	FSL	HM	FSL	HM	FSL	HM	FSL
Att. Fusion [FK20]	15.46	16.37	28.22	31.57	30.73	39.02	37.39	36.88	51.68	47.18	57.91	52.19	4.35	5.82	6.17	8.13	10.67	10.78
Perceiver [Jae+21]	17.97	18.51	29.92	33.58	33.65	40.73	44.12	33.73	48.60	40.47	55.33	47.86	17.34	12.53	25.75	21.50	29.88	26.46
MBT [Nag+21]	14.70	21.96	27.26	34.95	30.12	38.93	39.65	27.99	46.55	34.53	50.04	39.73	14.26	12.63	23.26	22.38	26.86	26.03
TCAF [Mer+22a]	19.54	20.01	26.09	32.22	28.95	36.43	44.61	35.90	46.29	37.39	54.19	47.61	16.50	13.01	22.79	21.81	24.78	23.33
ProtoGan [Kum+19]	10.74	14.08	25.17	28.87	29.85	34.80	37.95	28.08	42.42	33.63	51.01	40.68	2.77	4.40	2.67	7.81	4.05	8.81
SLDG [BLH20]	16.83	17.57	20.79	25.17	24.11	29.48	39.92	28.91	36.47	28.56	34.31	26.96	13.57	10.30	22.29	19.16	27.81	25.35
TSL [Xia+21]	18.73	22.44	19.49	29.50	21.93	31.29	44.51	35.17	51.08	42.42	60.93	55.63	9.53	10.77	10.97	12.77	10.39	12.18
HiP [Car+22]	19.27	18.64	26.82	30.67	29.25	35.13	21.79	34.88	36.44	42.23	50.69	43.29	13.80	10.31	18.10	16.25	19.37	17.06
Zorro [Rec+23]	18.88	21.79	29.56	35.17	32.06	40.66	44.35	34.52	51.86	42.59	58.89	49.06	14.56	11.94	23.14	21.94	27.35	26.33
AVCA [Mer+22b]	6.29	10.29	15.98	20.50	18.08	28.27	43.61	31.24	49.19	36.70	50.53	39.17	12.83	12.22	20.09	21.65	26.02	26.76
AV-DIFF	**20.31**	**22.95**	**31.19**	**36.56**	**33.99**	**41.39**	**51.50**	**39.89**	**59.96**	**51.45**	**64.18**	**57.39**	**18.47**	**13.80**	**26.96**	**23.00**	**30.86**	**27.81**

Table 3.2: **Our benchmark study for audio-visual (G)FSL**: 1,5,10-shot performance of our AV-DIFF and compared methods on (G)FSL. The harmonic mean (HM) of the mean class accuracies for base and novel classes are reported for GFSL. For the FSL performance, only the test subset of the novel classes is considered. Base, novel, and 20-shots performances are included in the suppl. material.

3.5 Experiments

In this section, we first provide the implementation details for obtaining the presented results (Section 3.5.1). We then report results for our proposed AV-DIFF in our benchmark study (Section 3.5.2). Finally, we analyse the impact of different components of AV-DIFF (Section 3.5.3).

3.5.1 Implementation Details

AV-DIFF uses features extracted from pre-trained audio and visual classification networks as inputs (details provided in the suppl. material). AV-DIFF is trained using $d_{dim} = 300$ and $d_{out} = 64$. Our fusion network has $L = 5, 4, 8$ transformer layers, the layer after which the attention changes is set to $Z = 3, 2, 5$ on ActivityNet-FSL, UCF-FSL and VGGSound-FSL respectively. We train all models on a single NVIDIA RTX 2080-Ti GPU. The first substage uses 30 epochs while the second one uses 20 epochs. We use the Adam optimizer [KB14], and $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of $1e^{-5}$. We use a learning rate of $7e^{-5}$ for UCF-FSL and ActivityNet-FSL, and $6e^{-5}$ for VGGSound-FSL. For ActivityNet-FSL and UCF-FSL, we use a scheduler that reduces the learning rate by a factor of 0.1 when the performance has not improved for 3 epochs. We use a batch size of 32 for ActivityNet-FSL, and 64 for UCF-FSL and VGGSound-FSL. Each epoch consists of 300 batches. As ActivityNet-FSL has very long videos, we randomly trim the number of features during training to 60. During evaluation, we also trim the videos to a maximum length of 300 features, and the trimmed features are centered in the middle of the video. To reduce the bias towards base classes, we use calibrated stacking [Cha+16] on the search space composed of the interval [0,1] with a step size of 0.1. This value is obtained on the validation dataset.

3.5.2 Audio-Visual GFSL Performance

For each of the models featuring in our benchmark, we report results for three different numbers of shots, i.e. 1-shot, 5-shot, 10-shot on all three datasets in Table 3.2. AV-DIFF outperforms all the methods across all shots and datasets for few-shot learning (FSL) and generalised few-shot learning (HM).

For 1-shot, AV-DIFF achieves a HM/FSL of 20.31%/22.95% vs. HM of 19.54% for TCAF and FSL score of 22.44% for TSL on VGGSound-FSL. On 5-shot, our model obtains a HM/FSL of 31.19%/36.56% vs. 29.92% for the Perceiver and FSL of 35.17% for Zorro. Furthermore, AV-DIFF yields slightly better results than the Perceiver in both HM and FSL for 10 shots, with HM/FSL of 33.99%/41.39% vs. 33.65%/40.73% for the Perceiver. Thus, combining our hybrid attention and the diffusion model is superior to systems that rely solely on powerful attention mechanisms without incorporating generative modeling (Perceiver, TCAF) and systems that incorporate generative modelling, but that do not employ powerful attention mechanisms (TSL, ProtoGan).

Similar trends are observed on UCF-FSL, while on ActivityNet-FSL, the ranking of methods changes dramatically. Methods that perform well on UCF-FSL and VGGSound-FSL, but which do not fully use the temporal information (e.g. Attention Fusion, ProtoGan and TSL) perform weakly on ActivityNet-FSL which contains videos with varying length, including some very long videos, making the setting more challenging. Our AV-DIFF can process temporal information effectively, resulting in robust state-of-the-art results on ActivityNet-FSL.

Interestingly, VGGSound-FSL contains the most classes among the datasets considered, resulting in a significantly lower N (suppl. material, Tab. 1) than FSL. This also lowers the HM (computed from B, N). On VGGSound-FSL, methods tend to be biased towards novel classes (N $\geq$ B) due to calibration [Cha+16]. In this case, HM $\leq$ N $\leq$ FSL. Moreover, some baselines that were also used in audio-visual zero-shot learning [Mer+22a; Mer+22b] (e.g. TCAF) exhibit significant increases in performance even in the 1-shot setting. This is expected as for 1-shot learning, one training example is used from each novel class. This reduces the bias towards base classes, leading to more balanced B and N scores, and thereby better HM and FSL results. Base, novel, and 20-shot performances are included in the suppl. material.

3.5.3 AV-DIFF Model Ablations

Here, we analyse the benefits of the main components of AV-DIFF, i.e. our proposed audio-visual fusion mechanism, and the diffusion model for feature generation. Furthermore, we analyse the importance of using multiple modalities, and the effect of different semantic representations.

Audio-visual fusion mechanism. Table 3.3 ablates our cross-modal fusion mechanism for generating rich audio-visual representations. As shown in Section 3.4.1, AV-DIFF uses two types of attention: $\mathbf{A}_{self}$+$\mathbf{A}_c$ for the first few layers and $\mathbf{A}$ for the later layers. For

Model ↓	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
A	28.56	31.52	29.98	36.55	78.95	42.07	54.90	43.75	23.10	22.06	22.57	22.53
$\mathbf{A}_{cross} + \mathbf{A}_c$	28.44	32.48	30.33	36.85	82.89	44.33	57.77	47.02	27.02	21.25	23.79	21.98
$\mathbf{A}_{self} + \mathbf{A}_c$	26.68	33.23	29.60	**37.06**	50.10	44.58	47.18	45.03	31.61	21.48	25.58	22.65
Alternate AV-DIFF	27.40	32.60	29.78	36.82	80.25	43.01	56.00	45.81	31.15	21.57	25.49	22.59
AV-DIFF	30.88	31.50	**31.19**	36.56	74.11	50.35	**59.96**	**51.45**	35.84	21.61	**26.96**	**23.00**

Table 3.3: Impact of different audio-visual fusion mechanisms in the 5-shot setting.

Model ↓	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
AV-GAN	27.80	31.75	29.64	36.53	83.79	36.20	50.56	37.33	35.12	19.53	25.10	21.35
AV-DIFF	30.88	31.50	**31.19**	**36.56**	74.11	50.35	**59.96**	**51.45**	35.84	21.61	**26.96**	**23.00**

Table 3.4: Influence of using different feature generators in the 5-shot setting.

Alternate AV-DIFF, we alternate the two types of attention used in AV-DIFF in subsequent layers. We also show our model with $\mathbf{A}_{cross}$+$\mathbf{A}_c$ which is the same attention used by the SOTA audio-visual GZSL framework [Mer+22a]. On ActivityNet-FSL, AV-DIFF obtains a HM/FSL of 26.96%/23.00% vs. 25.58%/22.65% for $\mathbf{A}_{self}$+$\mathbf{A}_c$. The same trend is seen on UCF-FSL. On VGGSound-FSL we outperform *Alternate AV-DIFF* on HM, but are slightly weaker than $\mathbf{A}_{self}$+$\mathbf{A}_c$ in FSL. Overall, our fusion mechanism is the best across both metrics and datasets.

Feature generation model. In Table 3.4, we investigate the impact of different generative models to produce audio-visual features for the novel classes. We compare the diffusion model in AV-DIFF to a GAN similar to the one used by TSL [Xia+21], which optimizes a Wasserstein GAN loss [ACB17]. On ActivityNet-FSL, we observe that AV-DIFF outperforms the GAN variant, with a HM/FSL of 26.96%/23.00% vs. 25.10%/21.35% for the GAN. The same can be seen on UCF-FSL and VGGSound-FSL. This shows that our generative diffusion model is better suited for audio-visual GFSL than a GAN.

Multi-modal input. We explore the impact of using multi-modal inputs for AV-DIFF in Table 3.5. For unimodal inputs, we adapt AV-DIFF to only employ full attention which is identical to self-attention in this case. On ActivityNet-FSL, using multi-modal inputs provides a significant boost in performance compared to unimodal inputs, with a HM/FSL of 26.96%/23.00% vs. 19.01%/17.84% when using only visual information. The same trend can be observed on UCF-FSL. In contrast, on VGGSound-FSL, using multi-modal inputs gives stronger GFSL but slightly weaker results in FSL than using the audio modality. This might be due to the focus on the audio modality in the data curation process for VGGSound. As a result, significant portions of the visual information can be unrelated to the labelled class. Overall, the use of multi-modal inputs from the audio and visual modalities significantly boosts the (G)FSL performance for AV-DIFF.

However, one interesting aspect is that using both modalities leads to better *B* and

Model ↓	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
Audio	28.30	30.56	29.39	**36.64**	55.31	39.18	45.87	44.44	13.74	15.23	14.45	17.58
Visual	7.83	8.92	8.35	9.51	67.13	30.70	42.14	30.98	20.80	17.49	19.01	17.84
AV-DIFF	30.88	31.50	**31.19**	36.56	74.11	50.35	**59.96**	**51.45**	35.84	21.61	**26.96**	**23.00**

Table 3.5: Influence of using multi-modal input in the 5-shot setting.

Model ↓	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
AV-DIFF av_{prot}	25.74	33.00	28.92	35.76	83.38	42.46	56.26	44.78	32.22	21.50	25.79	22.73
AV-DIFF	30.88	31.50	**31.19**	**36.56**	74.11	50.35	**59.96**	**51.45**	35.84	21.61	**26.96**	**23.00**

Table 3.6: Influence of different semantic class representations in the 5-shot setting.

N performances across all three datasets. For example, on ActivityNet-FSL, AV-DIFF obtains a B score of 35.84% and an N score of 21.61% compared to 20.80% and 17.49% when using only the visual modality. On UCF-FSL, AV-DIFF achieves a score of 74.11% for B and 50.35% for N compared to 67.13% and 39.18% for the visual and audio modalities respectively. Finally, on VGGSound-FSL, AV-DIFF achieves a B score of 30.88% and an N score of 31.50% compared to 28.30% and 30.56% for unimodal audio inputs. This shows that using multi-modal inputs decreases the bias towards either of the metrics, leading to a more robust and balanced system.

Semantic class representations. We consider using different semantic class representations in Table 3.6. In FSL, the most common semantic descriptor is word2vec [Mik+13] which is used to condition the audio-visual feature generation in AV-DIFF. However, related works (e.g. ProtoGan [Kum+19]), use prototypes which average the visual features of all the training videos in a class to obtain the semantic representation of that class. In the multi-modal setting, we can concatenate the audio and visual prototypes to obtain multi-modal prototypes av_{prot} which is used as a conditioning signal for our diffusion model. On ActivityNet-FSL, using word2vec embeddings leads to better results than using the audio-visual prototypes av_{prot}, with a HM/FSL of 26.96%/23.00% vs. 25.79%/22.73% for av_{prot}. The same can be seen on UCF-FSL and VGGSound-FSL, demonstrating that the word2vec embeddings provide a more effective conditioning signal.

3.6 Conclusion

In this work, we propose an audio-visual (generalised) few-shot learning benchmark for video classification. Our benchmark includes training and evaluation protocols on three datasets, namely VGGSound-FSL, UCF-FSL and ActivityNet-FSL, and baseline performances for ten state-of-the-art methods adapted from different fields. Moreover, we propose AV-DIFF which fuses multi-modal information with a hybrid attention

mechanism and uses a text-conditioned diffusion model to generate features for novel classes. AV-DIFF outperforms all related methods on the new benchmark. Finally, we provided extensive model ablations to show the benefits of our model's components. We hope that our benchmark will enable significant progress for audio-visual generalised few-shot learning.

4

Video-Adverb Retrieval with Compositional Adverb-Action Embeddings

Retrieving adverbs that describe an action in a video poses a crucial step towards fine-grained video understanding. We propose a framework for video-to-adverb retrieval (and vice versa) that aligns video embeddings with their matching compositional adverb-action text embedding in a joint embedding space. The compositional adverb-action text embedding is learned using a residual gating mechanism, along with a novel training objective consisting of triplet losses and a regression target. Our method achieves state-of-the-art performance on five recent benchmarks for video-adverb retrieval. Furthermore, we introduce dataset splits to benchmark video-adverb retrieval for unseen adverb-action compositions on subsets of the MSR-VTT Adverbs and ActivityNet Adverbs datasets. Our proposed framework outperforms all prior works for the generalisation task of retrieving adverbs from videos for unseen adverb-action compositions. Code and dataset splits are available at https://hummelth.github.io/ReGaDa/.

4.1 Introduction

Fine-grained video understanding is concerned with the detailed analysis of video content beyond action recognition. This is relevant for improving and potentially accelerating video search and retrieval. While there has been significant progress in action retrieval and recognition in videos [Alh+21; Mie+20; Pat+21; Tan+20a], the fine-grained understanding of actions remains challenging. In particular, it can be useful to perceive how an action is performed in order to better understand the action itself [Dou+20; DS22; Mol+23]. For instance, in addition to recognising the action *cutting*, it is useful to understand details about the execution of an action, e.g. *cutting slowly*. Specifically, we consider the bidirectional video-adverb retrieval task where we retrieve adverbs that match an action in a video and vice versa.

For bidirectional video-adverb retrieval, adverbs and action words can be combined in

a compositional manner. The same adverb can describe multiple actions, such as *cutting slowly* or *dancing slowly*. The compositional nature of the adverb-action pairings can also be exploited when learning adverb-action representations. Our proposed ReGAdA framework for video-adverb retrieval uses a **re**sidual **g**ating mechanism to compose **ad**verb-**a**ction (ReGAdA) representations for retrieval.

At its core, our framework learns to align adverb representations and video representations in a shared embedding space using a novel training objective which consists of a direct regression loss between the adverb and video representations and triplet losses. To obtain the adverb representation, the adverb and action are jointly embedded using a residual gating mechanism, which we adapted to the video-adverb retrieval task from [Vo+19]. It models the composition as a transformation of the adverb embedding based on the action by using a gate and a residual mechanism. The gate facilitates the preservation of meaningful information from the adverb embeddings based on the adverb-action composition. Our final composition is learned as a residual combination on top of the gated adverb embeddings. This allows our composed embeddings to be in the same "feature space" as the original adverb embeddings. Similar to previous works for this task, our model assumes knowledge of the ground-truth action class to perform video-adverb retrieval.

The compositional adverb-action embeddings and our proposed training objective prove beneficial for the retrieval performance, specifically for the retrieval of unseen adverb-action compositions. ReGAdA obtains state-of-the-art results on the five video-adverb retrieval benchmarks HowTo100M Adverbs [Dou+20; Mie+19], VATEX Adverbs [DS22; Wan+19a], ActivityNet Adverbs [Cab+15; DS22], MSR-VTT Adverbs [DS22; Xu+16], and Adverbs in Recipes [Mie+19; Mol+23]. Furthermore, we propose two additional splits for benchmarking the retrieval of unseen adverb-action compositions on the ActivityNet Adverbs and MSR-VTT Adverbs datasets.

To summarise, we make the following contributions: 1) Our proposed method for video-adverb retrieval uses a text encoder based on a gated residual mechanism and a novel training objective. 2) We evaluate ReGAdA on the challenging unseen video-adverb retrieval task and introduce new benchmark splits, compliant with zero-shot learning principles, for the retrieval of unseen adverb-action compositions based on the ActivityNet Adverbs and MSR-VTT Adverbs datasets. 3) Our framework outperforms prior work for both the seen and the unseen adverb-action composition retrieval tasks.

4.2 Related Work

Fine-grained action understanding in video retrieval. Early works for video understanding extended retrieval approaches for images to videos, by temporally aggregating frames in a video [DLS18; Ota+16; TTS16; Xu+15b]. With the availability of large video-text datasets [Ann+17; Bai+21; Kri+17; Mie+19; Onc+21; Wan+19a; Xu+16; ZXC18], different methods focused on sentence disambiguation [Che+19a; Wra+19], self-supervision

[Ala+20; Rou+21; ZY20], weakly supervised learning [Mie+20; Mie+19; Pat+21], multiple embedding experts [Gab+20; Liu+19a; MLS18], or the use of large pre-trained models [Lei+21; Luo+22; Par+22; Wu+23]. Video-action retrieval specifically aims at retrieving videos based on an action, e.g. using a verb to describe the same [HSR19; WD19]. Moreover, [Che+20b; Ge+22; Wra+19; Xu+15b; Zhu+19b] use nouns in addition to verbs for video-text retrieval. In a more general setting, [Mom+23] recently proposed to use a large language model to generate modified captions to improve verb understanding in video-language models. Different to these methods, we focus on adverbs in the video-adverb retrieval task.

Video-adverb retrieval. The video-adverb retrieval task was introduced by [Dou+20] along with the HowTo100M Adverbs dataset. [Dou+20] learns a shared representation between videos and adverbs, modelling adverb information as learned linear transformations on action class label word embeddings, similar to [NG18] for object attributes. Unlike [Dou+20], we choose to utilise semantic information from adverb embeddings in addition to action embeddings for modelling adverb-action compositions. [DS22] extends [Dou+20] to the low-data regime with pseudo-labelling. The recently proposed [Mol+23] tackles the task either as a classification or regression problem. Its video encoder builds on [Dou+20] with an additional projection following the attention while keeping the text representations frozen. The classification variant is trained with a cross-entropy loss for adverb classification, while the regression variant uses a regression target describing the change an adverb induced in an action embedding. Different from [Mol+23], we aim at learning the adverb-action representations and the video representations in a shared embedding space. Formulating the task as an alignment problem in a shared embedding space combined with compositional adverb-action representations significantly boosts the performance for video-adverb retrieval.

Learning with object attributes. Approaches for learning object-attribute pairs from images can be broadly categorized into classification [Li+20b; Man+21; MGH17; Nae+21; NG18] and retrieval approaches [Bor+13; CG14; ILA15; Nan+19; Vo+19; WJ13; WM10]. Our adverb-action compositions are most closely related to [Vo+19], which proposed a residual gating mechanism for learning compositional image-text embeddings. This mechanism proved particularly useful for retrieving images using both an image and a text query, the text describing a desired modification of the query image. We adapt a similar residual gating mechanism for learning compositional adverb-action embeddings by aligning the composition with action-focused video embeddings.

4.3 REGADA Framework for Video-Adverb Retrieval

In this section, we provide details about our proposed REGADA framework for video-adverb retrieval which is visualised in Fig. 4.1. We first describe the video-adverb retrieval task, and then provide details about our framework. Finally, we detail our training objective and the inference procedure for retrieval.

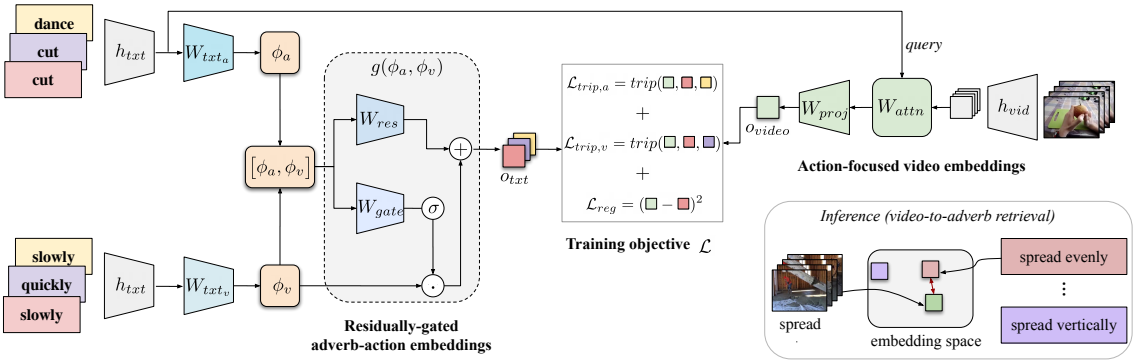

Figure 4.1: **Overview of our REGADA framework for video-adverb retrieval.** Our framework composes adverb-action embeddings with a gated residual between the adverbs ϕ_v and the concatenated action and adverb embeddings $[\phi_a, \phi_v]$. The training objective $\mathcal{L}$ aligns the learned text and video representations in a joint embedding space. For test time inference, outputs are obtained based on similarity in the embedding space.

Task setting and dataset. The adverb-to-video retrieval task aims at retrieving matching videos from a pool of videos for a given adverb. Similarly, for the video-to-adverb retrieval task, given a video, the aim is to retrieve the adverb that best describes the action depicted in the video from a pool of pre-set adverbs. We denote a dataset with N samples, A action classes and V adverb classes by $\mathcal{D} = \{X_{[i]}, y_{[i]}\}_{i=1}^{N}$, consisting of video data $X_{[i]}$, and ground-truth action and adverb labels $y_{[i]} = \{a_{[i]}, v_{[i]}\}$ with one-hot encodings for the action $a_{[i]} \in \mathbb{R}^A$ and adverb $v_{[i]} \in \mathbb{R}^V$. We define the sets of possible actions and adverbs as $\mathcal{A}$ and $\mathcal{V}$. The set of all possible adverb-action combinations is $\mathcal{C} = \mathcal{V} \times \mathcal{A}$.

Our REGADA framework learns to align video and adverb-action representations in a joint embedding space. It generates compositional textual representations for adverb-action pairs using a text encoder. Additionally, the visual information is processed in a video encoder to obtain visual representations that contain information about the adverb associated with a given action. In the following, we describe how we obtain class label embeddings for the actions and adverbs, and how the video and text encoders process the video features and class label embeddings.

Residually-gated adverb-action embeddings. We obtain word embeddings for the action $a \in \mathcal{A}$ and for the adverb $v \in \mathcal{V}$ from a pre-trained language encoder h_{txt}, giving $\theta_v = h_{txt}(v)$, and $\theta_a = h_{txt}(a)$ with $\theta_a, \theta_v \in \mathbb{R}^{d_\theta}$. We then apply two linear maps $W_{txt_a}, W_{txt_v} : \mathbb{R}^{d_\theta} \rightarrow \mathbb{R}^{d_{dim}}$, such that $\phi_a = W_{txt_a}(\theta_a)$ and $\phi_v = W_{txt_v}(\theta_v)$. The action and adverb embeddings are then further processed jointly in our text encoder. Additionally, the action word embedding θ_a serves as a query vector in the video encoder's attention for generating an action-focused video embedding.

Our text encoder uses a residual gating mechanism which is based on [Vo+19]. Given ϕ_a and ϕ_{v_j} as inputs, the output of the text encoder is defined as:

$$o_{txt_j} = g(\phi_a, \phi_{v_j}) = \omega_g * \sigma(W_{gate}(\phi_a, \phi_{v_j})) \odot \phi_{v_j} + \omega_r * W^{res}(\phi_a, \phi_{v_j}), \tag{4.1}$$

where $j \in \{1, \cdots, V\}$, ω_g, ω_r are learnable scalar weights for balancing the gating mechanism and the residual, $\odot$ is an element-wise product, and σ the sigmoid function. W_{res} and W_{gate} are modelled using MLPs with N_r and N_g layers respectively. For those, the input consisting of adverb and action embeddings, is first passed through a concatenation operator and batch normalisation [IS15] is applied. The subsequent layers consist of a linear map followed by dropout [Sri+14] with probability $drop_g$ and a Leaky ReLU [Xu+15a]. The final layer is a linear projection to $\mathbb{R}^{d_{dim}}$.

We tackle video-adverb retrieval by aligning text and videos in a learned shared embedding space. Our residual gating mechanism models the composition as a transformation of the adverb embedding based on the action. The gating mechanism thereby allows to retain information from adverbs when actions do not provide useful semantic information.

Action-focused video embeddings. A pre-trained video classification network h_{vid} is used to extract a sequence of visual features $\boldsymbol{x}_{[i]} = \{x_1, ..., x_t, ..., x_T\}_i$, where $\boldsymbol{x}_{[i]} = h_{vid}(\mathcal{X}_{[i]})$ and $x_t \in \mathbb{R}^{d_x}$. We use T to denote the number of temporal segments in a video clip.

Given a sequence of video features $\boldsymbol{x}_{[i]}$ and its associated action word embedding $\theta_{a_{[i]}}$ (for easier readability, we omit the subscripts $_{[i]}$), we obtain action-focused video embeddings using a similar mechanism as the one proposed in [Dou+20]. The video embeddings are obtained using weak action-level ground-truth in the multi-head attention mechanism [Vas+17]. The action word embedding θ_a serves as the query in the attention to focus on parts of the video that are relevant to the given action, and ignore the temporal segments that may be relevant to other actions.

For the multi-head attention, we map the video features $\{x_t\}_{t\in[1,T]}$ to keys and values using linear mappings $W_k : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_{head_x} H_x}$, $W_v : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_{head_x} H_x}$ with H_x heads and a dimension of d_{head_x} per head. We also map the action word embeddings θ_a to queries with $W_q : \mathbb{R}^{d_\theta} \rightarrow \mathbb{R}^{d_{head_x} H_x}$. For each attention head j, we have

$$p^j_{attn} = g^{DL}_{attn}\left(softmax\left(\frac{W^j_q(\theta_a)^T W^j_k(\boldsymbol{x})}{\sqrt{d_{head_x}}}\right)\right) W^j_v(\boldsymbol{x}), \tag{4.2}$$

where g^{DL}_{attn} denotes dropout with probability $drop_{attn}$.

We apply a linear mapping $W_{attn} : \mathbb{R}^{d_{head_x} H_x} \rightarrow \mathbb{R}^{d_{dim}}$ to aggregate the per-head attention giving the output video embedding $o_{attn} = W_{attn}([p^1_{attn}, \cdots, p^H_{attn}])$. The final output is obtained with an MLP, $W_{proj} : \mathbb{R}^{d_{dim}} \rightarrow \mathbb{R}^{d_{dim}}$,

$$o_{video} = W_{proj}(o_{attn}), \tag{4.3}$$

where each of the N_{proj} layers of W_{proj} consists of a linear layer $W^l_{proj} : \mathbb{R}^{d_{dim}} \rightarrow \mathbb{R}^{d_{dim}}$, layer normalisation [BKH16] g^{LN}_{proj}, ReLU [NH10] g^{ReLU}_{proj}, and dropout g^{DL}_{proj} with probability $drop_{proj}$.

Training objectives. Our REGADA framework is trained with triplet losses based on [Dou+20] and with a direct regression loss between the video and text embeddings. We consider the triplet loss function $trip(a, p, n) = max(0, \|a - p\|_2 - \|a - n\|_2 + \mu)$, with the

anchor embedding a, the embeddings for the positive and negative samples p and n, and the margin μ. The **action triplet loss** encourages the alignment of the video representation o_{video} and text embeddings with the matching action as opposed to a sampled negative action $\phi_{\bar{a}}$. For this, we use the video embedding o_{video} as the anchor, the text embedding with ground truth action ϕ_a and adverb ϕ_v as the positive sample, and the text embedding of the same adverb but different action $\phi_{\bar{a}_i}$ as a negative:

$$\mathcal{L}_{trip,a} = \frac{1}{n}\sum_{i=1}^{n} trip(o_{video_i}, g(\phi_{a_i}, \phi_{v_i}), g(\phi_{\bar{a}_i}, \phi_{v_i})) \quad \text{for } \phi_{\bar{a}_i} \neq \phi_{a_i}. \tag{4.4}$$

We use an **adverb triplet loss** to push text embeddings containing the adverb antonym $\phi_{\bar{v}}$ away from the ground-truth text embedding:

$$\mathcal{L}_{trip,v} = \frac{1}{n}\sum_{i=1}^{n} trip(o_{video_i}, g(\phi_{a_i}, \phi_{v_i}), g(\phi_{a_i}, \phi_{\bar{v}_i})). \tag{4.5}$$

By restricting the negative samples for adverbs to their antonyms, the loss does not punish potential ambiguities of actions in videos (e.g. a drawer being opened slowly can at the same time be opened partially but not quickly). Our **regression loss** directly minimises the distance between the output video and text embeddings:

$$\mathcal{L}_{reg} = \frac{1}{n}\sum_{i=1}^{n}(o_{video_i} - g(\phi_{a_i}, \phi_{v_i}))^2. \tag{4.6}$$

The final loss is computed as the weighted sum of the above losses according to

$$\mathcal{L} = \lambda_a * \mathcal{L}_{trip,a} + \lambda_v * \mathcal{L}_{trip,v} + \lambda_{reg} * \mathcal{L}_{reg}, \tag{4.7}$$

with hyperparameters $\lambda_a, \lambda_v, \lambda_{reg} \in \mathbb{R}$.

Retrieving adverbs and videos (inference). Similar to [Dou+20], we evaluate our method on adverb-to-video and video-to-adverb retrieval given the ground-truth action a. For video-to-adverb retrieval, given a video $\boldsymbol{x}$ and action query a, we embed the video to obtain o_{video}, and we obtain embeddings for j adverb-action combinations o_{txt_j} for $j \in \{1, \cdots, V\}$. Using the cosine similarity metric we rank all the text embeddings o_{txt_j} by their similarity to the query video embedding o_{video} and we consider the highest-ranked pair as the retrieved adverb.

For adverb-to-video retrieval, given an adverb v and action a that are embedded to o_{txt}, we define the set of test videos containing action a as Γ. We rank all video embeddings o_{video_j} for videos in Γ using the similarity computed between each o_{video_j} and o_{txt} and select the video which is closest to o_{txt}.

4.4 Video-Adverb Retrieval Benchmarks

In this section, we provide details about the datasets used in our experiments. In particular, we use five datasets for video-adverb retrieval. Furthermore, we propose two new dataset splits for the task of retrieving adverbs from videos for unseen adverb-action compositions.

Video-adverb retrieval datasets. HowTo100M Adverbs [Dou+20] consists of 5,824 video clips with annotations for 6 adverbs and 72 actions. In the following, we refer to HowTo100M Adverbs as **HowTo100M**. The recently proposed **Adverbs in Recipes** dataset has 10 adverbs, 48 actions and 7,003 videos. VATEX Adverbs [DS22] dataset has, with 34 adverbs and 135 actions, the largest variety of annotated adverbs and actions, consisting of 14,617 videos. We refer to VATEX Adverbs as **VATEX**. ActivityNet Adverbs [DS22] consists of 3,099 videos with 20 adverbs and 114 actions. We refer to it as **ActivityNet**. MSR-VTT Adverbs [DS22] is made up of 1,824 videos with 18 adverbs and 106 actions. In the following, we call this dataset **MSR-VTT**.

Unseen adverb-action compositions splits. We strive to explore the ability to recognise adverbs for novel adverb-action combinations. [DS22] proposed a dataset split for unseen compositions at test time for the VATEX dataset. Using the available videos in VATEX from [Mol+23], we replicate this split for the S3D video and text features used in this work, by omitting unavailable videos. We additionally propose new splits for unseen compositions on the ActivityNet and MSR-VTT datasets. We exclude HowTo100M Adverbs and Adverbs in Recipes, as both are subsets of HowTo100M which was used for pre-training the text and S3D video model. Hence, this would not comply with zero-shot learning principles.

To create splits for ActivityNet and MSR-VTT, we follow the protocol in [DS22]: We first split the set of possible adverb-action compositions into two non-overlapping sets, so that all adverbs and all actions are present in both sets, but individual compositions are only contained in one of the sets. We additionally constrain the compositions for each set so that for a given adverb-action composition, its antonym-action composition is assigned to the same set. We assign the videos from one of the sets to the training set and split the videos of the other half into two different sets, assigning half of the instances in each composition to the test set and the other to an unlabelled set (which is used to train [DS22] with pseudo-labelling). Table 4.1 shows details about the replicated split for VATEX, and for our proposed splits based on ActivityNet and MSR-VTT (full details are provided in the supplementary material).

Dataset	# tr (s)	# t (s)	# tr (p)	# t (p)
VATEX	6603	3293	319	316
MSR-VTT	987	454	225	225
ActivityNet	1490	848	635	543

Table 4.1: Statistics of the proposed dataset splits for the retrieval of unseen adverb-action compositions on the MSR-VTT and ActivityNet datasets. (tr: train, t: test, s: video samples, p: adverb-action pairs)

4.5 Experiments

In this section, we provide details about the baselines, implementation details, and evaluation metrics used in this work. Video-adverb retrieval results on five benchmarks are presented in Section 4.5.1, and we provide model ablation studies in Section 4.5.2. In Section 4.5.4, we investigate the transfer to unseen adverb-action compositions during inference.

Baselines. We report results for the **Prior** and **S3D pre-trained** baselines from [Mol+23]. **Prior** does not require any training but it uses the data distribution and adverb frequency for scoring. **S3D pre-trained** is also training-free and uses the similarity between frozen video and text representations from the S3D backbone jointly trained on video and text. **TIRG [Vo+19]** employs a similar residual gating mechanism as ReGAdA for image-text retrieval. To adapt it to the video domain, we use the same video encoder as our method. Different from ReGAdA, it models the composition as a transformation of the action embedding and uses a classification-based training objective. We also compare our framework to **Action Modifier** [Dou+20] and to the recently proposed **AC** frameworks [Mol+23]. AC tackles the task either as a classification (AC_{CLS}) or regression (AC_{REG}) problem.

Implementation details. We use the video and text features provided by [Mol+23] which were extracted using a frozen S3D model that was jointly pre-trained on video-text pairs from HowTo100M [Mie+19]. Here, $d_x = 1024$, T is the length of the video in seconds, and $d_\theta = 512$. ReGAdA uses an internal embedding dimension $d_{dim} = 400$. We use $N_g = 2$, except for HowTo100M and Adverbs in Recipes where $N_g = 3$ and $N_g = 4$ respectively. Additionally, we set $N_r = 2$ except for Adverbs in Recipes where we use $N_r = 3$. The dropout probability in the residual gating mechanism is $drop_g = 0.6$ for all datasets but Adverbs in Recipes and HowTo100M where we use $drop_g = 0.7$. The loss hyperparameters are chosen as $\lambda_a = 1$ for all datasets and $\lambda_v = 2.0$ for all datasets, except for $\lambda_v = 1.5$ on Adverbs in Recipes. Furthermore, we use a $\lambda_{reg} = 1.0$ for all dataset except for HowTo100M where $\lambda_{reg} = 1.5$. We train with a batch size of 512, and employ the Adam [KB14] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 10^{-5}. Our method is trained for 2000 epochs using a learning rate of 10^{-5} for all datasets with the exception of HowTo100M where we use $3 * 10^{-5}$. We follow [Mol+23], and train all baselines for 1000 epochs using a learning rate of 10^{-4}. We conduct all experiments on a single Nvidia 2080-Ti GPU.

Evaluation metrics. We follow [Mol+23], and report mean Average Precision (mAP) scores for adverb-to-video-retrieval, in particular **mAP M** ("adverb-to-video (all)" in [Dou+20]) and **mAP W**. mAP M is computed by ranking videos that contain the same ground-truth action according to their similarity to the adverb-action text embedding. For mAP W, the class scores are reweighed according to their support size in the test set. For video-to-adverb retrieval, we report binary antonym accuracy **Acc-A**. This is equivalent to ranking adverb-action embeddings according to their similarity to the embedded video

	HowTo100M [Dou+20]			Adverbs in Recipes [Mol+23]			ActivityNet [DS22]			MSR-VTT [DS22]			VATEX [DS22]		
	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A
Priors	0.446	0.354	0.786	0.491	0.263	0.854	0.217	0.159	0.745	0.308	0.152	0.723	0.216	0.086	0.752
S3D pre-tr.	0.339	0.238	0.560	0.389	0.173	0.735	0.118	0.070	0.560	0.194	0.075	0.603	0.122	0.038	0.586
TIRG [Vo+19]	0.441	0.476	0.721	0.485	0.228	0.835	0.186	0.111	0.709	0.297	0.113	0.700	0.195	0.065	0.735
Act. M. [Dou+20]	0.406	0.372	0.796	0.509	0.251	0.857	0.184	0.125	0.753	0.233	0.127	0.731	0.139	0.059	0.751
AC_{CLS}† [Mol+23]	0.562	0.420	0.786	0.606	0.289	0.841	0.130	0.096	0.741	0.305	0.131	0.751	0.283	0.108	0.754
AC_{REG}† [Mol+23]	0.555	0.423	0.799	0.613	0.244	0.847	0.119	0.079	0.714	0.282	0.114	0.774	0.261	0.086	0.755
REGADA	**0.567**	**0.528**	**0.817**	**0.704**	**0.418**	**0.874**	**0.239**	**0.175**	**0.771**	**0.378**	**0.228**	**0.786**	**0.290**	**0.113**	**0.817**

Table 4.2: Results for adverb-to-video (mAP W/M) and video-to-adverb retrieval (Acc-A). Higher is better for all metrics. † refers to updated results provided by the authors.

and calculating the mAP by restricting the set of adverbs to the target adverb and its antonym ("video-to-adverb (antonym)" in [Dou+20]). Similar to [Mol+23], we report the best metrics independently. This means that models corresponding to each result may originate from different epochs.

4.5.1 Comparison with the State of the Art

In Table 4.2, we present adverb-to-video retrieval and video-to-adverb retrieval results with our REGADA framework on five benchmark datasets. It can be observed that REGADA outperforms the baselines across all datasets. In particular, we see more significant improvements of our framework over the prior methods for the adverb-to-video retrieval metrics (mAP W and mAP M) compared to video-to-adverb retrieval (Acc-A). For instance, on the HowTo100M dataset REGADA outperforms AC_{CLS} for adverb-to-video retrieval with mAP M and mAP W scores of 0.528 and 0.567 compared to 0.420 and 0.562. For the video-to-adverb retrieval measure Acc-A, REGADA obtains a score of 0.817 compared to 0.786 with AC_{REG}.

The most recent and strongest competitor [Mol+23] optimises its systems using two different losses. The best results obtained from these two models are reported for each dataset and metric, showing no clear pattern as to which model variant is stronger. Our REGADA framework consistently outperforms both model variants [Mol+23] on all metrics and datasets. We hypothesise that our framework's strong performance can be attributed to its compositional embeddings which is a key element of REGADA.

4.5.2 Model Ablations

This section analyses the impact of using different input text information, losses, and components in the text encoder on the overall video-adverb retrieval performance of REGADA.

Input to the text encoder. The gating mechanism in REGADA represents the composition as a residual on top of the adverb and allows the adverb information to be retained, leveraging the action as auxiliary information. We refer to the adverb as the *main* and the action as the *auxiliary* modality in REGADA. We investigate if a compositional adverb-action word embedding ϕ_{comp}, which directly embeds an adverb-action label pair

Text Input		HowTo100M [Dou+20]			Adverbs in Recipes [Mol+23]			ActivityNet [DS22]			MSR-VTT [DS22]			VATEX [DS22]		
main	*auxiliary*	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A
ϕ_a	ϕ_v	0.485	0.390	0.824	0.436	0.221	0.872	0.225	0.147	0.763	0.336	0.144	0.780	0.245	0.078	0.807
ϕ_{comp}	ϕ_v	0.498	0.454	0.827	0.518	0.322	0.877	0.220	0.150	0.751	0.350	0.144	0.771	0.255	0.084	0.808
ϕ_{comp}	ϕ_a	0.503	0.467	**0.830**	0.524	0.365	**0.881**	0.222	0.147	0.758	0.348	0.146	0.763	0.255	0.090	0.806
ϕ_v	ϕ_a	**0.567**	**0.528**	0.817	**0.704**	**0.418**	0.874	**0.239**	**0.175**	**0.771**	**0.378**	**0.228**	**0.786**	**0.290**	**0.113**	**0.817**

Table 4.3: Effect of using different types of input information for the text encoder in ReGAdA.

Loss			HowTo100M [Dou+20]			Adverbs in Recipes [Mol+23]			ActivityNet [DS22]			MSR-VTT [DS22]			VATEX [DS22]		
$\mathcal{L}_{trip,a}$	$\mathcal{L}_{trip,v}$	$\mathcal{L}_{reg}$	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A
✓	✗	✗	0.361	0.228	0.697	0.429	0.214	0.836	0.162	0.104	0.582	0.259	0.138	0.714	0.133	0.047	0.677
✗	✓	✗	0.340	0.236	0.740	0.430	0.213	0.846	0.128	0.079	0.664	0.260	0.127	0.737	0.166	0.062	0.743
✗	✗	✓	0.470	0.378	0.743	0.468	0.234	0.839	0.202	0.140	0.729	0.288	0.186	0.743	0.182	0.074	0.700
✓	✓	✗	0.367	0.246	0.755	0.468	0.239	0.851	0.157	0.098	0.674	0.273	0.116	0.737	0.174	0.062	0.756
✓	✓	✓	**0.567**	**0.528**	**0.817**	**0.704**	**0.418**	**0.874**	**0.239**	**0.175**	**0.771**	**0.378**	**0.228**	**0.786**	**0.290**	**0.113**	**0.817**

Table 4.4: Impact of using different losses to train ReGAdA. For losses that are not used, the corresponding scalar weight in $\mathcal{L}$ is set to zero.

(e.g. "*cut quickly*") with h_{text}, can be used as the main modality instead. Table 4.3 shows the impact of using different main and auxiliary modalities. ReGAdA obtains scores of 0.290 and 0.113 for mAP W and mAP M on VATEX compared to 0.245 and 0.078 when using ϕ_a as main modality and ϕ_v as auxiliary. This confirms that capturing information about the adverb is crucial for solving the task. Acc-A is less affected by the type of input information, ReGAdA obtains 0.817 compared to 0.806 when using ϕ_{comp} as main and ϕ_a as auxiliary modality. Overall, using ϕ_v as main and ϕ_a as auxiliary modality is most effective across datasets.

Losses. In Table 4.4, we show the impact of our three loss functions, $\mathcal{L}_{trip,a}$, $\mathcal{L}_{trip,v}$, and $\mathcal{L}_{reg}$. On VATEX, ReGAdA obtains a mAP W and mAP M of 0.290 and 0.113 compared to 0.182 and 0.074 when using only $\mathcal{L}_{reg}$. For Acc-A, ReGAdA obtains a score of 0.817 compared to 0.756 for $\mathcal{L}_{trip,a} + \mathcal{L}_{trip,v}$. The regression loss $\mathcal{L}_{reg}$ boosts the performance on all datasets significantly. Our novel loss combination gives the best video-adverb retrieval performance by better aligning adverb-action compositions and video representations. Previous work either only used triplet losses [Dou+20; DS22] or used a fixed textual regression target [Mol+23].

Residual gating mechanism in the text encoder. Table 4.5 analyses the contributions of the components of the residual gating mechanism, such as the residual branch, the sigmoid, and weight sharing between the gated and residual branches. On VATEX, ReGAdA achieves the best results. Interestingly, sharing weights between the gated and residual branches yields only slightly weaker results, with a mAP-W score of 0.288 compared to 0.290 with ReGAdA. For mAP M and Acc-A, ReGAdA obtains 0.113 and 0.817 compared to 0.111 and 0.815 when not using the residual. While some configurations can achieve better results in selected metrics, ReGAdA yields consistent state-of-the-art results across all metrics, confirming our model design choices.

Components			HowTo100M [Dou+20]			Adverbs in Recipes [Mol+23]			ActivityNet [DS22]			MSR-VTT [DS22]			VATEX [DS22]		
R	σ	SW	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A
✓	✓	✓	0.535	0.433	0.811	0.689	0.404	0.875	**0.256**	**0.190**	**0.771**	0.374	0.182	0.766	0.288	0.109	0.808
✓	✗	✗	0.512	0.496	0.811	0.501	0.269	0.862	0.234	0.171	0.770	0.360	0.194	0.780	0.260	0.098	0.804
✗	✓	✗	0.516	0.477	**0.817**	0.562	0.296	**0.877**	0.228	0.169	0.765	0.367	0.161	0.783	0.283	0.111	0.815
✓	✓	✗	**0.567**	**0.528**	**0.817**	**0.704**	**0.418**	0.874	0.239	0.175	**0.771**	**0.378**	**0.228**	**0.786**	**0.290**	**0.113**	**0.817**

Table 4.5: Impact of different components in the residually-gated text encoder. R: With residual branch W_{res}; σ: With sigmoid; SW: Sharing weights between W_{res} and W_{gate}.

Figure 4.2: Example results for ReGaDa (Ours) on the VATEX dataset compared to those from AC_{REG}. The two left examples are success cases for our model. The third and fourth example show bidirectionally performed actions that are labelled with only one of the adverbs. The right-most example shows a wrongly labelled video. Full videos are available at: https://hummelth.github.io/ReGaDa

4.5.3 Qualitative Results

We show qualitative results for ReGaDa on the VATEX dataset in Figure 4.2. In particular, success cases for ReGaDa which AC_{REG} retrieved a wrong adverb are shown below in the first and second columns. The third and fourth columns show videos with actions performed forwards/backwards, and upwards/downwards but labelled with only one of the adverbs. This makes both outputs plausible. The right-most column shows an example of a wrongly labelled video for which our model retrieves the correct adverb. This confirms ReGaDa's strong generalisation capabilities. In general, we observe that ReGaDa better captures directional movements or speed than AC_{REG}. It is also superior at disentangling the diverse visual effect of adverbs on different actions (e.g. crawl vs. bend backwards). This can potentially be attributed to the compositional nature of our learned adverb-action representations.

4.5.4 Generalisation to Unseen Adverb-Action Compositions

We additionally evaluate the ReGaDa framework on video-to-adverb retrieval for unseen adverb-action compositions, i.e. compositions that were not seen during training. We consider the existing VATEX benchmark and our proposed MSR-VTT and ActivityNet splits for this task (see Section 4.4). Following [DS22], we report binary antonym classification accuracy for video-to-adverb retrieval. We provide additional baseline results with the CLIP [Rad+21] model (details for this are provided in the supplementary material). In Table 4.6, we observe that ReGaDa significantly outperforms AC_{REG} on VATEX with an accuracy of 61.7 compared to 54.9. On ActivityNet, ReGaDa obtains a score of 58.4, outperforming [DS22] with a score of 57.0. This is impressive given that [DS22] was additionally trained on pseudo-labelled data. CLIP obtains an antonym accuracy of only

Model	VATEX	ActivityNet	MSR-VTT
CLIP [Rad+21]	54.5	55.1	57.0
Act. Mod. [DS22]	53.8	57.0	56.0
AC_{CLS} [Mol+23]	54.3	55.1	53.7
AC_{REG} [Mol+23]	54.9	53.9	59.0
REGADA	**61.7**	**58.4**	**61.0**

Table 4.6: Retrieval of unseen adverb-action compositions on the VATEX, ActivityNet and MSR-VTT benchmarks. [DS22] uses pseudo-labelling.

54.5 on VATEX, showing a limited fine-grained retrieval capability of CLIP. We provide a further analysis of exploiting different word embeddings for unseen compositions in the supplementary material. Overall, our model yields better results than any prior framework for both seen (c.f. Table 4.2) and unseen compositions.

4.6 Conclusion

In this work, we proposed a framework for video-adverb retrieval that uses a residual gating mechanism to generate compositional adverb-action representations from adverb and action word embeddings. Along with a novel training objective, our model achieves state-of-the-art results on five video-adverb retrieval benchmarks. Moreover, we introduce two additional dataset splits to benchmark the retrieval of unseen adverb-action compositions. Our proposed framework outperforms all prior works on this task, confirming that our text encoder results in better generalisation abilities.

5

EgoCVR: An Egocentric Benchmark for Fine-Grained Composed Video Retrieval

In Composed Video Retrieval, a video and a textual description which modifies the video content are provided as inputs to the model. The aim is to retrieve the relevant video with the modified content from a database of videos. In this challenging task, the first step is to acquire large-scale training datasets and collect high-quality benchmarks for evaluation. In this work, we introduce `EgoCVR`, a new evaluation benchmark for fine-grained Composed Video Retrieval using large-scale egocentric video datasets. `EgoCVR` consists of 2,295 queries that specifically focus on high-quality temporal video understanding. We find that existing Composed Video Retrieval frameworks do not achieve the necessary high-quality temporal video understanding for this task. To address this shortcoming, we adapt a simple training-free method, propose a generic re-ranking framework for Composed Video Retrieval, and demonstrate that this achieves strong results on `EgoCVR`. Our code and benchmark are freely available at https://github.com/ExplainableML/EgoCVR.

5.1 Introduction

Recent advances in Vision-Language Models (VLMs) have enabled video search through free-form textual descriptions. However, expressing complex queries, especially those involving subtle transformations or actions, remains challenging with purely text-based searches. In the image domain, Composed Image Retrieval (CIR) [Liu+21b; Vo+19; Wu+21] has emerged as a related task where a user provides a reference image and a textual description of the desired modification. In the video domain, the corresponding task is coined as Composed Video Retrieval (CVR) [Ven+24] where the aim is to retrieve videos from a database given a reference video and a textual query that describes how the reference video should be modified. For example, a user searching through a long video might provide a short reference clip showing construction work along with a textual description such as "`make the person cut with a jigsaw instead`" to pinpoint the

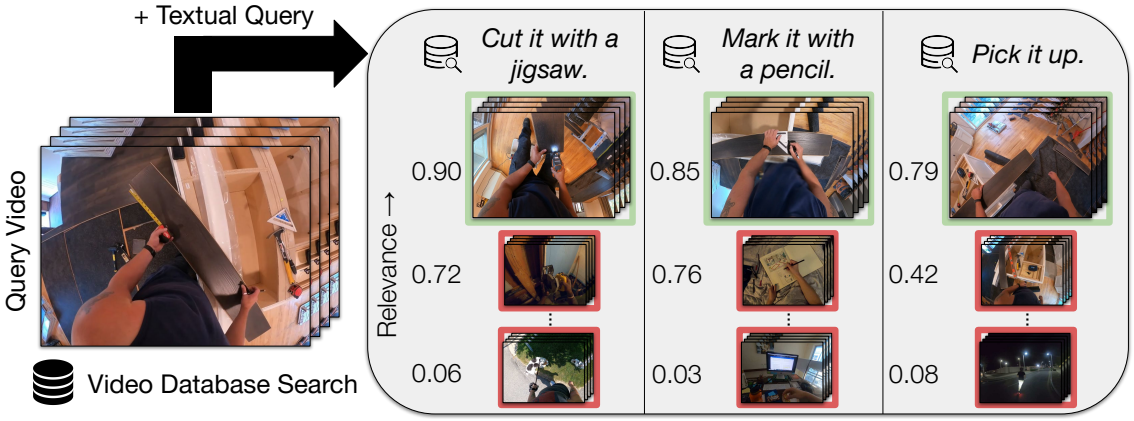

Figure 5.1: The goal of the Composed Video Retrieval (CVR) task is to retrieve the correct video using both a query video and a textual video modification instruction that describes the semantic changes required from the query video.

precise video they are looking for (See Figure 5.1). CVR remains a relatively under-explored area, posing unique challenges due to the added complexity of effectively utilizing the temporal information inherent in videos. CVR is extremely challenging because it requires understanding both the visual and textual inputs and composing them to retrieve the desired video efficiently.

A major step towards tackling the CVR challenge is the introduction of the large-scale WebVid-CoVR training set and a smaller evaluation benchmark. These datasets are automatically collected by using existing video-text datasets looking for pairs that differ only in a single word in the caption, and using a Large Language Model (LLM) [Tou+23] to generate the textual instruction. The final training set contains over 1.6M triplets, which is extremely useful for the CVR task. However, the evaluation set quality is quite limited due to the automatic dataset construction. For instance, most of the modifications predominantly focus on the color, shape, and adding/removing objects from the scene that do not require temporal understanding (see Figure 5.3). Therefore, the task can be tackled with a single image rather than a video, e.g. a vision-language model [Li+22] trained on the image level achieves state of the art.

In this work, we propose to create an evaluation set for the Composed Video Retrieval task that requires holistic video understanding to obtain strong performance. To achieve this, we propose `EgoCVR`, a manually curated and high-quality evaluation set with 2,295 videos sourced from the Ego4D[Gra+22] dataset. Our `EgoCVR` dataset consists of a query and target clip sourced from the same long video and the textual modifier asking for a subtle change in the action being performed in the clip. As a result, models need to have strong video understanding to be able to achieve strong performance in our evaluation setting.

Furthermore, we evaluate on our new benchmark several methods designed for cross-modal retrieval, consisting of vision-language models such as CLIP[Rad+21], BLIP [Li+22], the video-based method LanguageBind [Zhu+24], as well as the egocentric video model

EgoVLP [Lin+22b; Pra+23], by adapting them to the CVR task. Naively adapting vision-language models to perform the CVR task, even if the model was finetuned on the large benchmark, does not work well, e.g. $\text{BLIP}_{\text{CoVR}}$ finetuned for the CVR task on 1.6M triplets performs poorly on `EgoCVR`.
To address this shortcoming, we propose to adapt a training-free method proposed for Composed Image Retrieval [Kar+24] to the Composed Video Retrieval task. When employed with a generic re-ranking strategy, this approach, which we name `TFR-CVR`, achieves the best results among all considered methods in various evaluation settings. To summarise, we make the following contributions:

- We propose `EgoCVR`, a benchmark with 2,295 queries, to evaluate vision-language models for the task of Composed Video Retrieval.
- We evaluate several vision-language models with varying configurations on our benchmark and find that existing models, even when finetuned for Composed Video Retrieval, have several shortcomings on the action-focused `EgoCVR` benchmark.
- Finally, we demonstrate that our proposed training-free `TFR-CVR` method, along with a generic re-ranking framework, achieves strong performance on the `EgoCVR` benchmark.

5.2 Related Work

Video-Language Models and Retrieval. Early work on video retrieval often focused on extending retrieval approaches from images to videos by aggregating image features within a video [DLS18; Ota+16; TTS16; Xu+15b]. However, with the introduction of large-scale video-text datasets [Bai+21; Mie+19; Wan+19a; Xu+16], and contrastive language-image pre-training [Li+23; Li+22; Rad+21], there have been several models proposed for the task of video-text retrieval [Bai+22; Gir+23; Luo+22; Zhu+24]. Due to the growing popularity of egocentric video datasets [Gra+22; Gra+24], video-language models have been proposed that specifically focus on this setting [Lin+22b; Pra+23; Zha+23b]. However, while there has been growing interest in developing video-based foundation models [Che+23; Xu+23], these have all been focused on captioning and video-text retrieval. Different from this, we show how existing video-text models can be utilised for fine-grained Composed Video Retrieval.
Composed Image Retrieval. The task of Composed Image Retrieval (CIR) has found significant application in conditional search [Han+17; Vo+19; Wu+21], where users perform interactive dialogue to refine a given query image toward retrieving specific items. Classical techniques often employ custom models that project text-image pairs into a common embedding space [ALK21; Bal+22; CB20; CGB20; LKH21; Vo+19] or use cross-modal attention mechanisms [Del+22]. With the advent of vision-language foundation models [Bom+21; Jia+21; Rad+21], interest in CIR has surged, especially in zero-shot

settings that avoid the need for task-specific models. Recent works either attempt to train models that avoid the necessity for paired triplets [Bai+24; Bal+23; CL23; Gu+24b; Sai+23; Tan+24] or train models on large datasets that then generalise to a wide variety of scenarios [Gu+24a; Lev+24; Liu+23; Ven+24]. There have also been several datasets and benchmarks proposed for Composed Image Retrieval including large-scale generic datasets such as CIRR [Liu+21b], CIRCO [Bal+23], as well as fine-grained evaluation benchmarks focusing on fashion [Han+17; Wu+21], fine-grained attributes [VCM23], sketches [Gat+24] or birds [For+19]. In this work, we are inspired by both CIR methods [Kar+24; SYG23] as well as CIR benchmarks [Liu+21b; VCM23] in curating a fine-grained Composed Video Retrieval dataset, as well as proposing general methods that can tackle this task.

Composed Video Retrieval. To the best of our knowledge, the only existing benchmark available for Composed Video Retrieval is WebVid-CoVR [Ven+24]. Further, the only models tailored for it are BLIP models finetuned on the training set of WebVid-CoVR [Tha+24; Ven+24]. Concurrent to our work, the task of *video detours* [Ash+24] was introduced, which focused on retrieving and localising temporal segments within long videos using free-form textual queries and the query video, specifically for instructional videos. In this work, we propose a fine-grained evaluation benchmark for Composed Video Retrieval with two evaluation settings, along with a training-free method using video-specific models for this task.

5.3 EgoCVR: An Egocentric Benchmark Dataset for Composed Video Retrieval

In Section 5.3.1, we first formally define the task of Composed Video Retrieval, while in Section 5.3.2, we describe our dataset construction methodology in detail.

5.3.1 Problem Definition

The Composed Video Retrieval task was first introduced by Ventura et al. [Ven+24]. Let $\mathcal{V}$ denote the space of videos and $\mathcal{T}$ the space of textual instructions. Given a query video $q_v \in \mathcal{V}$ and textual instruction $q_t \in \mathcal{T}$, the goal of composed video retrieval (CVR) is to identify the modified video $v \in \mathcal{D}$ from a database of videos (gallery) $\mathcal{D} = \{v_1, \ldots, v_n\}$, where n is the number of videos in $\mathcal{D}$, that most closely represents the semantic modifications described by q_t. The task can be formalized as a scoring function $\Phi : \mathcal{V} \times \mathcal{T} \times \mathcal{D} \to \mathbb{R}$. This function measures the similarity between the query video q_v, the modification text q_t, and each candidate video v_i in the database, $0 \leq i \leq n$. The video with the highest score according to Φ is deemed the optimal retrieval result.

The scoring function Φ is implemented by representing videos and text within a shared embedding space. We denote the video encoder as $\Psi_v : \mathcal{V} \to \mathbb{R}^d$ and the text encoder as $\Psi_t : T \to \mathbb{R}^d$, where d is the dimension of the embedding space. The video encoder Ψ_v processes either single frames (with averaged frame-level embeddings) or frame sequences

using a temporal video encoder. The text encoder Ψ_t embeds the modification instructions into the same space as Ψ_v. Text and video embeddings are then combined to form a multi-modal video-text embedding $q_{v,t}$ using a fusion function $\Psi_q : \{q_v, q_t\} \rightarrow \mathbb{R}^d$. Candidate videos from the database $\mathcal{D}$ are also encoded using Ψ_v. Finally, the cosine similarity is used as a matching score between the query embedding $q_{v,t}$ and each candidate video embedding $v_i, v_i \in \mathcal{D}, 0 \leq i \leq n$.

5.3.2 From Egocentric Videos to Composed Video Retrieval

We collect videos and the corresponding annotations with the narrations (in free-form text) from the Ego4D Forecasting Hand and Object (FHO) task[1]. As this task focuses on understanding and anticipating human-object interactions, we ensure that collected videos contain frequent and diverse interactions with clear visual quality and a broad range of everyday objects. The FHO task provides short video narrations describing actions and object interactions. An example of a narration is "`#C C trims the blue cardboard to a circular shape with the scissors in her right hand.`".

The dataset includes 155k narrations, each associated with 2-8 second video clips extracted from 1,250 long-form videos. We reduce the 155k densely annotated clips to 9k distinct clips by automatically filtering out clips with temporal overlap, ensuring a higher likelihood of single, focused actions within each clip. Our dataset annotation process aims to find pairs of videos that have subtle differences. While previous work [Ven+24] applied an automated video-matching process by searching for single-word differences in video captions, `EgoCVR` is created using a careful manual annotation process outlined in the following.

We manually search for possible video pairs within long videos, i.e. video pairs originate from the same source video. Creating annotations from the same source video allows for fine-grained comparisons where the primary difference between video pairs is the controlled textual modification. We identify matching pairs through similarity in their narrations, i.e. the narrations differ in a single semantic concept like actions (e.g. rinsing vs. rubbing) or objects (e.g. knife vs. spoon). During annotation, an emphasis was put on creating pairs where temporal modifications are prioritised. We manually disregard annotation pairs through visual inspection when i) the narrations do not accurately describe the clip, ii) the narrated actions or objects are visible only for a fraction of the clip, or iii) the presence of multiple actions would result in ambiguous samples. When multiple videos with the same narration are present (i.e. "`#C C puts down the piece of cloth.`" and "`#C C puts down the cloth.`"), we group the clips together. This allows us to create samples in `EgoCVR` with multiple ground truth targets, even for narrations that do not perfectly match with an exact textual search. This annotation process resulted in a total of 2,295 queries with an average of 1.2 ground-truth targets per query.

[1]The task can be viewed here: https://ego4d-data.org/docs/tutorials/FHO_Overview/

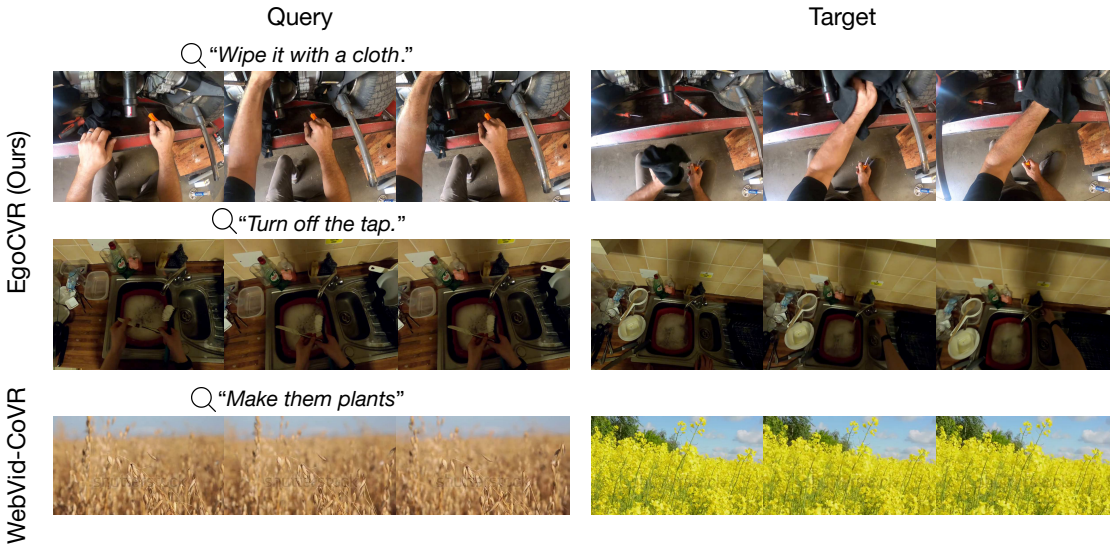

Figure 5.2: Samples consisting of visual and text queries along with the target video from our test set `EgoCVR` (top two rows) and WebVid-CoVR-Test set [Ven+24] (bottom row).

Creating Textual Video Modification Instructions. We create textual video modification instructions from the video narrations of paired clips. Ideal modification instructions clearly describe the most prominent change that needs to be applied to the query video to get to the desired target video. We design these modifications to be as concise as possible while still conveying all the relevant information. Instructions in `EgoCVR` provide only the minimum necessary semantic difference between the query and the target videos. For instance, the instruction "`Rinse it instead.`" does not provide information on which object to rinse. To create the text modifications, we utilise the reasoning capabilities of LLMs to generate concise instructions that describe the transformation from the provided query clip narration to the target clip narration. As LLM, we employ GPT-4 [Ope23] and provide the LLM with a list of 15 in-context examples [Don+24] together with a clear instruction prompt (more details in the supplementary). We illustrate examples from `EgoCVR`, as well as how it contrasts with typical samples from WebVid-CoVR in Figure 5.2.

Visual Distractors. We additionally collect distractor video clips for each annotated target video similar to the CIRR image subsets [Liu+21b]. We automatically source the distractor clips from the same long-form video provided by the Ego4D FHO task. Our collected distractor clips ensure high visual similarity (i.e. identical camera wearer and scene) and prevent trivial retrieval shortcuts based solely on visual similarity. To obtain the distractor clips, for each target video in `EgoCVR`, we filter out clips from the Ego4D FHO annotations originating from different long-form videos, clips used as query-target video annotations, and clips depicting the same action as the target. We then rank potential distractor clips by their narration's CLIP similarity to the target video narration. Finally, we sample up to 6 distractor clips per target video. To represent various semantic similarity levels, we

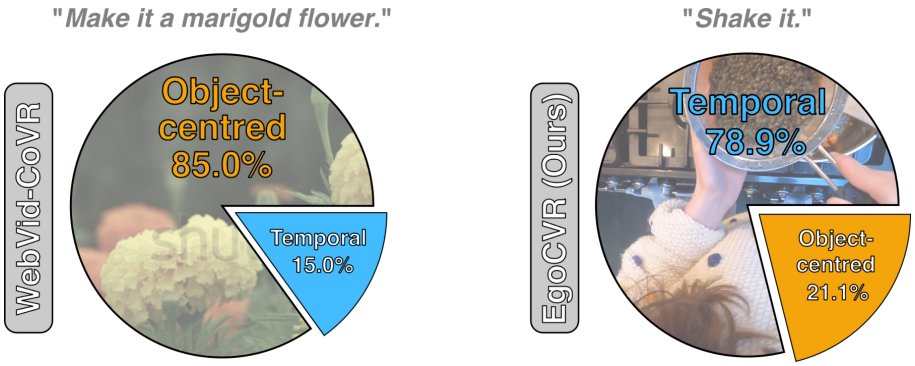

Figure 5.3: EgoCVR focuses to a significantly greater extent on temporal and action-related modifications (blue) as opposed to object-centred modifications (orange) when compared to the previously existing WebVid-CoVR-Test benchmark [Ven+24].

sample one clip from the bottom 10 % of similarity scores, four from the middle 80 %, and one from the top 10 %. We sample a total of 10,522 distractors with an average of 4.2 distractors per target video.

Dataset Statistics. EgoCVR is created with the intent to explore the video understanding capabilities of current vision-language models. Our annotation process ensures i) high-quality annotated video pairs and ii) a strong focus on temporal events. We analyse the instructions of EgoCVR and WebVid-CoVR-Test regarding their focus on temporal events. We consider instructions as temporal if the change from query to target video, described using the modification text, directly changes the depicted action or requires temporal video understanding (i.e. *Pick* it *up* instead.). In contrast, object-centred modifications require no action understanding but manipulating given objects (i.e. Cut the *carrot* instead.). To obtain this information, we instruct GPT-4 to assess whether a given instruction focuses on temporal events or objects (see the supplementary for more details). Our EgoCVR evaluation benchmark consists of 2,295 samples, from which 1,811 focus on temporal events (78.9 %) and 484 on object-centred changes (21.1 %). As visualised in Figure 5.3, this starkly contrasts with WebVid-CoVR-Test, where 85 % of samples focus on object-centred modifications.

To assess the variety of actions and objects in EgoCVR, we apply part-of-speech (POS) tagging on the instructions. For actions, we count the occurrences of unique verbs for temporal modifications, while for objects, we count the occurrences of unique direct objects for object-centred modifications. With 179 different actions and 121 unique objects, we find a great variety of actions and objects present in EgoCVR. Video clips in EgoCVR have a length of 3.9-8.1 seconds with an average length of 7.9 seconds. Modification instructions in EgoCVR are designed with an average of four words to be precise and concise, i.e. we ensure that the instruction only describes the transformation and not the target video itself. Instructions vary from short two-word (e.g. "Shake it.") to longer and more detailed instructions (e.g. "Use the other hand to pick up a different object from the shelf.").

5.4 Training-Free Re-Ranking Composed Video Retrieval

We adopt several vision-language methods for composed video retrieval. To show that our proposed benchmark focuses on temporal actions, we employ both image processing and video processing models in the evaluation. As image processing models, we employ two widely used image-language models, namely **CLIP** [Rad+21] and **BLIP** [Li+22]. We also adapt the video-language models **EgoVLPv2** [Pra+23] and **LanguageBind** [Zhu+24] that were specifically designed to learn temporal video representations. EgoVLPv2 was specifically pre-trained on egocentric videos, while LanguageBind aligns various modalities such as video, infrared, depth and audio to a frozen language encoder after pre-training. We also employ the recently introduced composed video retrieval framework $\textbf{BLIP}_{\textbf{CoVR}}$ [Ven+24] and $\textbf{BLIP}_{\textbf{CoVR-ECDE}}$ [Tha+24] which leverages the BLIP model for cross-attention between visual and textual encoders. They were specifically finetuned on WebVid-CoVR for CVR, leading to top performance on the WebVid-CoVR test set. We also evaluate CIReVL [Kar+24] on our benchmark since it is a training-free method. Below, we only discuss the self-adapted methods TF-CVR and `TFR-CVR`.

Composed Video Retrieval by Language. We use a methodology very similar to **CIReVL** [Kar+24], which has successfully been applied for Composed Image Retrieval. Given a video captioning model such as LaViLa [Zha+23b], we can obtain the textual caption of the query video. We name this approach training-free CVR (**TF-CVR**). Specifically, given a query video q_v, and a video captioning model Ψ_C, we obtain its textual representation as $c_q = \Psi_C(q_v) \in \mathcal{T}$. However, this video caption only captures the reference video, not the specified textual modification q_t. While the two texts could be combined naively using concatenation, we use an LLM to combine the video caption and textual modifier into a coherent target caption, similar to CIReVL [Kar+24]. Formally, given access to an LLM Ψ_R, we generate a target video caption as $c_q^t = \Psi_R(p \circ c_q \circ q_t)$, which queries the LLM with a concatenation of the template prompt p, the generated video caption c_q and modification instruction q_t. The template prompt p consists of a few in-context examples to guide the LLM and a short task description. Concrete examples of this process are shown in the supplementary material. Given this generated target caption c_q^t, TF-CVR searches the video database $\mathcal{D}$ alongside c_q^t using a text-video model (e.g. EgoVLPv2 [Pra+23], LanguageBind [Zhu+24]). The retrieved target V_q^t is:

$$V_q^t = \underset{v \in \mathcal{D}}{\texttt{argmax}} \frac{\Psi_V(v)^\intercal \Psi_T(c_q^t)}{||\Psi_V(v)|| \cdot ||\Psi_T(c_q^t)||} \quad . \tag{5.1}$$

Re-Ranking for Composed Video Retrieval. While the proposed approach, TF-CVR, is simple and effective, a major drawback of the method is that solely relying on text could potentially lead to the selection of semantically similar yet visually unrelated video clips. Therefore, we first apply a visual filtering step to select a candidate video database $D' \subset D$. This is performed by selecting the n_c most similar video clips to the provided reference

video q_v. This is described more formally as:

$$D' = \underset{v \in \mathcal{D}}{\texttt{top}n_c} \left(\frac{\Psi_V(q_v)^\intercal \Psi_V(v)}{||\Psi_V(q_v)|| \cdot ||\Psi_V(v)||} \right) \quad . \tag{5.2}$$

After filtering, we apply our proposed approach TF-CVR, except now, the video gallery is restricted to D'. We refer to this method as training-free re-ranking CVR (`TFR-CVR`). We demonstrate the efficacy of this method in Section 5.5.2, especially in settings with a large video gallery. Note that the visual filtering applied in this step can use a different visual encoder than the text-video retrieval step, allowing us to leverage the complementary abilities of different models.

5.5 Experiments

We explain the two proposed evaluation settings and metrics in Section 5.5.1. Further, we discuss the results obtained on `EgoCVR` in Section 5.5.2 along with the ablations, analyses and qualitative examples performed on `EgoCVR`.

5.5.1 Evaluation Settings and Implementation Details

Global and Local Settings. We consider two possible evaluation settings for `EgoCVR`. The first is the standard composed image/video retrieval setting, where the gallery comprises a long list of videos. We refer to this strategy as the *global* search. In the *global* setting, the query is searched in the pool of videos, which contains all the other video queries, along with their video distractors. Each query tuple has a search gallery of at least 10,661 video clips, with a maximum of 12,526. The second setting is the *local* search and is obtained by restricting the gallery to only clips from the same video sequence. This strategy simulates the scenario when searching in a long video for a specific moment. Each query tuple has a gallery of a maximum of 10 clips with an average of 6.4.

Evaluation Metrics. We employ the widely used recall metrics, namely Recall@1, Recall@5 and Recall@10 for the *global* setting, while for the *local* setting we employ Recall@1, Recall@2 and Recall@3, since the length of the gallery is considerably smaller. When a query has more than one target video, we consider the target as true positive only once when one of the target videos is retrieved.

Implementation Details. We perform our experiments using the publicly available official implementations of various vision-language models [Li+22; Pra+23; Rad+21; Zhu+24], using their default configurations to extract both visual and textual features. We employ the ViT-L/14 [Dos+21] version of the CLIP model provided by OpenCLIP [Ilh+21] which was pre-trained on DataComp-1B [Gad+23], as well as, the BLIP-Large variant finetuned on COCO [Lin+14] and the BLIP model finetuned on WebVid-CoVR [Ven+24] by Ventura et al. [Ven+24]. We use the fully finetuned video encoder for LanguageBind [Zhu+24] and EgoVLPv2 [Pra+23] with full projection. Unless otherwise noted, we use $n_c = 15$ and employ EgoVLPv2 as the text encoder Ψ_T in `TFR-CVR` (Equation 5.1). CLIP and

Method	Video Model	Textual Input	Visual Input	Fusion Strategy	Global R@1	Global R@5	Global R@10	Local R@1	Local R@2	Local R@3
Random	✗	✗	✗	-	0.01	0.05	0.1	25.3	38.2	50.7
CLIP	✗	✓	✗	-	0.7	1.7	2.7	33.5	48.8	61.8
BLIP	✗	✓	✗	-	0.4	1.4	2.7	32.5	46.9	59.7
EgoVLPv2	✓	✓	✗	-	1.7	3.9	7.2	<u>41.0</u>	<u>57.3</u>	<u>69.0</u>
LanguageBind	✓	✓	✗	-	0.9	2.7	4.2	34.2	51.1	64.1
CLIP	✗	✗	✓	-	7.4	33.2	<u>55.3</u>	26.1	43.4	57.7
BLIP	✗	✗	✓	-	6.5	32.6	<u>55.3</u>	26.5	43.7	57.5
EgoVLPv2	✓	✗	✓	-	7.6	32.5	49.6	27.5	44.3	59.1
LanguageBind	✓	✗	✓	-	6.1	33.1	53.4	26.1	42.9	57.7
CLIP	✗	✓	✓	Avg	7.5	33.6	**55.6**	26.4	43.7	57.9
BLIP	✗	✓	✓	Avg	8.7	32.9	52.8	29.5	45.9	61.0
EgoVLPv2	✓	✓	✓	Avg	<u>9.5</u>	<u>34.9</u>	52.1	30.7	51.3	66.0
LanguageBind	✓	✓	✓	Avg	6.1	33.2	53.5	26.1	43.1	57.8
$\text{BLIP}_{\text{CoVR}}$ [Ven+24]	✗	✓	✓	Cross-Attention	5.4	15.2	24.3	33.1	49.5	62.9
$\text{BLIP}_{\text{CoVR-ECDE}}$ [Tha+24]	✗	✓	✓	Cross-Attention	6.0	16.3	24.8	33.4	49.3	63.0
CIReVL [Kar+24]	✗	✓	✓	Captioning	2.0	6.8	10.6	33.6	49.7	61.4
TFR-CVR (Ours)	✓	✓	✓	Captioning	**14.1**	**39.5**	54.4	**44.2**	**61.0**	**73.2**

Table 5.1: Results on both the global and local evaluation settings on EgoCVR. Our proposed TFR-CVR achieves state-of-the-art results in both the global and local settings. We also report several baselines that only use the text, the reference video, or a naive average of the visual and textual embeddings (Fusion Strategy Avg). The best and the second best results are in **bold** and <u>underlined</u>, respectively.

BLIP visual representations for the videos are obtained by averaging embeddings from 15 uniformly sampled image frames.

5.5.2 Benchmark Evaluation and Model Ablations

We explore the potential of different query modalities in fine-grained composed video retrieval for EgoCVR. Specifically, we use three methods for video ranking: retrieval using only text query (**text-only**), retrieval using only visual query (**visual-only**), and retrieval using both text and visual queries (**visual-text**).

Global Setting. The results for the global evaluation setting are presented in Table 5.1. We notice the absolute performance of the Recall@k ($k \in \{1, 5, 10\}$) values being quite low due to having thousands of candidate video clips in the gallery for each query. However, we observe that relying solely on the text performs extremely poorly for all methods, as they fail to achieve an R@1 of even 2%. We notice that methods relying on visual features demonstrate competitive performance (up to 7.6% in R@1). In this setting, our proposed TFR-CVR achieves the best results (R@1 of 14.1%) due to the combination of LanguageBind-based candidate filtering using visual features, followed by re-ranking using the generated target caption. It is also notable that the BLIP finetuning methods ($\text{BLIP}_{\text{CoVR}}$ and $\text{BLIP}_{\text{CoVR-ECDE}}$, which attain state-of-the-art results on WebVid-CoVR as well as several CIR benchmarks, does not generalise to our proposed EgoCVR benchmark, achieving an R@1 value of only 5.4% and 6.0% respectively. We also observe that CIReVL [Kar+24], which was tailored

Method	Temporal	R@1	R@5	R@10
LanguageBind	✗	**6.4**	25.3	38.0
	✓	6.1	**33.1**	**53.4**
EgoVLPv2	✗	6.9	24.0	34.8
	✓	**9.5**	**34.9**	**52.1**
$\text{BLIP}_{\text{CoVR}}$	✗	4.1	12.5	19.6
	✓	**5.4**	**15.2**	**24.3**
TFR-CVR	✗	10.2	27.4	39.2
	✓	**14.1**	**39.5**	**54.4**

Table 5.2: Results in terms of R@1, R@5 and R@10 on the global setting that emphasize the importance of temporal information on EgoCVR.

Method	R@1	R@5	R@10
LanguageBind (Stage 1)	6.1	33.1	53.4
EgoVLPv2 (Stage 1)	7.6	32.5	49.6
TF-CVR (Stage 2)	4.4	12.9	18.3
TFR-CVR (EgoVLPv2 → TF-CVR)	12.2	35.1	49.5
TFR-CVR (LanguageBind → TF-CVR)	**14.1**	**39.5**	**54.4**

Table 5.3: Results in terms of R@1, R@5 and R@10 demonstrating the importance of the two-stage (filtering and re-ranking) process for our proposed TFR-CVR on the global setting of EgoCVR. When applying re-ranking, we show the model that is applied for visual filtering before using TF-CVR.

for the task of CIR, does not generalise directly to videos, obtaining an R@1 score of 2%.

Local Setting. In the local setting, we notice diminishing returns from methods that rely solely on visual features, performing only marginally better than random selection. This is due to the fact that, in this setting, all the videos in the gallery are by design very similar. Therefore, the visual similarity is not too helpful. Textual search performs much better because in the local setting, the textual information is the main indicator in finding the videos in the gallery. TFR-CVR performs text-video retrieval with a full caption that also captures the information from the source video, achieves the best result (R@1 of 44.2%).

Benefits of Temporal Information. We demonstrate the benefits of using temporal information through processing the whole video compared to using only a single frame sampled from the middle of the video. The results are reported in Table 5.2. We observe that temporal information improves performance significantly across all methods (up to 12.1 percentage points in terms of R@5), confirming that our benchmark benefits and requires temporal understanding. This is particularly noticeable on the R@5 and R@10 metrics, where we observe TFR-CVR improving from 27.4% to 39.5% and from 39.2% to 54.3%, respectively, emphasising the importance of using temporal information for this task.

Benefits of Re-Ranking. We also demonstrate the efficacy of our two-stage approach, TFR-CVR (i.e. first selecting candidates using visual similarity and re-ranking them using text similarity), on the global setting in Table 5.3. We notice that only using visual similarity or textual search alone is insufficient, while combining the two steps leads to the best-performing results (last two rows). Additionally, we highlight the benefits of TFR-CVR in drawing complementary knowledge from distinct models. For instance, TF-CVR employs the textual encoder from EgoVLPv2. Using LanguageBind for visual filtering in TFR-CVR (last row) instead of EgoVLPv2 improves the retrieval results across all metrics.

In Figure 5.4, we illustrate the resulting order of the videos obtained after re-ranking. We can notice that in the first stage, while all videos are visually similar, the correct video is ranked lower, while after the second stage, the target video is moved to the first position,

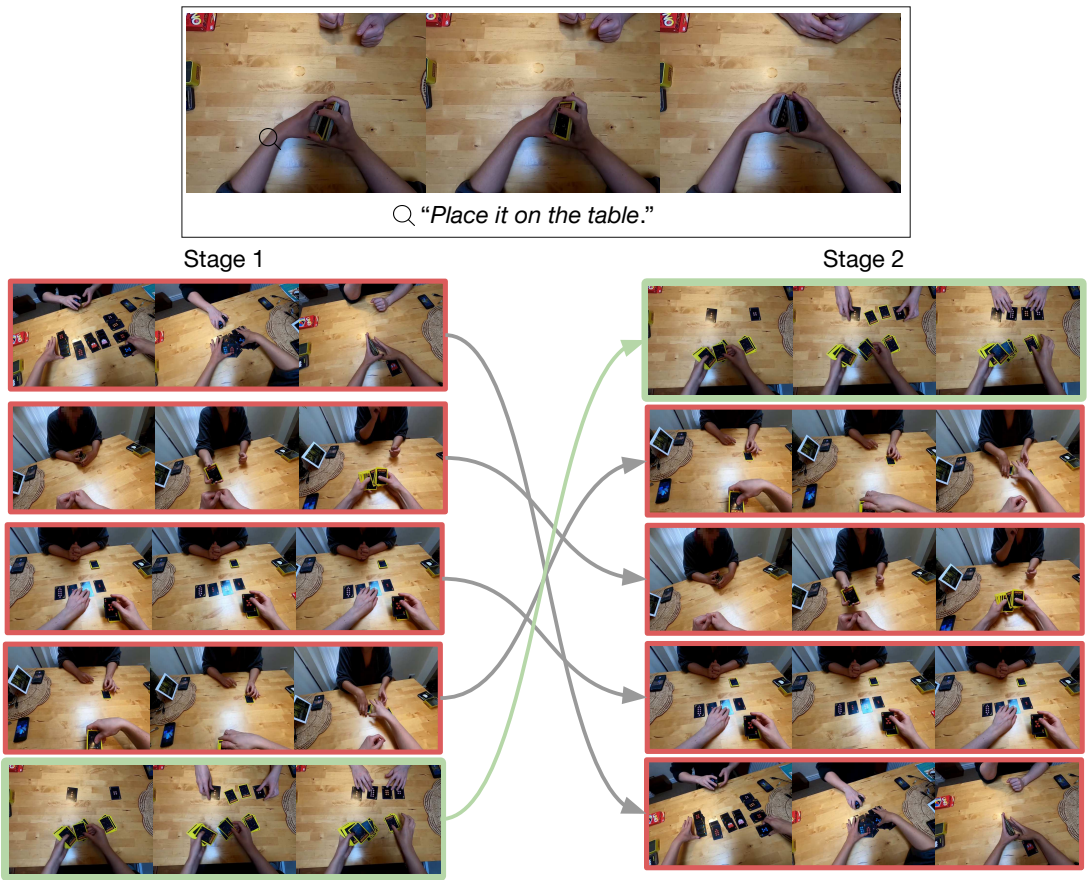

Figure 5.4: The first and the second stage ranking results of the TFR-CVR method. The arrows indicate how the ranking was changed. The correct video is showcased in green.

resulting in a correct retrieval.

Effect of the Number of Re-Ranking Candidates. Our approach involves re-ranking the candidates chosen by the first stage of visual filtering. We study the impact of the number of neighbours chosen after the first stage in Figure 5.5. We observe that the performance stops fluctuating once we select a sufficient number of candidates n_c for re-ranking ($n_c > 10$). Once the number of candidates becomes too large ($n_c > 30$), the performance starts diminishing and eventually loses the benefits of the visual filtering. In our experiments, we use $n_c = 15$. However, the final results are stable within a large range of selected candidates.

Effect of Text-Caption. We also investigate the benefits of using an LLM to generate a plausible caption of the video, along with its shortcomings in Table 5.4. This is achieved by resorting to text-video retrieval with different textual inputs. We experiment with the input textual modification, the predicted caption (as the result of video captioning and LLM reformulation described in Section 5.4), as well as using the ground-truth target caption provided by Ego4D. The ground-truth target caption serves as a useful upper bound on text-only search. We notice that the video captioning and LLM reformulation consistently improve the results for all the models. In the global setting, the improvement is especially

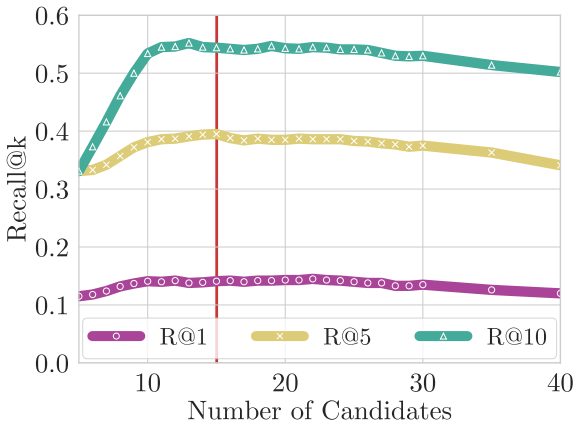

Figure 5.5: Effect of the number of candidates n_c for the visual re-ranking step of TFR-CVR. The vertical line denotes the value of n_c used in our experiments.

noticeable as the R@1 result increases at least twice (from 0.7% to 2.1% for the CLIP model) due to having a complete caption instead of a brief text. While the ground truth caption naturally improves the results further, it must be noted that improving the caption further does not offer much room for improvement. For instance, in the local setting, the R@1 result improves from 44.2% to 51.7% when our method is employed. Future work on EgoCVR would benefit both from improving the underlying text-video models through better foundation models [Lin+23a; Zha+24], as well as, from developing methods that can cohesively utilise the reference video and the textual instruction simultaneously.

Qualitative Examples. We demonstrate the benefits of our re-ranking approach in Figure 5.4. We observe that relying on visual features to select the top candidates results in visually similar clips, without capturing subtle actions accurately. After re-ranking with

Method	Text Source	Global			Local		
		R@1	R@5	R@10	R@1	R@2	R@3
CLIP	Instruction	0.7	1.7	2.7	33.5	48.8	61.8
	Pred. Caption	1.5	4.2	7.5	34.0	49.8	63.7
	GT Caption	2.1	5.9	9.3	35.1	52.0	69.4
LanguageBind	Instruction	0.9	2.7	4.2	34.2	51.1	64.1
	Pred. Caption	1.7	5.7	8.2	36.6	52.2	64.8
	GT Caption	3.3	8.0	11.5	39.2	56.6	69.1
TFR-CVR	Instruction	12.8	35.3	53.4	41.0	57.3	69.0
	Pred. Caption	14.1	39.5	54.4	44.2	61.0	73.2
	GT Caption	18.5	44.7	58.5	51.7	69.7	81.8

Table 5.4: Text-only retrieval results obtained with CLIP [Rad+21], LanguageBind [Zhu+24], and TFR-CVR on EgoCVR. As text query alternatives, we switch among the *instruction*, the caption prediction from video captioning [Zha+23b] combined with LLM reformulation (*Pred. Caption*), and lastly the ground-truth narration (*GT Caption*) available from Ego4D. The video query is used by TFR-CVR for visual re-ranking on the global evaluation.

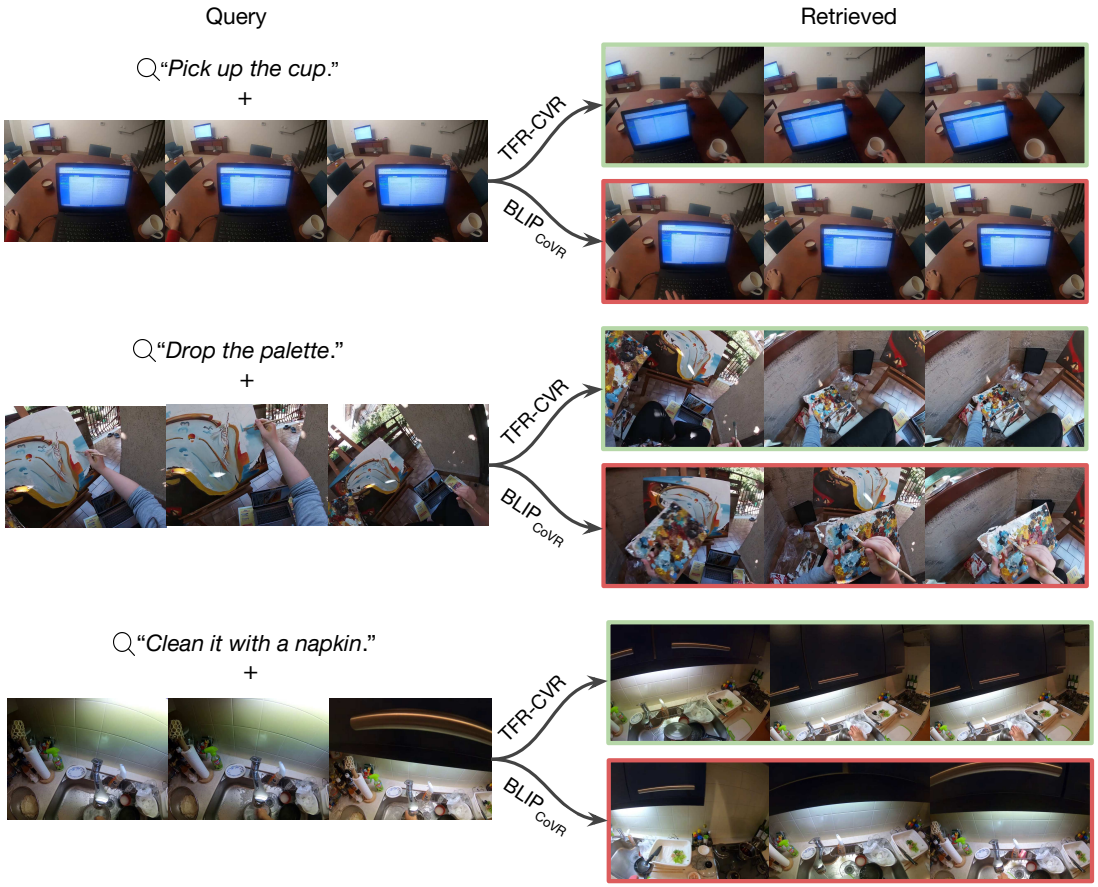

Figure 5.6: Qualitative examples of composed video retrieval ranking on EgoCVR. For each example, we show the queries along with the clip retrieved by TFR-CVR and $\text{BLIP}_{\text{CoVR}}$ [Ven+24]. The target videos are enclosed in green rectangles.

the text-video retrieval using the predicted caption, fine-grained actions are accurately captured in the final ranking.

We additionally illustrate qualitative examples in Figure 5.6, comparing the retrieved samples of the TFR-CVR and $\text{BLIP}_{\text{CoVR}}$ methods. We observe that our proposed method performs better than $\text{BLIP}_{\text{CoVR}}$, retrieving all targets. Notably, these examples require fine-grained action understanding. While TFR-CVR returns the correct clip, $\text{BLIP}_{\text{CoVR}}$ retrieves visually similar clips, however, they do not display the correct action. This highlights the inherent limitations of an image-based model despite being finetuned for Composed Video Retrieval.

Limitations. We provide a high-quality evaluation benchmark for CVR. However, collecting a training set, even through an automated process, would allow finetuning models for CVR instead of adapting existing vision-language models in a training-free manner. Furthermore, our evaluation benchmark also consists of egocentric videos, however, it can be employed to assess the generalization of any CVR model. Despite the aforementioned limitations, we believe that our proposed benchmark serves as an intriguing validation ground for adapting existing vision-language models and a valuable evaluation set for

high-quality temporal action understanding. Expanding the scope of the benchmark to include different types would also increase the diversity and applicability of the findings.

5.6 Conclusion

In this work, we introduce the EgoCVR benchmark for the task of fine-grained Composed Video Retrieval. We demonstrate that existing text-video and Composed Video Retrieval methods do not directly generalise to EgoCVR. Therefore, we introduce our method TFR-CVR, which uses existing video and language models in a modular fashion to achieve strong results on EgoCVR. We also show the shortcomings of existing vision-language models, even when they are explicitly finetuned for Composed Video Retrieval. We hope that our benchmark and method inspire further work on fine-grained action understanding and retrieval.

6

Discussion and Conclusion

This thesis tackles video understanding by developing settings and methods that use language to bridge the gap between visual content and semantic understanding. The research presented investigates a variety of video understanding tasks, unified by their use of language to enhance performance and generalisability. Specifically, Chapters 2 and 3 explore audio-visual low-shot learning scenarios, employing semantic class labels as semantic guidance. Chapter 2 tackles audio-visual generalised zero-shot learning, focusing on recognising videos from previously unseen classes, while Chapter 3 extends this to the few-shot learning given only a limited number of labelled examples for certain novel classes. Chapters 4 and 5 shift the focus to fine-grained video understanding. Chapter 4 tackles video-adverb retrieval by composing adverbs and verbs for fine-grained retrieval, while Chapter 5 introduces a novel benchmark and method for composed video retrieval, highlighting the usefulness of natural language instructions in guiding desired video content.

This chapter first discusses the contributions of this thesis in Sections 6.1 and 6.2. Afterwards, Section 6.3 highlights the limitations of the work presented in this thesis and proposes directions for future research.

6.1 Discussion of Contributions of Individual Chapters

In the following, we discuss the contributions of each of the Chapters 2 to 5 individually.

6.1.1 Temporal and Cross-Modal Attention for Audio-Visual Zero-Shot Learning

A significant portion of existing video classification research, including work on zero-shot learning, has focused on uni-modal approaches [Bra+20; CZ17; Fei+19], relying solely on visual information. However, a more comprehensive and realistic approach to video understanding requires incorporating audio and visual cues. Therefore, this thesis first addresses the task of audio-visual generalised zero-shot learning for video classification, as detailed in Chapter 2. This chapter specifically focuses on the generalised setting, where

models are evaluated not only on their ability to classify unseen classes but also on their ability to retain performance on previously seen classes.

Unlike prior work from Mercea et al. [Mer+22b], which relied on temporally-averaged features, the proposed Temporal and Cross-modal Attention Framework (TCAF) is designed to leverage the full temporal context within each modality. Moreover, TCAF introduced a novel cross-modal attention mechanism, explicitly guiding each modality to attend to the other modality. This inductive bias proved beneficial for learning representations that generalise well to unseen classes. In addition, TCAF also contains a novel loss formulation, simplifying the individual loss terms compared to previous work. The improvements in training objective, temporal information usage, and architectural design resulted in audio-visual representations that align better with the semantic information provided by class-label word embeddings, ultimately leading to superior zero-shot classification performance compared to previous methods.

6.1.2 Text-to-Feature Diffusion for Audio-Visual Few-Shot Learning

Chapter 3 extends the audio-visual generalised zero-shot setting from Chapter 2 to audio-visual generalised few-shot learning. While generalised zero-shot learning evaluates a model's ability to generalise to entirely unseen classes, acquiring large labelled datasets for all possible classes is often infeasible. Few-shot learning addresses this by assuming the availability of a limited number of labelled examples for novel classes.

This work introduces this novel task of generalised few-shot learning for audio-visual video classification, comprehensive training and evaluation protocols, and multiple new baselines. Three new benchmarks are proposed, designed to be compatible with the base and novel classes used in the zero-shot setting of Chapter 2, allowing for better analysis and comparisons across these learning paradigms. A state-of-the-art transformer framework, AV-DIFF, is presented, utilising a hybrid attention mechanism to fuse information from audio and visual streams across their temporal dimensions. This fusion mechanism demonstrated superior performance in the few-shot setting, outperforming standard multi-modal full-attention and the cross-attention employed in TCAF (Chapter 2). A key challenge in few-shot learning is mitigating bias introduced by the limited number of training examples. To address this, a novel text-to-feature diffusion model was proposed. This model is conditioned on semantic class representations derived from word embeddings, generating synthetic multi-modal features that augment the training. This approach increases the number and diversity of samples, leading to a more robust and less biased classifier. Semantic class representations from class-label word embeddings proved more effective than conditioning on audio-visual prototypes obtained through class-wide averaging of features, highlighting the benefits of language-derived semantics in enhancing generalisation.

6.1.3 Video-Adverb Retrieval with Compositional Adverb-Action Embeddings

Chapter 4 shifts the focus from video classification to improving fine-grained understanding of actions. Modelling and recognising adverbs that precisely describe how an action is performed in a video is crucial for better understanding events in videos beyond simply identifying *which* action is occurring.

This chapter introduces the framework ReGaDa for bidirectional video-adverb retrieval by learning a shared embedding space between video embeddings and corresponding compositional adverb-action embeddings. Previous work tackled this as a classification or regression problem [Mol+23] or by only utilising semantic action representations when learning a shared embedding space [Dou+20; DS22]. ReGaDa, however, utilises, in addition to action word embeddings, the semantic information from adverb embeddings.

One of the core innovations of this work lies in how the compositional adverb-action text embeddings are learned. ReGaDa explicitly exploits the compositionality of actions and adverbs when learning adverb-action text representations. For this, ReGaDa uses a residual gating mechanism to compose adverb and action word embeddings in a learned manner. Adverbs are used as gated residuals on top of a learned composition, allowing us to preserve relevant semantic adverb information for the task. This is coupled with a novel training objective that combines triplet losses with a regression target that directly encourages the alignment of video and composed text embeddings.

A particular effort was also put into measuring and evaluating generalisation to unseen adverb-action compositions, as it is likely that practical systems encounter specific compositions that they have not encountered during training. Given the scarcity of suitable benchmarks prior to this work, two novel dataset splits, designed according to zero-shot learning principles, are introduced to assess video-adverb retrieval for unseen compositions specifically. Achieving state-of-the-art performance on five established video-adverb retrieval benchmarks, ReGaDa also outperforms existing methods in generalisation, highlighting its ability to capture the compositional relationship between adverbs and actions effectively.

6.1.4 EgoCVR: An Egocentric Benchmark for Fine-Grained Composed Video Retrieval

Chapter 5 continues to focus on fine-grained video understanding but explores with composed video retrieval (CVR) a more complex task in which natural language plays a much more important role. The goal of this task is to retrieve a target video from a database, given a video and a textual description that specifies how the target video differs from the reference. This requires reasoning across the visual content of the reference video and the compositional changes indicated by the text query. For true CVR, the task should not be solved with a single frame. The previous benchmark [Ven+24], however, focuses primarily on modifications at the object level, for example replacing the object in the

reference video. As these types of modifications neglect the temporal aspects inherent in videos, Chapter 5 introduces the egocentric benchmark `EgoCVR` to allow for a more faithful evaluation of methods on their cross-modal temporal video understanding. Our results suggest that current VLMs have only limited fine-grained understanding and temporal reasoning capabilities. To address this, a modular training-free re-ranking framework, `TFR-CVR`, is proposed. It leverages an LLM for the core compositional reasoning by converting the complex vision-language problem into a language-based reasoning task. Through its modular design, it is possible to combine the strength of different off-the-shelf foundation models, upgrade models with better versions over time, and for interpretable post-hoc introspection of failure cases. `TFR-CVR` demonstrates improved performance compared to existing approaches on the newly proposed `EgoCVR` benchmark, highlighting the effectiveness of this language-based reasoning. Notably, our training-free approach outperforms prior methods on CVR trained on two million videos.

6.2 Discussion of Collective Contributions

Collectively, Chapters 2 to 5 advance video understanding by utilising language across a spectrum of tasks and learning scenarios. The unifying theme across all chapters is the strategic use of language, whether in class labels, adverbial modifiers, or complex textual instructions, to shape, constrain, and enhance the learning process. While the individual contributions of each chapter address distinct tasks, ranging from audio-visual classification to fine-grained compositional retrieval, their collective impact highlights advancements in three key areas:

1. Advancing multi-modal learning for video understanding by developing new methods to fuse and align representations across modalities effectively.
2. Enhancing compositional understanding in video and language within and across modalities to achieve a more fine-grained video understanding.
3. Establishing new benchmarks and evaluation protocols to facilitate further research in video understanding.

The first core contribution of this thesis lies in its **advancement of multi-modal learning for video understanding**, particularly how models integrate and leverage vision, audio, and language. In Chapters 2 and 3, we developed novel transformer-based attention mechanisms for effective cross-modal and temporal representation fusion. By introducing constraints on the attention process, these works effectively introduce inductive biases that shape the learned representations to the requirements of the given task, learning to selectively attend to relevant cues across modalities and over time. TCAF (Chapter 2) constrains the attention to pure cross-attention, while AV-DIFF (Chapter 3) employs a hybrid attention mechanism, applying first within-modality and then cross-modality fusion. This highlights the challenges and potential of multi-modal fusion, especially when

dealing with limited data. The need for balancing within- and cross-modality interactions suggests that no single fusion strategy is globally optimal. Therefore, adaptive methods that can be tailored to specific tasks and datasets are required. Additionally, limited data can amplify overfitting, where models might capture spurious correlations instead of learning robust, generalisable relationships. In addition, the thesis improves on training objectives to promote better alignment between modalities. TCaF (Chapter 2) simplifies the loss formulation compared to prior work, leading to improved zero-shot learning performance, while ReGAdA adds an additional regression target between video and composed text embeddings that encourages the alignment of representations in the shared embedding space. This illustrates a broader implication: practical multi-modal learning benefits not only from architectural designs that improve cross-modal interaction but also from training objectives that directly enforce alignment and shared understanding between modalities. Effective training objectives ensure that models learn meaningful multi-modal representations rather than relying on weak correlations. By incorporating constraints that guide the learning process, these approaches develop stronger semantic correspondences between modalities, enhancing generalisation and robustness in downstream applications.

Secondly, the thesis contributes to **enhancing compositional understanding in video and language**. In Chapters 4 and 5, we address the need for fine-grained video understanding by focusing on compositional tasks. Enhancing compositional video understanding requires more than a multi-modal fusion of information. It requires addressing the interplay between dynamic visual content and linguistic semantics. The work in this thesis goes beyond identifying *what* is present but studies the nuances of *how* actions are performed and *how* videos are modified. By exploring both intra-modal compositionality within language (Chapter 4) and inter-modal compositionality between vision and language (Chapter 5), we developed models that enable more fine-grained video understanding.

ReGAdA (Chapter 4) demonstrates the effectiveness of explicitly modelling compositional relationships within language. By introducing a residual gating mechanism to learn how actions and adverbs combine, we could better preserve the nuanced effects of adverbial modifiers on actions and, in turn, improve the alignment between the linguistic compositions and the action-focused visual representations. Chapter 5, with the `TFR-CVR` framework, extends compositional understanding to the interaction between modalities. CVR itself is inherently compositional because it requires understanding how a textual modifier transforms the content of a reference video. This requires reasoning about the temporal dynamics of the video and how the text alters those dynamics. However, the results presented in Chapter 5 demonstrate that current VLMs, despite their successes in other areas, often struggle with such fine-grained compositional video understanding. They may identify relevant objects or actions but fail to reason accurately across modalities. `TFR-CVR` overcomes this limitation by shifting the core compositional reasoning burden to an LLM. Instead of relying on the VLM to directly perform the complex cross-modal comparison, `TFR-CVR` transforms the problem into a language-based reasoning task. This approach effectively decouples the visual understanding (performed by the off-the-shelf

video captioning model) from the compositional reasoning (performed by the LLM).

Finally, a core contribution of this thesis lies in **establishing new benchmarks and evaluation protocols** that will enable future research in video understanding. The benchmarks introduced in this thesis were not created in isolation but were driven by a need to overcome specific shortcomings that hindered an effective evaluation of model capabilities. For instance, in Chapter 5, we identify a crucial gap in the existing evaluation benchmark [Ven+24] for composed video retrieval: a lack of focus on temporal understanding. Instead, it predominantly focuses on object-level modifications. The introduction of `EgoCVR`, focusing on egocentric videos, directly addresses this by providing a more challenging and realistic setting for evaluating a model's ability to reason about temporal events. As egocentric videos capture the continuous flow of actions and interactions from a first-person perspective, they allow us to reason about temporal context in a way that time-invariant object modifications do not. `EgoCVR` provides a more challenging setting for evaluating the ability of models to understand actions and their modifications and encourages the development of models with better temporal understanding capabilities. Similarly, the few-shot learning benchmarks and protocols established in Chapter 3 move beyond the often unrealistic assumption of abundant labelled data for all classes. We established a novel few-shot learning setting for audio-visual video classification, including three benchmarks, multiple baselines, and a comprehensive evaluation protocol. This work enables a more relevant evaluation of a model's generalisability in a multi-modal setting. By ensuring compatibility of the newly proposed audio-visual few-shot setting and the previous audio-visual zero-shot setting of Chapter 2, the thesis contributes to better comparisons between results obtained in the zero-shot setting and the new few-shot setting. Complementing this, we introduced in Chapter 4 additional zero-shot evaluation splits for video-adverb retrieval. This enables a more precise examination of how models generalise to unseen adverb-action compositions, advancing fine-grained action understanding beyond action recognition. Collectively, these benchmarks and protocols encourage the development of models with enhanced temporal reasoning, generalisation, and compositional understanding capabilities that better capture the complexities of real-world video data.

6.3 Limitations, Future Directions & Conclusion

This thesis has explored various directions to improve video understanding. While significant progress has been made, the presented work is not without limitations. First, Chapters 2 to 4 rely on pre-trained, frozen feature extractors for video input, and Chapters 2 and 3 additionally use them for audio. Although computationally efficient, this approach restricts a model's ability to adapt to the specific datasets and tasks addressed in each chapter. The features, pre-trained on large, general video and audio datasets, might not fully capture the specific visual, auditory, or temporal characteristics crucial for each task. This limitation is particularly relevant when a task requires detailed information that is

not necessarily captured by the pre-training objectives of the feature extractor. Training or fine-tuning the feature extractors simultaneously while aligning different modalities in the shared space would increase the computational requirements significantly but could potentially lead to improvements in performance and task-specific feature representation.

Furthermore, Chapters 2 to 4 use word embeddings to represent the semantic content of class labels associated with the videos. While Chapters 2 and 3 use a singular representation for the audio-visual class label (e.g. *playing piano*), Chapter 4 uses separate embeddings for actions (e.g. *cut*) and adverbs (e.g. *quickly*). While word embeddings provide an efficient way to capture semantic relationships between words, they may not fully encapsulate the complex interplay between visual information and the semantic concept encoded in the word embeddings. Visual variations of the same action can be significant. A video for *playing piano* can, for instance, display vastly different visual appearances, like close-ups of hands or wide shots of a stage. A single word embedding might not fully capture these nuances. Additionally, combining actions and adverbs presents a challenge in compositionality because the visual effects of adverbs vary a lot between different actions. A potential solution lies in representing the class labels, or their compositions, with more detailed descriptions by leveraging external sources or LLMs.

Another limitation is present in Chapter 5, where we introduce a high-quality egocentric benchmark for composed video retrieval. `EgoCVR` specifically focuses on temporal video understanding, an aspect not prioritized in previous benchmarks. Our findings show that existing VLMs struggle with understanding fine-grained composed videos. However, no high-quality training dataset existed that would have allowed finetuning models for the task. While Wu et al. [Wu+25] recently introduced with FineCVR-1M a large-scale dataset with a more considerable emphasis on temporal understanding, it does not comprise egocentric videos. Consequently, our `TFR-CVR`, designed explicitly for egocentric content, outperforms methods trained on FineCVR-1M on `EgoCVR`. This highlights that although the field is gaining traction, a promising future research direction could be the development of a large-scale and high-quality egocentric training dataset.

Video understanding is a fast growing field with promising future research directions. Three future directions are highlighted in the following:

- **Creating large-scale and high-quality datasets focused on temporal reasoning:** Progress in research is often facilitated through the development of new datasets, tasks, and challenges. The often leading performance of proprietary models on video understanding tasks is partly attributed to their training on massive, non-public video datasets. This underscores the importance of data scale and the quality of datasets, particularly for videos. Despite that, we have not yet reached a plateau regarding the benefits of scaling video models and video training data. Furthermore, current benchmarks still show a substantial gap between current video systems and humans [Dan+25; Hey+24; LIF25; Li+25; Zha+25], especially in understanding and reasoning about fine-grained temporal dynamics. This persistent gap underscores

the need for the development of higher-quality, large-scale video training datasets to address these limitations.

- **Developing the next generation of video representations:** This thesis has leveraged pre-trained video encoders, relying on extracted video features or tokens as input. The processing of videos as sequential image tokens from sampled video frames using ViT-based architectures has become the standard in large-scale video architectures. However, the vast information content of video data significantly impacts the computational cost of processing long sequences. Therefore, an important future direction lies in the investigation and development of stronger video representations and specialised architectural designs for videos. Several promising avenues include, for instance, the work of Carreira et al. [Car+24], which explores the learning of a video model from a single continuous video stream, with the goal of creating a model that can further adapt to the environment after deployment. Steenkiste et al. [Ste+24] try to learn scene-grounded video representations by decoupling tokens from the image grid that can encode and track scene content through time. Advancements like these will be crucial for addressing the increasing demands of real-world applications and the necessity for efficient and strong video representations.
- **VideoLLMs:** A currently emerging trend that is not covered in this thesis is multi-modal LLMs (MLLMs) [Che+24; Ope23; Tea+24; Wan+24; Ye+24] focused on video understanding. These models aim to equip language models with the ability to process and reason on different input modalities, such as images, audio, and videos. To enable the language model to process these modalities, an encoder extracts tokens, and a connector aligns those modality tokens with the LLM's text-based tokens. This results in powerful models for open-world video understanding that can handle various multi-modal tasks like captioning, question answering, retrieval, reasoning, and classification. For processing videos, these models usually use sample image frames that are processed with encoders shared between image and video processing. Due to the sheer volume of video information and the high computational costs of encoding long token sequences, MLLMs often sparsely sample video frames. Developing video MLLMs that can capture both fine-grained details and process long-form content is, therefore, still an open research problem. Future research should focus on developing more efficient VideoLLMs that capture temporal dynamics without losing fine-grained understanding.

In conclusion, this thesis demonstrates how integrating language information of various levels of complexity enables better generalisation and more nuanced and ultimately more robust and accurate video understanding. Importantly, it shows that guidance with language is not just an auxiliary feature but a fundamental tool for structuring representations, guiding retrieval, and improving model generalisation. While this thesis has contributed to improving video understanding, it also opened up avenues for future research for the development of more intelligent video systems.

Bibliography

[Abu+16] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. "Youtube-8m: A Large-Scale Video Classification Benchmark". 2016. arXiv: 1609.08675 (cit. on pp. 31, 105).

[AL18] J. Adler and S. Lunz. "Banach Wasserstein Gan". In: *NeurIPS*. 2018 (cit. on p. 30).

[Afo+22] T. Afouras, Y. M. Asano, F. Fagan, A. Vedaldi, and F. Metze. "Self-Supervised Object Detection from Audio-Visual Correspondence". In: *CVPR*. 2022 (cit. on pp. 13, 29).

[Afo+18] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. "Deep Audio-Visual Speech Recognition". In: *IEEE TPAMI*. 2018 (cit. on pp. 13, 29).

[ACZ20] T. Afouras, J. S. Chung, and A. Zisserman. "ASR Is All You Need: Cross-modal Distillation for Lip Reading". In: *ICASSP*. 2020 (cit. on pp. 13, 29).

[Afo+20] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman. "Self-Supervised Learning of Audio-Visual Objects from Video". In: *ECCV*. 2020 (cit. on pp. 13, 29).

[Aka+15a] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. "Label-Embedding for Image Classification". In: *IEEE TPAMI*. 2015 (cit. on p. 13).

[Aka+15b] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. "Evaluation of Output Embeddings for Fine-Grained Image Classification". In: *CVPR*. 2015 (cit. on p. 13).

[Ala+20] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman. "Self-Supervised Multimodal Versatile Networks". In: *NeurIPS*. 2020 (cit. on p. 43).

[Alh+21] T. Alhersh, H. Stuckenschmidt, A. U. Rehman, and S. B. Belhaouari. "Learning Human Activity from Visual Data Using Deep Learning". In: *IEEE Access*. 2021 (cit. on p. 41).

[Alw+20] H. Alwassel, D. Mahajan, L. Torresani, B. Ghanem, and D. Tran. "Self-Supervised Learning by Cross-Modal Audio-Video Clustering". In: *NeurIPS*. 2020 (cit. on pp. 13, 29).

[Ann+17] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell. "Localizing Moments in Video with Natural Language". In: *ICCV*. 2017 (cit. on pp. 2, 42).

[ALK21] M. U. Anwaar, E. Labintcev, and M. Kleinsteuber. "Compositional Learning of Image-Text Query for Image Retrieval". In: *WACV*. 2021 (cit. on p. 55).

[AZ18] R. Arandjelovic and A. Zisserman. "Objects That Sound". In: *ECCV*. 2018 (cit. on pp. 13, 29).

[ACB17] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein Generative Adversarial Networks". In: *ICML*. 2017 (cit. on p. 38).

[Arn+21] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid. "ViViT: A Video Vision Transformer". In: *ICCV*. 2021 (cit. on p. 2).

[Asa+20] Y. Asano, M. Patrick, C. Rupprecht, and A. Vedaldi. "Labelling Unlabelled Videos from Scratch with Multi-Modal Self-Supervision". In: *NeurIPS*. 2020 (cit. on pp. 19, 99, 104).

[Ash+24] K. Ashutosh, Z. Xue, T. Nagarajan, and K. Grauman. "Detours for Navigating Instructional Videos". In: *CVPR*. 2024 (cit. on p. 56).

[AVT16] Y. Aytar, C. Vondrick, and A. Torralba. "Soundnet: Learning Sound Representations from Unlabeled Video". In: *NeurIPS*. 2016 (cit. on pp. 13, 29).

[BKH16] J. L. Ba, J. R. Kiros, and G. E. Hinton. "Layer Normalization". 2016. arXiv: 1607.06450 (cit. on pp. 17, 45).

[Bai+24] Y. Bai, X. Xu, Y. Liu, S. Khan, F. Khan, W. Zuo, R. S. M. Goh, and C.-M. Feng. "Sentence-Level Prompts Benefit Composed Image Retrieval". In: *ICLR*. 2024 (cit. on p. 56).

[Bai+21] M. Bain, A. Nagrani, G. Varol, and A. Zisserman. "Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval". In: *ICCV*. 2021 (cit. on pp. 2–4, 42, 55).

[Bai+22] M. Bain, A. Nagrani, G. Varol, and A. Zisserman. "A CLIP-Hitchhiker's Guide to Long Video Retrieval". 2022. arXiv: 2205.08508 (cit. on p. 55).

[Bal+23] A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo. "Zero-Shot Composed Image Retrieval with Textual Inversion". In: *ICCV*. 2023 (cit. on p. 56).

[Bal+22] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo. "Effective Conditioned and Composed Image Retrieval Combining Clip-Based Features". In: *CVPR Workshops*. 2022 (cit. on p. 55).

[BWT21] G. Bertasius, H. Wang, and L. Torresani. "Is Space-Time Attention All You Need for Video Understanding?" 2021. arXiv: 2102.05095 (cit. on p. 2).

[BZP19] M. Bishay, G. Zoumpourlis, and I. Patras. "Tarn: Temporal Attentive Relation Network for Few-Shot and Zero-Shot Action Recognition". In: *BMVC*. 2019 (cit. on p. 30).

[Bla+22] A. Blattmann, R. Rombach, K. Oktay, J. Müller, and B. Ommer. "Semi-Parametric Neural Image Synthesis". In: *NeurIPS*. 2022 (cit. on p. 30).

[BLH20] Y. Bo, Y. Lu, and W. He. "Few-Shot Learning of Video Action Recognition Only Based on Video Contents". In: *WACV*. 2020 (cit. on pp. 32, 36, 108).

[BV19] W. Boes and H. Van hamme. "Audiovisual Transformer Architectures for Large-Scale Classification and Synchronization of Weakly Labeled Audio Events". In: *ACM MM*. 2019 (cit. on pp. 14, 29).

[Boj+17] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. "Enriching Word Vectors with Subword Information". In: *TACL*. 2017 (cit. on p. 110).

[Bom+21] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. "On the Opportunities and Risks of Foundation Models". 2021. arXiv: 2108.07258 (cit. on p. 55).

[Bor+13] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. "Large-Scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs". In: *Acm Mm*. 2013 (cit. on p. 43).

[Bra+20] B. Brattoli, J. Tighe, F. Zhdanov, P. Perona, and K. Chalupka. "Rethinking Zero-Shot Video Classification: End-to-end Training for Realistic Applications". In: *CVPR*. 2020 (cit. on pp. 12, 69).

[Bro+20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. "Language Models Are Few-Shot Learners". In: *NeurIPS*. 2020 (cit. on p. 110).

[Cab+15] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. "Activitynet: A Large-Scale Video Benchmark for Human Activity Understanding". In: *CVPR*. 2015 (cit. on pp. 1, 31, 42, 109).

[Cao+20] K. Cao, J. Ji, Z. Cao, C.-Y. Chang, and J. C. Niebles. "Few-Shot Video Classification via Temporal Alignment". In: *CVPR*. 2020 (cit. on pp. 9, 28, 30).

[Car+22] J. Carreira, S. Koppula, D. Zoran, A. Recasens, C. Ionescu, O. Henaff, E. Shelhamer, R. Arandjelovic, M. Botvinick, O. Vinyals, et al. "Hierarchical Perceiver". 2022. arXiv: 2202.10890 (cit. on pp. 32, 36, 108).

[CZ17] J. Carreira and A. Zisserman. "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset". In: *CVPR*. 2017 (cit. on pp. 1, 2, 69).

[Car+24] J. Carreira, M. King, V. Patraucean, D. Gokay, C. Ionescu, Y. Yang, D. Zoran, J. Heyward, C. Doersch, Y. Aytar, et al. "Learning from one continuous video stream". In: *CVPR*. 2024 (cit. on p. 76).

[Cha+16] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha. "An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild". In: *ECCV*. 2016 (cit. on pp. 20, 36, 37, 107).

[CG14] C.-Y. Chen and K. Grauman. "Inferring Analogous Attributes". In: *CVPR*. 2014 (cit. on p. 43).

[Che+21a] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman. "Localizing Visual Sounds the Hard Way". In: *CVPR*. 2021 (cit. on pp. 13, 29).

[Che+20a] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. "Vggsound: A Large-Scale Audio-Visual Dataset". In: *ICASSP*. 2020 (cit. on pp. 9, 31).

[Che+19a] H. Chen, G. Ding, Z. Lin, S. Zhao, and J. Han. "Cross-Modal Image-Text Retrieval with Semantic Consistency". In: *ACM MM*. 2019 (cit. on p. 42).

[CL23] J. Chen and H. Lai. "Pretrain like You Inference: Masked Tuning Improves Zero-Shot Composed Image Retrieval". 2023. arXiv: 2311.07622 (cit. on p. 56).

[Che+20b] S. Chen, Y. Zhao, Q. Jin, and Q. Wu. "Fine-Grained Video-Text Retrieval with Hierarchical Graph Reasoning". In: *CVPR*. 2020 (cit. on p. 43).

[Che+23] S. Chen, H. Li, Q. Wang, Z. Zhao, M. Sun, X. Zhu, and J. Liu. "Vast: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset". In: *NeurIPS*. 2023 (cit. on pp. 3, 55).

[Che+19b] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang. "A Closer Look at Few-Shot Classification". 2019. arXiv: 1904.04232 (cit. on p. 30).

[CB20] Y. Chen and L. Bazzani. "Learning Joint Visual Semantic Matching Embeddings for Language-Guided Retrieval". In: *ECCV*. 2020 (cit. on p. 55).

[CGB20] Y. Chen, S. Gong, and L. Bazzani. "Image Search with Text Feedback by Visiolinguistic Attention Learning". In: *CVPR*. 2020 (cit. on p. 55).

[Che+21b] Y. Chen, Y. Xian, A. S. Koepke, Y. Shan, and Z. Akata. "Distilling Audio-Visual Knowledge by Compositional Contrastive Learning". In: *CVPR*. 2021 (cit. on pp. 13, 29).

[Che+24] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao, et al. "VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs". 2024. arXiv: 2406.07476 (cit. on p. 76).

[DG07] S. Danafar and N. Gheissari. "Action recognition for surveillance applications using optic flow and SVM". In: *ACCV*. 2007 (cit. on p. 2).

[Dan+25] R. Dang, Y. Yuan, W. Zhang, Y. Xin, B. Zhang, L. Li, L. Wang, Q. Zeng, X. Li, and L. Bing. "ECBench: Can Multi-modal Foundation Models Understand the Egocentric World? A Holistic Embodied Cognition Benchmark". 2025. arXiv: 2501.05031 (cit. on p. 75).

[Del+22] G. Delmas, R. S. Rezende, G. Csurka, and D. Larlus. "ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity". In: *ICLR*. 2022 (cit. on p. 55).

[Dev+19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *ACL*. 2019 (cit. on p. 14).

[DN21] P. Dhariwal and A. Nichol. "Diffusion Models Beat Gans on Image Synthesis". In: *NeurIPS*. 2021 (cit. on p. 30).

[Don+15] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. "Long-term recurrent convolutional networks for visual recognition and description". In: *CVPR*. 2015 (cit. on p. 2).

[DLS18] J. Dong, X. Li, and C. G. Snoek. "Predicting Visual Features from Text for Image and Video Caption Retrieval". In: *IEEE TMM*. 2018 (cit. on pp. 42, 55).

[Don+24] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, L. Li, and Z. Sui. "A Survey on In-Context Learning". In: *EMNLP*. 2024 (cit. on p. 58).

[Dos+21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *ICLR*. 2021 (cit. on pp. 2, 14, 61).

[Dou+20] H. Doughty, I. Laptev, W. Mayol-Cuevas, and D. Damen. "Action Modifiers: Learning from Adverbs in Instructional Videos". In: *CVPR*. 2020 (cit. on pp. 8, 41–43, 45–51, 71, 109, 110, 112).

[DS22] H. Doughty and C. G. Snoek. "How Do You Do It? Fine-grained Action Understanding with Pseudo-Adverbs". In: *CVPR*. 2022 (cit. on pp. 41–43, 47, 49–52, 71, 109–112).

[Dou+18] M. Douze, A. Szlam, B. Hariharan, and H. Jégou. "Low-Shot Learning with Large-Scale Diffusion". In: *CVPR*. 2018 (cit. on p. 30).

[Ess+21] P. Esser, R. Rombach, A. Blattmann, and B. Ommer. "Imagebart: Bidirectional Context with Multinomial Diffusion for Autoregressive Image Synthesis". In: *NeurIPS*. 2021 (cit. on p. 30).

[Far+09] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. "Describing Objects by Their Attributes". In: *CVPR*. 2009 (cit. on p. 13).

[FK20] H. M. Fayek and A. Kumar. "Large Scale Audiovisual Learning of Sounds with Weakly Labeled Data". In: *IJCAI*. 2020 (cit. on pp. 13, 20, 29, 32, 36, 99, 108).

[Fei+19] C. Feichtenhofer, H. Fan, J. Malik, and K. He. "Slowfast networks for video recognition". In: *ICCV*. 2019 (cit. on p. 69).

[For+19] M. Forbes, C. Kaeser-Chen, P. Sharma, and S. Belongie. "Neural Naturalist: Generating Fine-Grained Image Comparisons". In: *EMNLP*. 2019 (cit. on p. 56).

[Fro+13] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. "Devise: A Deep Visual-Semantic Embedding Model". In: *NeurIPS*. 2013 (cit. on p. 13).

[Gab+20] V. Gabeur, C. Sun, K. Alahari, and C. Schmid. "Multi-Modal Transformer for Video Retrieval". In: *ECCV*. 2020 (cit. on pp. 14, 29, 43).

[Gad+23] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, et al. "Datacomp: In Search of the next Generation of Multimodal Datasets". In: *NeurIPS*. 2023 (cit. on p. 61).

[Gan+20] C. Gan, D. Huang, P. Chen, J. B. Tenenbaum, and A. Torralba. "Foley Music: Learning to Generate Music from Videos". In: *ECCV*. 2020 (cit. on pp. 13, 29).

[GG19] R. Gao and K. Grauman. "Co-Separating Sounds of Visual Objects". In: *ICCV*. 2019 (cit. on pp. 13, 29).

[Gat+24] P. Gatti, K. G. Parikh, D. P. Paul, M. Gupta, and A. Mishra. "Composite Sketch+ Text Queries for Retrieving Objects with Elusive Names and Complex Interactions". In: *AAAI*. 2024 (cit. on p. 56).

[GEB15] L. A. Gatys, A. S. Ecker, and M. Bethge. "A Neural Algorithm of Artistic Style". 2015. arXiv: 1508.06576 (cit. on p. 30).

[Ge+22] Y. Ge, Y. Ge, X. Liu, D. Li, Y. Shan, X. Qie, and P. Luo. "Bridging Video-Text Retrieval with Multiple Choice Questions". In: *CVPR*. 2022 (cit. on p. 43).

[Gem+17] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events". In: *ICASSP*. 2017 (cit. on p. 9).

[Gir+23] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. "Imagebind: One Embedding Space to Bind Them All". In: *CVPR*. 2023 (cit. on p. 55).

[GM18] S. Goldstein and Y. Moses. "Guitar Music Transcription from Silent Video." In: *BMVC*. 2018 (cit. on pp. 13, 29).

[Goo+20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative Adversarial Networks". In: *Communications of the ACM* (2020) (cit. on p. 30).

[Gra+22] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. "Ego4d: Around the World in 3,000 Hours of Egocentric Video". In: *CVPR*. 2022 (cit. on pp. 1, 54, 55).

[Gra+24] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, et al. "Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives". In: *CVPR*. 2024 (cit. on pp. 1, 55).

[Gu+24a] G. Gu, S. Chun, W. Kim, H. Jun, Y. Kang, and S. Yun. "CompoDiff: Versatile Composed Image Retrieval With Latent Diffusion". In: *TMLR*. 2024 (cit. on p. 56).

[Gu+24b] G. Gu, S. Chun, W. Kim, Y. Kang, and S. Yun. "Language-Only Efficient Training of Zero-Shot Composed Image Retrieval". In: *CVPR*. 2024 (cit. on p. 56).

[HSR19] M. Hahn, A. Silva, and J. M. Rehg. "Action2vec: A Crossmodal Embedding Approach to Action Learning". 2019. arXiv: 1901.00484 (cit. on p. 43).

[Han+17] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis. "Automatic Spatially-Aware Fashion Concept Discovery". In: *ICCV*. 2017 (cit. on pp. 55, 56).

[HG17] B. Hariharan and R. Girshick. "Low-Shot Visual Recognition by Shrinking and Hallucinating Features". In: *ICCV*. 2017 (cit. on pp. 29–31).

[HG16] D. Hendrycks and K. Gimpel. "Gaussian Error Linear Units (Gelus)". 2016. arXiv: 1606.08415 (cit. on p. 17).

[Her+17] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. "CNN Architectures for Large-Scale Audio Classification". In: *ICASSP*. 2017 (cit. on pp. 19, 31, 105).

[Hey+24] J. Heyward, J. Carreira, D. Damen, A. Zisserman, and V. Pătrăucean. "Perception Test 2024: Challenge Summary and a Novel Hour-Long VideoQA Benchmark". 2024. arXiv: 2411.1994 (cit. on p. 75).

[HJA20] J. Ho, A. Jain, and P. Abbeel. "Denoising Diffusion Probabilistic Models". In: *NeurIPS*. 2020 (cit. on p. 34).

[Hum+24] T. Hummel, S. Karthik, M.-I. Georgescu, and Z. Akata. "EgoCVR: An Egocentric Benchmark for Fine-Grained Composed Video Retrieval". In: *ECCV*. 2024 (cit. on pp. 4, 10).

[Hum+23] T. Hummel, O.-B. Mercea, A. S. Koepke, and Z. Akata. "Video-Adverb Retrieval with Compositional Adverb-Action Embeddings". In: *BMVC*. 2023 (cit. on pp. 4, 10).

[IR20] V. Iashin and E. Rahtu. "A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-Modal Transformer". In: *BMVC*. 2020 (cit. on pp. 14, 29).

[Ilh+21] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt. *OpenCLIP*. Version 0.1. 2021-07. DOI: 10.5281/zenodo.5143773. URL: https://doi.org/10.5281/zenodo.5143773 (cit. on p. 61).

[IS15] S. Ioffe and C. Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *ICML*. 2015 (cit. on pp. 15, 45).

[ILA15] P. Isola, J. J. Lim, and E. H. Adelson. "Discovering States and Transformations in Image Collections". In: *CVPR*. 2015 (cit. on p. 43).

[Iso+17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. "Image-to-Image Translation with Conditional Adversarial Networks". In: *CVPR*. 2017 (cit. on p. 30).

[Jae+21] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira. "Perceiver: General Perception with Iterative Attention". In: *ICML*. 2021 (cit. on pp. 20, 32, 36, 99, 100, 103, 108).

[JCZ19] A. Jamaludin, J. S. Chung, and A. Zisserman. "You Said That?: Synthesising Talking Faces from Audio". In: *IJCV*. 2019 (cit. on p. 13).

[Jia+21] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. "Scaling up Visual and Vision-Language Representation Learning with Noisy Text Supervision". In: *ICML*. 2021 (cit. on p. 55).

[Kan+20] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis. "Decoupling Representation and Classifier for Long-Tailed Recognition". In: *ICLR*. 2020 (cit. on p. 30).

[Kar+14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. "Large-Scale Video Classification with Convolutional Neural Networks". In: *CVPR*. 2014 (cit. on pp. 2, 19, 31, 105).

[Kar+24] S. Karthik, K. Roth, M. Mancini, and Z. Akata. "Vision-by-Language for Training-Free Compositional Image Retrieval". In: *ICLR*. 2024 (cit. on pp. 55, 56, 60, 62).

[Kay+17] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. "The kinetics human action video dataset". 2017. arXiv: 1705.06950 (cit. on p. 2).

[Ker+21] A. Kerrigan, K. Duarte, Y. Rawat, and M. Shah. "Reformulating Zero-Shot Action Recognition for Multi-Label Actions". In: *NeurIPS*. 2021 (cit. on p. 12).

[KC22] S. Kim and D.-W. Choi. "Better Generalized Few-Shot Learning Even without Base Data". 2022. arXiv: 2211.16095 (cit. on p. 30).

[KB14] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". 2014. arXiv: 1412.6980 (cit. on pp. 19, 36, 48).

[Koe+20] A. S. Koepke, O. Wiles, Y. Moses, and A. Zisserman. "Sight to Sound: An End-to-End Approach for Visual Piano Transcription". In: *ICASSP*. 2020 (cit. on pp. 13, 29).

[KWZ19] A. S. Koepke, O. Wiles, and A. Zisserman. "Visual Pitch Estimation". In: *SMC*. 2019 (cit. on pp. 13, 29).

[KTT18] B. Korbar, D. Tran, and L. Torresani. "Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization". In: *NeurIPS*. 2018 (cit. on pp. 13, 29).

[Kri+17] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles. "Dense-Captioning Events in Videos". In: *ICCV*. 2017 (cit. on p. 42).

[Kue+11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. "HMDB: A large video database for human motion recognition". In: *ICCV*. 2011 (cit. on p. 2).

[Kum+19] S. Kumar Dwivedi, V. Gupta, R. Mitra, S. Ahmed, and A. Jain. "Protogan: Towards Few Shot Learning for Action Recognition". In: *ICCVW*. 2019 (cit. on pp. 29, 30, 32, 36, 39, 108).

[Lap05] I. Laptev. "On space-time interest points". In: *IJCV*. 2005 (cit. on p. 2).

[LKH21] S. Lee, D. Kim, and B. Han. "Cosmo: Content-style Modulation for Image Retrieval with Text Feedback". In: *CVPR*. 2021 (cit. on p. 55).

[Lei+21] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu. "Less Is More: Clipbert for Video-and-Language Learning via Sparse Sampling". In: *CVPR*. 2021 (cit. on pp. 4, 43).

[Lev+24] M. Levy, R. Ben-Ari, N. Darshan, and D. Lischinski. "Data Roaming and Early Fusion for Composed Image Retrieval". In: *AAAI*. 2024 (cit. on p. 56).

[LIF25] C. Li, E. W. Im, and P. Fazli. "VidHalluc: Evaluating Temporal Hallucinations in Multimodal Large Language Models for Video Understanding". In: *CVPR*. 2025 (cit. on p. 75).

[Li+20a] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang. "Unicoder-vl: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training". In: *AAAI*. 2020 (cit. on p. 14).

[Li+23] J. Li, D. Li, S. Savarese, and S. Hoi. "Blip-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models". In: *ICML*. 2023 (cit. on pp. 3, 55).

[Li+22] J. Li, D. Li, C. Xiong, and S. Hoi. "Blip: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation". In: *ICML*. 2022 (cit. on pp. 3, 54, 55, 60, 61, 115, 117).

[Li+24] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo, et al. "MVBench: A comprehensive multi-modal video understanding benchmark". In: *CVPR*. 2024 (cit. on p. 2).

[Li+19a] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. "Visualbert: A Simple and Performant Baseline for Vision and Language". 2019. arXiv: 1908.03557 (cit. on p. 14).

[Li+19b] X. Li, Q. Sun, Y. Liu, Q. Zhou, S. Zheng, T.-S. Chua, and B. Schiele. "Learning to Self-Train for Semi-Supervised Few-Shot Classification". In: *NeurIPS*. 2019 (cit. on p. 30).

[Li+25] Y. Li, J. Niu, Z. Miao, C. Ge, Y. Zhou, Q. He, X. Dong, H. Duan, S. Ding, R. Qian, et al. "OVBench: How Far is Your Video-LLMs from Real-World Online Video Understanding?" In: *CVPR*. 2025 (cit. on p. 75).

[Li+20b] Y.-L. Li, Y. Xu, X. Mao, and C. Lu. "Symmetry and Group in Attribute-Object Compositions". In: *CVPR*. 2020 (cit. on p. 43).

[Lin+23a] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan. "Video-Llava: Learning United Visual Representation by Alignment before Projection". 2023. arXiv: 2311.10122 (cit. on p. 65).

[Lin+22a] C.-C. Lin, K. Lin, L. Wang, Z. Liu, and L. Li. "Cross-Modal Representation Learning for Zero-Shot Action Recognition". In: *CVPR*. 2022 (cit. on p. 12).

[Lin+22b] K. Q. Lin, J. Wang, M. Soldan, M. Wray, R. Yan, E. Z. XU, D. Gao, R.-C. Tu, W. Zhao, W. Kong, et al. "Egocentric Video-Language Pretraining". In: *NeurIPS*. 2022 (cit. on p. 55).

[Lin+23b] K. Q. Lin, P. Zhang, J. Chen, S. Pramanick, D. Gao, A. J. Wang, R. Yan, and M. Z. Shou. "Univtg: Towards unified video-language temporal grounding". In: *ICCV*. 2023 (cit. on p. 4).

[Lin+14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. "Microsoft Coco: Common Objects in Context". In: *ECCV*. 2014 (cit. on p. 61).

[LW20] Y.-B. Lin and Y.-C. F. Wang. "Audiovisual Transformer with Instance Attention for Audio-Visual Event Localization". In: *ACCV*. 2020 (cit. on pp. 14, 29).

[Liu+21a] S. Liu, H. Fan, S. Qian, Y. Chen, W. Ding, and Z. Wang. "HiT: Hierarchical Transformer with Momentum Contrast for Video-Text Retrieval". In: *ICCV*. 2021 (cit. on p. 14).

[Liu+18] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S. J. Hwang, and Y. Yang. "Learning to Propagate Labels: Transductive Propagation Network for Few-Shot Learning". 2018. arXiv: 1805.10002 (cit. on p. 30).

[Liu+19a] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman. "Use What You Have: Video Retrieval Using Representations from Collaborative Experts". In: *BMVC*. 2019 (cit. on p. 43).

[Liu+19b] Y. Liu, J. Guo, D. Cai, and X. He. "Attribute Attention for Semantic Disambiguation in Zero-Shot Learning". In: *CVPR*. 2019 (cit. on p. 13).

[Liu+23] Y. Liu, J. Yao, Y. Zhang, Y. Wang, and W. Xie. "Zero-Shot Composed Text-Image Retrieval". In: *BMVC*. 2023 (cit. on p. 56).

[Liu+21b] Z. Liu, C. Rodriguez-Opazo, D. Teney, and S. Gould. "Image Retrieval on Real-Life Images with Pre-Trained Vision-and-Language Models". In: *ICCV*. 2021 (cit. on pp. 53, 56, 58).

[Lu+19] J. Lu, D. Batra, D. Parikh, and S. Lee. "Vilbert: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks". In: *NeurIPS*. 2019 (cit. on p. 14).

[Luo+22] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li. "CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval and Captioning". In: *Neurocomputing*. 2022 (cit. on pp. 43, 55).

[Maj+22] S. Majumder, C. Chen, Z. Al-Halah, and K. Grauman. "Few-Shot Audio-Visual Learning of Environment Acoustics". In: *NeurIPS*. 2022 (cit. on p. 28).

[Man+21] M. Mancini, M. F. Naeem, Y. Xian, and Z. Akata. "Open World Compositional Zero-Shot Learning". In: *CVPR*. 2021 (cit. on pp. 43, 110).

[Maz+21] P. Mazumder, P. Singh, K. K. Parida, and V. P. Namboodiri. "Avgzslnet: Audio-visual Generalized Zero-Shot Learning by Reconstructing Label Features from Multi-Modal Embeddings". In: *WACV*. 2021 (cit. on pp. 12, 14, 20).

[Mer+22a] O.-B. Mercea, T. Hummel, A. S. Koepke, and Z. Akata. "Temporal and Cross-Modal Attention for Audio-Visual Zero-Shot Learning". In: *ECCV*. 2022 (cit. on pp. 4, 10, 28, 29, 31–33, 36–38, 107, 108, 110).

[Mer+23] O.-B. Mercea, T. Hummel, A. S. Koepke, and Z. Akata. "Text-to-feature diffusion for audio-visual few-shot learning". In: *DAGM GCPR*. 2023 (cit. on pp. 4, 10).

[Mer+22b] O.-B. Mercea, L. Riesch, A. S. Koepke, and Z. Akata. "Audio-Visual Generalised Zero-Shot Learning with Cross-Modal Attention and Language". In: *CVPR*. 2022 (cit. on pp. 5, 8, 12, 14, 19, 20, 28, 29, 31, 33, 36, 37, 70, 99, 100, 102–104, 107, 108, 110).

[Mie+20] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman. "End-to-End Learning of Visual Representations from Uncurated Instructional Videos". In: *CVPR*. 2020 (cit. on pp. 41, 43).

[MLS18] A. Miech, I. Laptev, and J. Sivic. "Learning a Text-Video Embedding from Incomplete and Heterogeneous Data". 2018. arXiv: 1804.02516 (cit. on p. 43).

[Mie+19] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. "HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips". In: *ICCV*. 2019 (cit. on pp. 3, 42, 43, 48, 55).

[Mik+13] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient Estimation of Word Representations in Vector Space". In: *ICLR*. 2013 (cit. on pp. 13, 35, 39, 110).

[Min+20] S. Min, H. Yao, H. Xie, C. Wang, Z.-J. Zha, and Y. Zhang. "Domain-Aware Visual Bias Eliminating for Generalized Zero-Shot Learning". In: *CVPR*. 2020 (cit. on p. 107).

[MO14] M. Mirza and S. Osindero. "Conditional Generative Adversarial Nets". 2014. arXiv: 1411.1784 (cit. on p. 30).

[MGH17] I. Misra, A. Gupta, and M. Hebert. "From Red Wine to Red Tomato: Composition with Context". In: *CVPR*. 2017 (cit. on p. 43).

[Mol+23] D. Moltisanti, F. Keller, H. Bilen, and L. Sevilla-Lara. "Learning Action Changes by Measuring Verb-Adverb Textual Relationships". In: *CVPR*. 2023 (cit. on pp. 41–43, 47–52, 71, 109–113).

[Mom+23] L. Momeni, M. Caron, A. Nagrani, A. Zisserman, and C. Schmid. "Verbs in Action: Improving Verb Understanding in Video-Language Models". 2023. arXiv: 2304.06708 (cit. on p. 43).

[Nae+21] M. F. Naeem, Y. Xian, F. Tombari, and Z. Akata. "Learning Graph Embeddings for Compositional Zero-Shot Learning". In: *CVPR*. 2021 (cit. on pp. 43, 110).

[Nag+22] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang. "Zero-shot temporal action detection via vision-language prompting". In: *ECCV*. 2022 (cit. on p. 4).

[NG18] T. Nagarajan and K. Grauman. "Attributes as Operators: Factorizing Unseen Attribute-Object Compositions". In: *ECCV*. 2018 (cit. on p. 43).

[Nag+21] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun. "Attention Bottlenecks for Multimodal Fusion". In: *NeurIPS*. 2021 (cit. on pp. 27–29, 32, 36, 108).

[NH10] V. Nair and G. E. Hinton. "Rectified Linear Units Improve Restricted Boltzmann Machines". In: *ICML*. 2010 (cit. on pp. 15, 45).

[Nan+19] Z. Nan, Y. Liu, N. Zheng, and S.-C. Zhu. "Recognizing Unseen Attribute-Object Pair with Generative Model". In: *AAAI*. 2019 (cit. on p. 43).

[Nar+21] M. Narasimhan, S. Ginosar, A. Owens, A. A. Efros, and T. Darrell. "Strumming to the Beat: Audio-conditioned Contrastive Video Textures". 2021. arXiv: 2104.02687 (cit. on pp. 13, 29).

[Nar+20] S. Narayan, A. Gupta, F. S. Khan, C. G. Snoek, and L. Shao. "Latent Embedding Feedback and Discriminative Features for Zero-Shot Classification". In: *ECCV*. 2020 (cit. on pp. 13, 29, 30).

[Ngu+24] T. Nguyen, Y. Bin, J. Xiao, L. Qu, Y. Li, J. Z. Wu, C.-D. Nguyen, S.-K. Ng, and L. A. Tuan. "Video-language understanding: A survey from model architecture, model training, and data perspectives". In: *ACL*. 2024 (cit. on pp. 2, 3).

[Onc+21] 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: *ICASSP*. 2021 (cit. on p. 42).

[Ope23] OpenAI. "GPT-4 Technical Report". 2023. arXiv: 2303.08774 (cit. on pp. 58, 76).

[Ota+16] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya. "Learning Joint Representations of Videos and Sentences with Web Image Search". In: *ECCV Workshops*. 2016 (cit. on pp. 42, 55).

[OE18] A. Owens and A. A. Efros. "Audio-Visual Scene Analysis with Self-Supervised Multisensory Features". In: *ECCV*. 2018 (cit. on pp. 13, 29).

[Owe+16] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. "Ambient Sound Provides Supervision for Visual Learning". In: *ECCV*. 2016 (cit. on pp. 13, 29).

[Owe+18] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. "Learning Sight from Sound: Ambient Sound Provides Supervision for Visual Learning". In: *IJCV*. 2018 (cit. on pp. 13, 29).

[Par+20] K. Parida, N. Matiyali, T. Guha, and G. Sharma. "Coordinated Joint Multimodal Embeddings for Generalized Audio-Visual Zero-Shot Classification and Retrieval of Videos". In: *WACV*. 2020 (cit. on pp. 12, 14, 20).

[Par+22] J. S. Park, S. Shen, A. Farhadi, T. Darrell, Y. Choi, and A. Rohrbach. "Exposing the Limits of Video-Text Models through Contrast Sets". In: *ACL*. 2022 (cit. on p. 43).

[Pat+20] M. Patrick, Y. M. Asano, R. Fong, J. F. Henriques, G. Zweig, and A. Vedaldi. "Multi-Modal Self-Supervision from Generalized Data Transformations". In: *NeurIPS*. 2020 (cit. on pp. 13, 27, 29).

[Pat+21] M. Patrick, P.-Y. Huang, Y. Asano, F. Metze, A. G. Hauptmann, J. F. Henriques, and A. Vedaldi. "Support-Set Bottlenecks for Video-Text Representation Learning". In: *ICLR*. 2021 (cit. on pp. 41, 43).

[PSM14] J. Pennington, R. Socher, and C. D. Manning. "GloVe: Global Vectors for Word Representation". In: *EMNLP*. 2014 (cit. on p. 110).

[Per+21] T. Perrett, A. Masullo, T. Burghardt, M. Mirmehdi, and D. Damen. "Temporal-Relational Crosstransformers for Few-Shot Action Recognition". In: *CVPR*. 2021 (cit. on p. 30).

[Pra+23] S. Pramanick, Y. Song, S. Nag, K. Q. Lin, H. Shah, M. Z. Shou, R. Chellappa, and P. Zhang. "EgoVLPv2: Egocentric Video-Language Pre-Training with Fusion in the Backbone". In: *ICCV*. 2023 (cit. on pp. 55, 60, 61, 116).

[QBL18] H. Qi, M. Brown, and D. G. Lowe. "Low-Shot Learning with Imprinted Weights". In: *CVPR*. 2018 (cit. on p. 30).

[Rad+21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. "Learning Transferable Visual Models from Natural Language Supervision". In: *ICML*. 2021 (cit. on pp. 3, 51, 52, 54, 55, 60, 61, 65, 110–112, 115, 117).

[Rad+19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. "Language Models Are Unsupervised Multitask Learners". In: *ICML*. 2019 (cit. on p. 14).

[RL17] S. Ravi and H. Larochelle. "Optimization as a Model for Few-Shot Learning". In: *ICLR*. 2017 (cit. on p. 30).

[Rec+23] A. Recasens, J. Lin, J. Carreira, D. Jaegle, L. Wang, J.-b. Alayrac, P. Luc, A. Miech, L. Smaira, R. Hemsley, et al. "Zorro: The Masked Multimodal Transformer". 2023. arXiv: 2301.09595 (cit. on pp. 32, 36, 108).

[Rom+22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *CVPR*. 2022 (cit. on p. 30).

[Rou+21] A. Rouditchenko, A. Boggust, D. Harwath, B. Chen, D. Joshi, S. Thomas, K. Audhkhasi, H. Kuehne, R. Panda, R. Feris, et al. "AVLnet: Learning Audio-Visual Language Representations from Instructional Videos". In: *Interspeech*. 2021 (cit. on p. 43).

[Roy+22] A. Roy, A. Shah, K. Shah, A. Roy, and R. Chellappa. "DiffAlign: Few-shot Learning Using Diffusion Based Synthesis and Alignment". 2022. arXiv: 2212.05404 (cit. on p. 30).

[Sai+23] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, and T. Pfister. "Pic2word: Mapping Pictures to Words for Zero-Shot Composed Image Retrieval". In: *CVPR*. 2023 (cit. on p. 56).

[SC21] D. Saxena and J. Cao. "Generative Adversarial Networks (GANs) Challenges, Solutions, and Future Directions". In: *ACM CSUR*. 2021 (cit. on p. 30).

[SLC04] C. Schuldt, I. Laptev, and B. Caputo. "Recognizing human actions: a local SVM approach". In: *ICPR*. 2004 (cit. on p. 2).

[Seo+22] P. H. Seo, A. Nagrani, A. Arnab, and C. Schmid. "End-to-end generative pretraining for multimodal video captioning". In: *CVPR*. 2022 (cit. on pp. 1, 4).

[SNS21] P. H. Seo, A. Nagrani, and C. Schmid. "Look before you speak: Visually contextualized utterances". In: *CVPR*. 2021 (cit. on p. 4).

[Ser+24] P. Sermanet, T. Ding, J. Zhao, F. Xia, D. Dwibedi, K. Gopalakrishnan, C. Chan, G. Dulac-Arnold, S. Maddineni, N. J. Joshi, et al. "RoboVQA: Multimodal long-horizon reasoning for robotics". In: *ICRA*. 2024 (cit. on p. 1).

[Sid04] H. Sidenbladh. "Detecting human motion with support vector machines". In: *ICPR*. 2004 (cit. on p. 2).

[SZ14] K. Simonyan and A. Zisserman. "Two-stream convolutional networks for action recognition in videos". In: *NeurIPS*. 2014 (cit. on p. 2).

[SSZ17] J. Snell, K. Swersky, and R. Zemel. "Prototypical Networks for Few-Shot Learning". In: *NeurIPS*. 2017 (cit. on p. 30).

[SZS12] K. Soomro, A. R. Zamir, and M. Shah. "UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild". 2012. arXiv: 1212.0402 (cit. on pp. 2, 31).

[Sri+14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *JMLR* (2014) (cit. on pp. 15, 45).

[Ste+24] S. van Steenkiste, D. Zoran, Y. Yang, Y. Rubanova, R. Kabra, C. Doersch, D. Gokay, E. Pot, K. Greff, D. Hudson, et al. "Moving off-the-grid: Scene-grounded video representations". In: *NeurIPS*. 2024 (cit. on p. 76).

[SLS20] K. Su, X. Liu, and E. Shlizerman. "Multi-Instrumentalist Net: Unsupervised Generation of Music from Body Movements". 2020. arXiv: 2012.03478 (cit. on pp. 13, 29).

[Su+19] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. "Vl-Bert: Pre-training of Generic Visual-Linguistic Representations". 2019. arXiv: 1908.08530 (cit. on p. 14).

[Sun+19a] C. Sun, F. Baradel, K. Murphy, and C. Schmid. "Learning Video Representations Using Contrastive Bidirectional Transformer". 2019. arXiv: 1906.05743 (cit. on p. 14).

[Sun+19b] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. "Videobert: A Joint Model for Video and Language Representation Learning". In: *ICCV*. 2019 (cit. on p. 14).

[SYG23] S. Sun, F. Ye, and S. Gong. "Training-Free Zero-Shot Composed Image Retrieval with Local Concept Reranking". 2023. arXiv: 2312.08924 (cit. on p. 56).

[Sun+18] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. "Learning to Compare: Relation Network for Few-Shot Learning". In: *CVPR*. 2018 (cit. on p. 30).

[TB19] H. Tan and M. Bansal. "Lxmert: Learning Cross-Modality Encoder Representations from Transformers". In: *EMNLP*. 2019 (cit. on p. 14).

[Tan+20a] S. Tanberk, Z. H. Kilimci, D. B. Tükel, M. Uysal, and S. Akyokuş. "A Hybrid Deep Model Using Deep Learning and Dense Optical Flow Approaches for Human Activity Recognition". In: *IEEE Access*. 2020 (cit. on p. 41).

[Tan+20b] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng. "Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains". In: *NeurIPS*. 2020 (cit. on pp. 15, 34).

[Tan+24] Y. Tang, J. Yu, K. Gai, Z. Jiamin, G. Xiong, Y. Hu, and Q. Wu. "Context-I2W: Mapping Images to Context-Dependent Words for Accurate Zero-Shot Composed Image Retrieval". In: *AAAI*. 2024 (cit. on p. 56).

[Tan+23] Y. Tang, J. Bi, S. Xu, L. Song, S. Liang, T. Wang, D. Zhang, J. An, J. Lin, R. Zhu, et al. "Video understanding with large language models: A survey". 2023. arXiv: 2312.17432 (cit. on pp. 2, 3).

[Tea+24] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context". 2024. arXiv: 2403.05530 (cit. on p. 76).

[Tha+24] O. Thawakar, M. Naseer, R. M. Anwer, S. Khan, M. Felsberg, M. Shah, and F. S. Khan. "Composed Video Retrieval via Enriched Context and Discriminative Embeddings". In: *CVPR*. 2024 (cit. on pp. 56, 60, 62).

[Tia+18] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. "Audio-Visual Event Localization in Unconstrained Videos". In: *ECCV*. 2018 (cit. on pp. 13, 29).

[TTS16] A. Torabi, N. Tandon, and L. Sigal. "Learning Language-Visual Embedding for Movie Understanding with Natural-Language". 2016. arXiv: 1609.08124 (cit. on pp. 42, 55).

[Tou+23] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models". 2023. arXiv: 2307.09288 (cit. on p. 54).

[Tra+15] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. "Learning Spatiotemporal Features with 3d Convolutional Networks". In: *ICCV*. 2015 (cit. on pp. 2, 19, 31, 105).

[VKK21] A. Vahdat, K. Kreis, and J. Kautz. "Score-Based Generative Modeling in Latent Space". In: *NeurIPS*. 2021 (cit. on p. 30).

[VH08] L. Van der Maaten and G. Hinton. "Visualizing Data Using T-SNE." In: *JMLR*. 2008 (cit. on p. 24).

[Vas+17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. "Attention Is All You Need". In: *NeurIPS*. 2017 (cit. on pp. 7, 14, 16, 34, 45).

[VCM23] S. Vaze, N. Carion, and I. Misra. "GeneCIS: A Benchmark for General Conditional Image Similarity". In: *CVPR*. 2023 (cit. on p. 56).

[Ven+24] L. Ventura, A. Yang, C. Schmid, and G. Varol. "CoVR: Learning Composed Video Retrieval from Web Video Captions". In: *AAAI*. 2024 (cit. on pp. 9, 53, 56–62, 66, 71, 74, 115–117, 120).

[Vin+16] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. "Matching Networks for One Shot Learning". In: *NeurIPS*. 2016 (cit. on p. 30).

[Vo+19] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays. "Composing Text and Image for Image Retrieval-an Empirical Odyssey". In: *CVPR*. 2019 (cit. on pp. 42–44, 48, 49, 53, 55, 112).

[Wah+11] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. "The Caltech-Ucsd Birds-200-2011 Dataset". In: *California Institute of Technology*. 2011 (cit. on p. 13).

[WS13] H. Wang and C. Schmid. "Action recognition with improved trajectories". In: *ICCV*. 2013 (cit. on p. 2).

[WZY21] X. Wang, L. Zhu, and Y. Yang. "T2vlad: Global-Local Sequence Alignment for Text-Video Retrieval". In: *CVPR*. 2021 (cit. on pp. 14, 29).

[WJ13] X. Wang and Q. Ji. "A Unified Probabilistic Approach Modeling Relationships between Attributes and Objects". In: *ICCV*. 2013 (cit. on p. 43).

[Wan+19a] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang. "VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research". In: *ICCV*. 2019 (cit. on pp. 42, 55, 109).

[Wan+18] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan. "Low-Shot Learning from Imaginary Data". In: *CVPR*. 2018 (cit. on p. 30).

[Wan+19b] Y. Wang, W.-L. Chao, K. Q. Weinberger, and L. van der Maaten. "Simpleshot: Revisiting Nearest-Neighbor Classification for Few-Shot Learning". 2019. arXiv: 1911.04623 (cit. on p. 30).

[WM10] Y. Wang and G. Mori. "A Discriminative Latent Model of Object Classes and Attributes". In: *ECCV*. 2010 (cit. on p. 43).

[Wan+24] Y. Wang, K. Li, X. Li, J. Yu, Y. He, G. Chen, B. Pei, R. Zheng, Z. Wang, Y. Shi, et al. "InternVideo2: Scaling foundation models for multimodal video understanding". In: *ECCV*. 2024 (cit. on pp. 3, 76).

[Wan+22] Y. Wang, K. Li, Y. Li, Y. He, B. Huang, Z. Zhao, H. Zhang, J. Xu, Y. Liu, Z. Wang, et al. "Internvideo: General video foundation models via generative and discriminative learning". 2022. arXiv: 2212.03191 (cit. on p. 3).

[WKZ18] O. Wiles, A. S. Koepke, and A. Zisserman. "X2face: A Network for Controlling Face Generation Using Images, Audio, and Pose Codes". In: *ECCV*. 2018 (cit. on p. 13).

[WD19] M. Wray and D. Damen. "Learning Visual Actions Using Multiple Verb-Only Labels". In: *BMVC*. 2019 (cit. on p. 43).

[Wra+19] M. Wray, D. Larlus, G. Csurka, and D. Damen. "Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings". In: *ICCV*. 2019 (cit. on pp. 42, 43).

[Wu+21] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris. "The Fashion IQ Dataset: Retrieving Images by Combining Side Information and Relative Natural Language Feedback". In: *CVPR*. 2021 (cit. on pp. 53, 55, 56).

[Wu+23] W. Wu, H. Luo, B. Fang, J. Wang, and W. Ouyang. "Cap4Video: What Can Auxiliary Captions Do for Text-Video Retrieval?" In: *CVPR*. 2023 (cit. on p. 43).

[Wu+25] Y. Wu, Z. Qi, Y. Wu, J. Sun, Y. Wang, and S. Wang. "Learning Fine-Grained Representations through Textual Token Disentanglement in Composed Video Retrieval". In: *ICLR*. 2025 (cit. on p. 75).

[Xia+16] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. "Latent Embeddings for Zero-Shot Classification". In: *CVPR*. 2016 (cit. on p. 110).

[Xia+21] Y. Xian, B. Korbar, M. Douze, L. Torresani, B. Schiele, and Z. Akata. "Generalized Few-Shot Video Classification with Video Retrieval and Feature Generation". In: *IEEE TPAMI*. 2021 (cit. on pp. 28–32, 36, 38, 108).

[Xia+18a] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. "Zero-Shot Learning—a Comprehensive Evaluation of the Good, the Bad and the Ugly". In: *IEEE TPAMI*. 2018 (cit. on pp. 13, 19).

[Xia+18b] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. "Feature Generating Networks for Zero-Shot Learning". In: *CVPR*. 2018 (cit. on p. 13).

[Xia+19] Y. Xian, S. Sharma, B. Schiele, and Z. Akata. "F-Vaegan-D2: A Feature Generating Framework for Any-Shot Learning". In: *CVPR*. 2019 (cit. on pp. 13, 29, 30).

[Xia+20] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer. "Audiovisual Slowfast Networks for Video Recognition". 2020. arXiv: 2001.08740 (cit. on pp. 27, 28).

[Xia+10] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. "Sun Database: Large-scale Scene Recognition from Abbey to Zoo". In: *CVPR*. 2010 (cit. on p. 13).

[Xie+19] G.-S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, J. Qin, Y. Yao, and L. Shao. "Attentive Region Embedding Network for Zero-Shot Learning". In: *CVPR*. 2019 (cit. on p. 13).

[Xie+18] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification". In: *ECCV*. 2018 (cit. on p. 2).

[Xu+15a] B. Xu, N. Wang, T. Chen, and M. Li. "Empirical Evaluation of Rectified Activations in Convolutional Network". 2015. arXiv: 1505.00853 (cit. on p. 45).

[Xu+23] H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang, et al. "mPLUG-2: A Modularized Multi-Modal Foundation Model across Text, Image and Video". In: *ICML*. 2023 (cit. on p. 55).

[Xu+21] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer. "Videoclip: Contrastive pre-training for zero-shot video-text understanding". In: *EMNLP*. 2021 (cit. on p. 3).

[Xu+17] H. Xu, Y. Gao, F. Yu, and T. Darrell. "End-to-end learning of driving models from large-scale video datasets". In: *CVPR*. 2017 (cit. on p. 1).

[Xu+16] J. Xu, T. Mei, T. Yao, and Y. Rui. "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language". In: *CVPR*. 2016 (cit. on pp. 42, 55, 109).

[Xu+15b] R. Xu, C. Xiong, W. Chen, and J. Corso. "Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified Framework". In: *AAAI*. 2015 (cit. on pp. 42, 43, 55).

[Xu+20] W. Xu, Y. Xian, J. Wang, B. Schiele, and Z. Akata. "Attribute Prototype Network for Zero-Shot Learning". In: *NeurIPS*. 2020 (cit. on p. 13).

[Yan+23] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid. "Vid2seq: Large-scale pretraining of a visual language model for dense video captioning". In: *CVPR*. 2023 (cit. on pp. 1, 2, 4).

[Ye+20] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha. "Few-Shot Learning via Embedding Adaptation with Set-to-Set Functions". In: *CVPR*. 2020 (cit. on p. 30).

[Ye+24] J. Ye, H. Xu, H. Liu, A. Hu, M. Yan, Q. Qian, J. Zhang, F. Huang, and J. Zhou. "mPLUG-Owl3: Towards long image-sequence understanding in multi-modal large language models". In: *ICLR*. 2024 (cit. on p. 76).

[Zha+23a] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. "Sigmoid loss for language image pre-training". In: *ICCV*. 2023 (cit. on p. 3).

[Zha+20] H. Zhang, L. Zhang, X. Qi, H. Li, P. H. Torr, and P. Koniusz. "Few-Shot Action Recognition with Permutation-Invariant Attention". In: *ECCV*. 2020 (cit. on pp. 9, 28).

[Zha+22] Y.-K. Zhang, D.-W. Zhou, H.-J. Ye, and D.-C. Zhan. "Audio-Visual Generalized Few-Shot Learning with Prototype-Based Co-Adaptation". In: *Interspeech*. 2022 (cit. on p. 28).

[Zha+24] L. Zhao, N. B. Gundavarapu, L. Yuan, H. Zhou, S. Yan, J. J. Sun, L. Friedman, R. Qian, T. Weyand, Y. Zhao, et al. "VideoPrism: A Foundational Visual Encoder for Video Understanding". 2024. arXiv: 2402.13217 (cit. on p. 65).

[Zha+25] Y. Zhao, L. Xie, H. Zhang, G. Gan, Y. Long, Z. Hu, T. Hu, W. Chen, C. Li, J. Song, et al. "MMVU: Measuring Expert-Level Multi-Discipline Video Understanding". In: *CVPR*. 2025 (cit. on p. 75).

[Zha+23b] Y. Zhao, I. Misra, P. Krähenbühl, and R. Girdhar. "Learning Video Representations from Large Language Models". In: *CVPR*. 2023 (cit. on pp. 55, 60, 65).

[Zho+19] H. Zhou, Z. Liu, X. Xu, P. Luo, and X. Wang. "Vision-Infused Deep Audio Inpainting". In: *ICCV*. 2019 (cit. on pp. 13, 29).

[ZXC18] L. Zhou, C. Xu, and J. Corso. "Towards Automatic Learning of Procedures from Web Instructional Videos". In: *AAAI*. 2018 (cit. on p. 42).

[Zhu+24] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, H. Wang, Y. Pang, W. Jiang, J. Zhang, Z. Li, et al. "LanguageBind: Extending Video-Language Pretraining to n-Modality by Language-Based Semantic Alignment". In: *ICLR*. 2024 (cit. on pp. 54, 55, 60, 61, 65, 116, 117).

[ZY18] L. Zhu and Y. Yang. "Compound Memory Networks for Few-Shot Video Classification". In: *ECCV*. 2018 (cit. on pp. 28, 30).

[ZY20] L. Zhu and Y. Yang. "Actbert: Learning Global-Local Video-Text Representations". In: *CVPR*. 2020 (cit. on p. 43).

[Zhu+19a] Y. Zhu, J. Xie, B. Liu, and A. Elgammal. "Learning Feature-to-Feature Translator by Alternating Back-Propagation for Generative Zero-Shot Learning". In: *ICCV*. 2019 (cit. on p. 13).

[Zhu+19b] D. Zhukov, J.-B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic. "Cross-Task Weakly Supervised Learning from Instructional Videos". In: *CVPR*. 2019 (cit. on p. 43).

A

Supplementary material: Temporal and cross-modal attention for audio-visual zero-shot learning

In the supplementary material, we provide additional details about baselines (Appendix A.1), and present further model ablations (Appendix A.2). Additionally, we study t-SNE visualisations for TCaF and [Mer+22b] (Appendix A.3), and provide a comparison of the computational complexity of TCaF and some of the baselines (Appendix A.4). Finally, we present further quantitative results for audio-visual (G)ZSL when using SeLaVi [Asa+20] features as inputs (Appendix A.5).

A.1 Additional Details about Baselines

In the following, we detail our adaptations of Attention Fusion [FK20] and of the Perceiver [Jae+21] to the (G)ZSL setting (which we briefly summarised in Section 2.4.2).

A.1.1 Attention Fusion

In order to use Attention Fusion [FK20] in the (G)ZSL setting, we take the same temporal audio and visual features as inputs as TCaF. Following TCaF, we embed the input features into the same feature dimension using A_{enc} and V_{enc}. Instead of directly mapping to the number of classes, as the authors originally proposed, A_{enc} and V_{enc} map the features to $\mathbb{R}^{d_{dim}}$. The embedded features are then temporally averaged to obtain a single $d_{dim}-$dimensional feature vector for each modality. The attention weight α, which is used for fusing both modalities, is computed using the channel-wise concatenation of the audio and visual embeddings through a linear layer $f_{attn} : \mathbb{R}^{2*d_{dim}} \rightarrow \mathbb{R}^{d_{dim}}$, followed by a sigmoid function. Both modalities are then fused to create the output token o_c through $o_c = \alpha \odot \phi_{a,avg} + (1 - \alpha) \odot \phi_{v,avg}$, where $\phi_{a,avg}$ and $\phi_{v,avg}$ are the temporally averaged audio and visual features. o_c is then projected using the same projection function O_{proj},

decoder D_o, and text embedding projections as in TCAF. We train Attention Fusion using the same learning rate and loss functions as TCAF.

A.1.2 Perceiver

The Perceiver [Jae+21] takes the same audio and visual features as input as TCAF. For consistency between frameworks, we again embedded the input features to the same feature dimension using A_{enc} and V_{enc}, and equip both TCAF and the Perceiver with the same temporal and modality information by adding positional embeddings as described in the main paper. Our goal was to directly compare our cross-attention mechanism with the Perceiver attention. Therefore, we adapted the cross-attention, self-attention and dense layer blocks of the Perceiver to use the same internal dimensions as TCAF. We also added a dropout layer at the end of dense layer blocks to match the dense blocks in TCAF. For the randomly initialised latent array, we use 64 latent tokens with dimension $\mathbb{R}^{d_{dim}}$ for all datasets. Increasing the number of latent tokens did not provide a boost in performance, but significantly increased the computational costs. One of the latent tokens is used as the output classification token c_o. We use one cross-attention block and one self-attention block per layer without weight sharing and use the same number of layers as TCAF. This results in just a slightly higher number of parameters for the Perceiver than for our TCAF. The output token c_o is projected using the projection function O_{proj} and the decoder D_o. The computations for the text embeddings are analogous to TCAF. We train the Perceiver using the same learning rate and loss functions as our model.

A.2 Additional Model Ablations

In this section, we first study the impact of using temporal embeddings (Appendix A.2.1) and of the number and design of the cross-attention layers in TCAF (Appendix A.2.2). Next, we evaluate the impact on performance when adding noise to the audio modality (Appendix A.2.3). Finally, we present results of transforming TCAF to [Mer+22b] (Appendix A.2.4).

A.2.1 Influence of Using Temporal Information

In the following, we investigate the influence of using temporal information when learning multi-modal video representation for (G)ZSL with TCAF. Since the operations in our audio-visual transformer layers (cf. Section 2.3.2) are invariant to permutation, the feature tokens are additionally equipped with temporal information through the addition of positional embeddings pos_t. Without temporal embeddings, the model is unable to put data from one time step in temporal relation to information from the other time steps. Temporal embeddings therefore allow the model to understand the concept of time.

Table A.1 shows results for training and evaluating TCAF with (+) and without (−) temporal embeddings (pos_t). The highest harmonic mean is achieved when using

Positional embeddings	VGGSound-GZSL^{cls}				UCF-GZSL^{cls}				ActivityNet-GZSL^{cls}			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
$-pos_t$	**15.78**	4.66	7.19	4.97	27.35	26.02	26.67	28.06	21.80	5.43	8.69	5.53
$+pos_t$ (TCAF)	12.63	**6.72**	**8.77**	**7.41**	**67.14**	**40.83**	**50.78**	**44.64**	**30.12**	**7.65**	**12.20**	**7.96**

Table A.1: Influence of temporal information provided through positional embeddings (pos_t) on the (G)ZSL performance on the VGGSound-GZSL^{cls}, UCF-GZSL^{cls}, and ActivityNet-GZSL^{cls} datasets.

Layer configurations	VGGSound-GZSL^{cls}				UCF-GZSL^{cls}				ActivityNet-GZSL^{cls}			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
1 layer w/o FF	19.70	4.47	7.29	4.66	63.30	26.45	37.31	27.85	15.10	4.59	7.04	4.63
1 layer	**17.95**	4.78	7.55	5.13	40.07	29.40	33.92	29.74	28.22	4.85	8.27	4.89
1/2*(all layers) w/o FF	11.33	4.25	6.18	4.59	38.72	23.17	28.99	23.28	8.13	3.35	4.75	3.40
1/2*(all layers)	12.08	4.69	6.75	5.12	**77.19**	30.18	43.40	34.18	28.65	6.04	9.98	6.25
1/2*(all layers) $+A_{self}$	14.62	4.56	6.96	4.97	53.05	34.83	42.05	35.84	**31.38**	5.93	9.97	6.51
all layers w/o FF	14.41	4.28	6.60	4.59	32.57	25.77	28.78	28.86	7.44	3.27	4.54	3.33
all layers	12.63	**6.72**	**8.77**	**7.41**	67.14	**40.83**	**50.78**	**44.64**	30.12	**7.65**	**12.20**	**7.96**

Table A.2: Varying the number of cross-attention layers in TCAF and the use of feed forward (FF) functions in the cross-attention layers.

temporal embeddings. For instance for ActivityNet-GZSL^{cls}, our model that does not use temporal embeddings ($-pos_t$) obtains only a HM of 8.69% and a ZSL score of 5.53%, compared to a HM of 12.20% and a ZSL score of 7.96% when using temporal embeddings. Similar observations can be made for VGGSound-GZSL^{cls} and UCF-GZSL^{cls}, showing the importance of temporal information for learning strong video representations.

A.2.2 Impact of Using Different Amounts of Cross-Attention Layers and of Varying the Cross-Attention Layer Design

In Table A.2, we present ablations on the number of cross-attention layers used in our model. Furthermore, we investigate the relevance of using feed forward functions (FF) in our cross-attention layers.

For TCAF, we used 8 cross-attention layers on VGGSound-GZSL^{cls} (all layers). On the UCF-GZSL^{cls} and ActivityNet-GZSL^{cls} datasets, we used 6 layers (all layers). We observe that using more layers is beneficial for GZSL and ZSL performance across all datasets. Moreover, we observe that, in general, eliminating the feed forward functions leads to a decrease in performance. Finally, using only half of the layers jointly with self-attention (1/2*(all layers) $+A_{self}$) leads to worse overall HM performance than using half of the layers without self-attention (1/2*(all layers)). This is in line with the experiments in the main paper, where adding the self-attention leads to worse results.

This ablation shows that using only cross-attention is beneficial even when using a different number of layers. Furthermore, using more cross-attention layers that are equipped with feed forward functions brings a boost in performance.

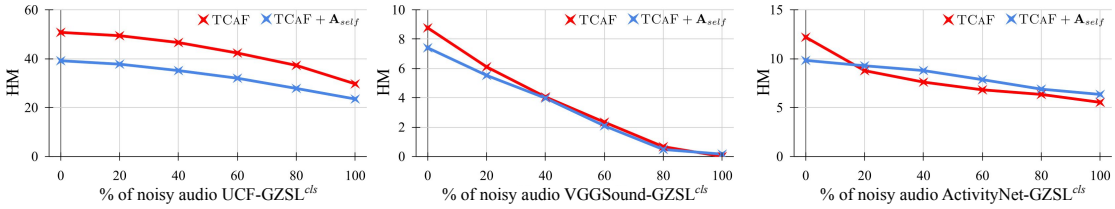

Figure A.1: Robustness of TCAF and TCAF +A_{self} to noise added to different proportions of the audio stream on UCF-GZSL^{cls}, VGGSound-GZSL^{cls} and ActivityNet-GZSL^{cls}.

A.2.3 Impact of Noise in Audio Stream on GZSL Performance

In this section, we study how the GZSL performance (HM) of TCAF decreases when noise is added to increasing temporal portions of the audio signal on all three datasets. We study both TCAF and TCAF +A_{self} in Figure A.1. It can be observed that an increase in the proportion of noise leads to a decrease in the GZSL performance for both models on all three datasets. Furthermore, it can be observed that TCAF is significantly more robust to perturbations on UCF-GZSL^{cls} and slightly more robust on VGGSound-GZSL^{cls}. On the other hand, we can observe that on ActivityNet-GZSL^{cls} the trend is reversed, with TCAF +A_{self} being slightly more robust. Overall, it can be argued that TCAF is more robust across all three datasets than TCAF +A_{self}.

A.2.4 Transforming TCAF into [Mer+22b]

Our TCAF builds on the AVCA [Mer+22b] framework for audio-visual GZSL. To highlight the benefits of TCAF compared to AVCA, we show results for transforming TCAF into AVCA [Mer+22b] in Table A.3.

TCAF exploits temporal information and obtains a HM performance of 8.77% on VGGSound-GZSL^{cls} compared to a HM of 7.65% (TCAF avg input) when using temporally averaged inputs. Moreover, TCAF uses an enhanced cross-modal attention to effectively gather multi-modal information. On the other hand, the attention mechanism of [Mer+22b] uses temporally averaged feature inputs, which leads to a HM of 6.82% on VGGSound-GZSL^{cls} ([Mer+22b]). Additionally, TCAF uses a single output branch and a classification token to aggregate the multi-modal information. In contrast, [Mer+22b] uses two branches and no classification token which leads to a HM of 6.27% (w/o class. token) on VGGSound-GZSL^{cls}. Finally, our training objective avoids triplet losses, i.e. there is no overhead to train with positive and negative pairs. Using triplet losses similar to those used in [Mer+22b] leads to a lower performance (TCAF + $l_{triplet}$) than TCAF. The same trend can be observed for the other datasets, proving that our architectural choices are more suitable for the audio-visual (G)ZSL task.

Model	VGGSound-GZSL^{cls}				UCF-GZSL^{cls}				ActivityNet-GZSL^{cls}			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
[Mer+22b]	12.63	6.19	8.31	6.91	63.15	30.72	41.34	37.72	16.77	7.04	9.92	7.58
TCAF +att from [Mer+22b]	10.08	5.16	6.82	5.41	39.47	28.85	33.33	29.79	5.58	2.37	3.33	2.43
TCAF avg input	11.69	5.69	7.65	6.16	12.00	20.46	15.13	20.59	16.43	3.26	5.44	3.42
w/o class. token	**18.40**	3.78	6.27	4.25	31.70	32.57	32.13	33.26	11.87	3.80	5.75	3.90
TCAF +$l_{triplet}$	14.51	4.78	7.19	5.06	**71.61**	35.91	47.83	40.00	18.74	6.58	9.74	6.63
TCAF	12.63	**6.72**	**8.77**	**7.41**	67.14	**40.83**	**50.78**	**44.64**	**30.12**	**7.65**	**12.20**	**7.96**

Table A.3: Transforming TCAF into [Mer+22b]

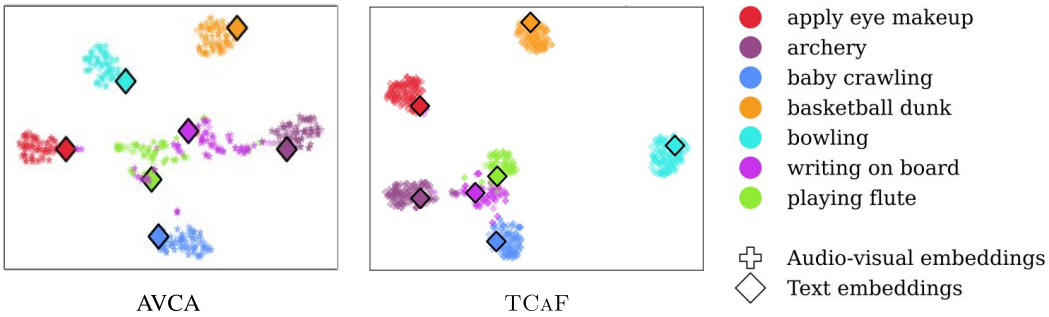

Figure A.2: t-SNE visualisations for five seen (*apply eye makeup, archery, baby crawling, basketball dunk, bowling*) and two unseen (*playing flute, writing on board*) test classes from the UCF-GZSL dataset, showing the difference between TCAF and [Mer+22b]. Textual class label embeddings are visualised with a square.

A.3 t-SNE Comparison Between TCAF and [Mer+22b]

We show t-SNE visualisations that highlight the difference between TCAF and [Mer+22b] in Figure A.2. It can be observed that in the case of [Mer+22b], the classes overlap more than in the case of TCAF. In particular, this can be observed for the unseen classes. Moreover, for [Mer+22b], the clusters are less concentrated than for TCAF.

A.4 Computational Complexity

The computational complexity increases with the length of the temporal sequence. Using the average duration of the data in UCF-GZSL^{cls} and a single forward pass for a batch of 256 samples, TCAF requires 51.8 GFLOPS vs 174.1 for [Jae+21] and 4.4 for [Mer+22b]. The Perceiver [Jae+21] uses a transformer architecture along with the temporal dimension, while [Mer+22b] does not use the temporal dimension. Thus, it can be observed that TCAF is more resource-efficient than the most similar baseline. TCAF was trained on a single NVIDIA 2080Ti GPU.

A.5 Additional Quantitative Results with SeLaVi [Asa+20] Features

In this section, we present additional results that show the performance of our TCAF with SeLaVi [Asa+20] input features from [Mer+22b]. On VGGSound-GZSL, TCAF obtains a HM of 7.33% compared to 6.31% for AVCA and a ZSL of 6.06% for TCAF vs. 6.00% for AVCA. Furthermore, on UCF-GZSL, TCAF significantly outperforms AVCA, with a HM of 31.72% compared to 27.15% and a ZSL performance of 24.81% compared to 20.01% for AVCA. On the other hand, on ActivityNet-GZSL, AVCA outperforms TCAF with a HM of 12.13% vs 10.71% for TCAF and a ZSL of 9.13% for AVCA vs 7.91% for TCAF. However, on ActivityNet-GZSL, TCAF outperforms Perceiver which is the most similar baseline to TCAF and which also uses temporal features.

Model	VGGSound-GZSL				UCF-GZSL				ActivityNet-GZSL			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
Att. Fusion	6.12	2.26	3.30	2.38	35.47	11.26	17.10	12.54	6.49	2.04	3.11	2.63
Perceiver	7.92	2.72	4.05	2.93	34.10	18.18	23.72	18.77	7.22	5.16	6.02	5.37
CJME	8.69	4.78	6.17	5.16	26.04	8.21	12.48	8.29	5.55	4.75	5.12	5.84
AVGZSLNet	**18.05**	3.48	5.83	5.28	52.52	10.90	18.05	13.65	8.93	5.04	6.44	5.40
AVCA	14.90	4.00	6.31	6.00	51.53	18.43	27.15	20.01	**24.86**	**8.02**	**12.13**	**9.13**
TCAF (ours)	9.64	**5.91**	**7.33**	**6.06**	**58.60**	**21.74**	**31.72**	**24.81**	18.70	7.50	10.71	7.91

Table A.4: Audio-visual (G)ZSL results when using SeLaVi [Asa+20] audio and visual features as inputs on the ActivityNet-GZSL, VGGSound-GZSL, and UCF-GZSL datasets.

B

Supplementary Material: Text-to-Feature Diffusion for Audio-Visual Few-Shot Learning

In Appendix B.1, we describe the procedure used to extract the audio and visual features that are used as inputs to our AV-Diff framework. In Appendix B.2, we provide additional experimental results for (G)FSL with 20 shots, along with reporting the GFSL performance on base and novel classes across all shots and datasets. Finally, we provide additional ablations on the hybrid attention and diffusion model.

B.1 Feature Extraction

We train AV-Diff on already pre-extracted temporal features for the audio and visual modalities. We used C3D [Tra+15] which was pretrained on Sports1M [Kar+14] and VGGish [Her+17] pre-trained on Youtube-8M [Abu+16] to extract audio and visual features respectively. Each audio feature is represented by a 128-dimensional vector corresponding to one second of audio data. To extract the visual features, we first resampled the videos to 25fps and then extracted a 4096-dimensional vector for 16 consecutive video frames.

B.2 Additional Experimental Results

We present (G)FSL results for 20 shots on the UCF-FSL, VGGSound-FSL and ActivityNet-FSL datasets in Appendix B.2.1. In Appendix B.2.2, we discuss the 1-,5-,10- and 20-shot (G)FSL performance on base and novel classes across all three datasets (which complements Section 3.5.2 of the main paper). Finally, Appendix B.2.3 shows additional ablations on the hybrid attention and diffusion model.

B.2.1 (G)FSL in the 20-Shot Setting

In Table B.1 (bottom), we provide additional (G)FSL results for the 20-shot setting with AV-DIFF and related methods. Similar to our observations in the main paper with 1, 5, and 10 shots, AV-DIFF achieves state-of-the-art performance for 20 shots, outperforming all related methods in the FSL and GFSL (HM) settings.

Similar to the conclusions for ActivityNet-FSL in the main paper, it can be observed that the ranking of baselines changes dramatically on ActivityNet-FSL, while AV-DIFF still remains the best, showing that our model is also more robust on 20 shots.

The HM and FSL performances on 20 shots for AV-DIFF and for the related methods are higher compared to the lower shots. The increase in performance for AV-DIFF from 10 to 20 shots is similar to the one from 5 to 10 shots. However, the most significant boost in performance happens between the 1-shot and 5-shot settings, showing that the gain in performance decreases as more training samples for novel classes are added. Similar trends can also be observed for the related methods.

B.2.2 Performance on Base and Novel Classes

In the main paper, we only presented the GFSL results in terms of the harmonic mean of the performance on the B (base) and N (novel) classes (Table 3.2 in the main paper). The harmonic mean is crucial as it evaluates how robust a system is, and it also provides higher scores to systems which are very balanced and which are less biased towards either B or N. In this section, we are going to analyse the performance of the components that are used to calculate the HM, namely the B and N performance, to have a better idea of the models' strengths and weaknesses. It can be seen in Table B.1 that in the majority of cases, AV-DIFF obtains state-of-the-art performance on B and N, but there are still some exceptions, as presented below.

In the 1-shot setting, it can be observed that MBT outperforms AV-DIFF on N in VGGSound-FSL and B in UCF-FSL, with scores of 21.34% and 79.89% compared to 21.25% and 77.94% for AV-DIFF. However, MBT is very biased towards one of the metrics. On VGGSound-FSL, the bias is towards N, and MBT obtains a very low score on B, only 11.21%, compared to 19.44% for AV-DIFF. The same applies on UCF-FSL, where MBT is very biased towards B. For B on VGGSound-FSL, AV-DIFF obtains a performance of 19.44% compared to 28.55% SLDG. While AV-DIFF scores similarly on both metrics in VGGSound-FSL, SLDG obtains a B score which is more than twice that of N, showing how unbalanced and biased SLDG is. An interesting observation that can be made in the 1-shot setting is that on VGGSound-FSL, AV-DIFF is not able to attain state-of-the-art performance in B or N, but it still performs overall much better than the systems that outperform AV-DIFF in these two metrics.

In the 5-shot setting, AV-DIFF is outperformed on B in both VGGSound-FSL and UCF-FSL by the Perceiver, with scores of 31.46% and 83.56% compared to 30.88% and 74.11% for AV-DIFF. Moreover, on VGGSound-FSL, AV-DIFF is also outperformed on N

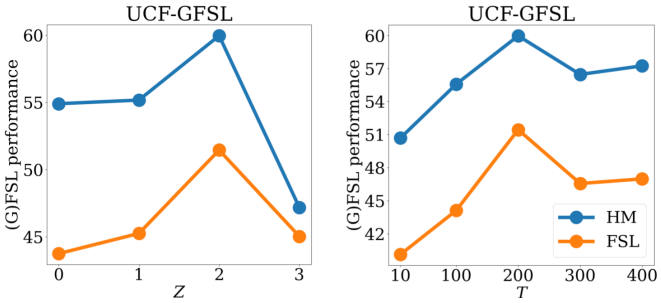

Figure B.1: (G)FSL performance (5-shot) for different numbers of self- (Z) and full attention layers (*left*), and different amounts of noise addition time steps T on UCF-FSL (*right*).

by MBT with scores of 31.79% vs 31.50% for AV-Diff. However, both MBT and Perceiver have a bigger bias towards one of the metrics, leading to a lower HM on VGGSound-FSL. On UCF-FSL, it can be clearly observed that Perceiver is biased towards B, obtaining a score which is more than twice that of N. For AV-Diff this is not the case, as scores for both B and N are much more balanced.

The same observations can be made in the 10- and 20-shot settings where sometimes AV-Diff is outperformed in one of the B or N, but still achieves a higher HM overall. While most of the baselines that outperform AV-Diff in one of the metrics are usually very biased towards that metric, this is not always the case. For example, in the 20-shot setting on UCF-FSL, Att. Fusion slightly outperforms AV-Diff on N with a score of 61.02% compared to 59.94% for AV-Diff. However, on B, AV-Diff significantly outperforms Att. Fusion with a score of 86.51% compared to 79.39% for Att. Fusion. While in this case Att. Fusion is very well balanced, it is still worse overall than AV-Diff, as it only slightly outperforms AV-Diff in N but it is significantly outperformed in B.

Interestingly, for different methods, the N score is sometimes higher than B. This is likely due to the use of calibrated stacking [Cha+16]. A similar behaviour has been observed by several other works, such as [Mer+22a; Mer+22b; Min+20]

Overall, AV-Diff is not necessarily the best in both B and N every single time. However, across all shots and datasets, AV-Diff achieves state-of-the-art GFSL performance in terms of the HM. This shows that AV-Diff is the most balanced and robust among all the methods, as it can consistently score very high on both B and N.

B.2.3 Ablation on Hybrid Attention and Diffusion.

In Figure B.1 (left), we analyse the impact of the number of self-attention layers Z and full-attention layers used. For values of $Z < 2$ the performance increases consistently and reaches a peak performance at $Z = 2$ for both metrics on UCF-FSL. It appears that changing the attention in late layers of the network is beneficial. Finally, we ablate over the timesteps T for adding noise to the original feature in the diffusion model in Figure B.1 (right). The (G)FSL performance maximizes for $T = 200$ on UCF-FSL which corresponds to the number of timesteps used in AV-Diff.

1-shot	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
Att. F. [FK20]	15.16	15.77	15.46	16.37	38.91	35.98	37.39	36.88	3.48	5.78	4.35	5.82
Perc. [Jae+21]	18.46	17.51	17.97	18.51	74.57	31.33	44.12	33.73	30.32	12.14	17.34	12.53
MBT [Nag+21]	11.21	21.34	14.70	21.96	79.89	26.37	39.65	27.99	17.07	12.24	14.26	12.63
TCAF [Mer+22a]	20.93	18.34	19.54	20.01	66.18	33.64	44.61	35.90	23.85	12.62	16.50	13.01
Proto [Kum+19]	8.85	13.65	10.74	14.08	60.12	27.72	37.95	28.08	2.02	4.40	2.77	4.40
SLDG [BLH20]	28.55	11.94	16.83	17.57	73.15	27.45	39.92	28.91	23.22	9.58	13.57	10.30
TSL [Xia+21]	17.09	20.72	18.73	22.44	68.18	33.04	44.51	35.17	8.96	10.18	9.53	10.77
HiP [Car+22]	23.39	16.39	19.27	18.64	16.20	33.26	21.79	34.88	25.02	9.53	13.80	10.31
Zorro [Rec+23]	17.49	20.51	18.88	21.79	67.85	32.94	44.35	34.52	19.67	11.55	14.56	11.94
AVCA [Mer+22b]	4.53	10.28	6.29	10.29	82.86	29.59	43.61	31.24	14.15	11.73	12.83	12.22
AV-DIFF	19.44	21.25	**20.31**	**22.95**	77.94	38.45	**51.50**	**39.89**	32.77	12.86	**18.47**	**13.80**

5-shot	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
Att. F. [FK20]	28.64	27.82	28.22	31.57	63.27	43.69	51.68	47.18	5.00	8.05	6.17	8.13
Perc. [Jae+21]	31.46	28.52	29.92	33.58	83.56	34.27	48.60	40.47	35.66	20.15	25.75	21.50
MBT [Nag+21]	23.86	31.79	27.26	34.95	80.61	32.72	46.55	34.53	25.36	21.48	23.26	22.38
TCAF [Mer+22a]	24.34	28.11	26.09	32.22	73.76	33.73	46.29	37.39	24.45	21.35	22.79	21.81
Proto [Kum+19]	25.27	25.08	25.17	28.87	63.69	31.79	42.42	33.63	1.61	7.81	2.67	7.81
SLDG [BLH20]	29.74	15.98	20.79	25.17	65.44	25.28	36.47	28.56	29.40	17.95	22.29	19.16
TSL [Xia+21]	15.02	27.75	19.49	29.50	68.80	40.62	51.08	42.42	9.93	12.27	10.97	12.77
HiP [Car+22]	30.01	24.18	26.82	30.67	33.65	39.74	36.44	42.23	21.98	15.39	18.10	16.25
Zorro [Rec+23]	29.06	30.07	29.56	35.17	69.13	41.49	51.86	42.59	25.72	21.03	23.14	21.94
AVCA [Mer+22b]	13.24	20.15	15.98	20.50	84.80	34.64	49.19	36.70	19.18	21.09	20.09	21.65
AV-DIFF	30.88	31.50	**31.19**	**36.56**	74.11	50.35	**59.96**	**51.45**	35.84	21.61	**26.96**	**23.00**

10-shot	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
Att. F. [FK20]	26.87	35.89	30.73	39.02	73.53	47.77	57.91	52.19	12.58	9.27	10.67	10.78
Perc. [Jae+21]	32.64	34.73	33.65	40.73	71.88	44.97	55.33	47.86	37.06	25.03	29.88	26.46
MBT [Nag+21]	26.76	34.43	30.12	38.93	84.07	35.62	50.04	39.73	29.06	24.98	26.86	26.03
TCAF [Mer+22a]	26.62	31.73	28.95	36.43	84.28	39.93	54.19	47.61	27.86	22.32	24.78	23.33
Proto [Kum+19]	30.48	29.26	29.85	34.80	70.28	40.03	51.01	40.68	2.63	8.81	4.05	8.81
SLDG [BLH20]	28.32	20.99	24.11	29.48	49.35	26.29	34.31	26.96	34.69	23.20	27.81	25.35
TSL [Xia+21]	17.96	28.15	21.93	31.29	74.31	51.63	60.93	55.63	9.31	11.76	10.39	12.18
HiP [Car+22]	28.43	30.12	29.25	35.13	75.54	38.14	50.69	43.29	24.32	16.10	19.37	17.06
Zorro [Rec+23]	28.48	36.68	32.06	40.66	82.88	45.67	58.89	49.06	30.11	25.05	27.35	26.33
AVCA [Mer+22b]	13.39	27.83	18.08	28.27	71.96	38.93	50.53	39.17	26.36	25.68	26.02	26.76
AV-DIFF	32.15	36.05	**33.99**	**41.39**	84.62	51.69	**64.18**	**57.39**	37.91	26.02	**30.86**	**27.81**

20-shot	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
Att. F. [FK20]	31.43	37.88	34.35	44.08	79.39	61.02	69.00	63.20	15.51	11.41	13.15	13.22
Perc. [Jae+21]	33.11	37.66	35.24	43.77	77.81	48.29	59.59	52.66	32.30	31.06	31.67	32.21
MBT [Nag+21]	28.41	37.95	32.49	43.19	81.73	42.35	55.80	44.58	36.21	28.60	31.96	30.76
TCAF [Mer+22a]	32.48	29.41	30.87	38.89	75.71	47.38	58.29	51.99	35.87	27.61	31.20	29.88
Proto [Kum+19]	31.44	32.66	32.04	38.42	61.07	49.32	54.57	50.48	25.05	8.17	12.32	14.65
SLDG [BLH20]	33.20	19.53	24.59	33.30	81.08	39.52	53.14	43.95	32.60	30.80	31.68	32.44
TSL [Xia+21]	18.21	29.32	22.47	32.07	76.82	49.44	60.16	52.02	9.68	15.01	11.77	15.78
HiP [Car+22]	32.03	29.83	30.89	38.46	71.59	43.43	54.06	48.07	33.78	17.59	23.13	20.67
Zorro [Rec+23]	29.84	39.46	33.98	43.63	87.82	48.46	62.45	57.10	34.15	28.55	31.10	30.31
AVCA [Mer+22b]	15.30	32.20	20.75	32.64	60.00	44.93	51.39	44.93	24.47	29.88	26.91	30.76
AV-DIFF	33.17	39.46	**36.04**	**44.79**	86.51	59.94	**70.81**	**65.72**	39.25	31.06	**34.68**	**32.89**

Table B.1: **Novel (N) and base (B) performance for audio-visual (G)FSL**: 1-shot, 5-shot, 10-shot, and 20-shot performance of AV-DIFF and compared methods on the VGGSound-FSL, UCF-FSL and ActivityNet-FSL datasets. The harmonic mean (HM) of the mean class accuracies for base and novel classes are reported for GFSL. The FSL performance considers only the test subset of novel classes.

C

Supplementary Material: Video-Adverb Retrieval with Compositional Adverb-Action Embeddings

C.1 Dataset Splits for Unseen Adverb-Action Compositions

In this section, we provide further details about our proposed dataset splits for unseen adverb-action compositions based on the ActivityNet Adverbs [Cab+15; Dou+20] and MSR-VTT Adverbs [Dou+20; Xu+16] datasets. In Table C.1, we include information about the number of unlabelled samples (i.e. videos) and the number of unlabelled pairs (i.e. adverb-action compositions) in the dataset splits. The unlabelled samples are not used by ReGaDa, but we designed the splits so that we can fairly evaluate previous work [DS22] that uses unlabelled samples for training. The number of unlabelled samples and unlabelled pairs usually ranges from 30% to 50% of the total number of training samples and training pairs. This is significant, as methods like [DS22] use more training data than ReGaDa while performing significantly worse as observed in Table 4.6 in the main paper. We refer to the ActivityNet Adverbs and MSR-VTT Adverbs datasets as ActivityNet and MSR-VTT respectively.

In addition to the ActivityNet Adverbs and MSR-VTT Adverbs datasets, we use the VATEX Adverbs dataset [Dou+20; Wan+19a], and in particular the corresponding splits for unseen adverb-action compositions introduced in [DS22]. However, we use the same pre-extracted features as the current state-of-the-art work [Mol+23]. As some of the videos used in the split in [DS22] are not available anymore, it is not possible to extract S3D features for those. Hence, this resulted in fewer samples in the dataset, the number of training samples being reduced from 6921 to 6603, unlabelled samples from 3469 to 3317, and test samples from 3457 to 3293. In the following, we refer to the VATEX Adverbs dataset as VATEX.

Dataset	# train samples	# unlabelled samples	# test samples	# pairs train	# pairs unlabelled	# pairs test
VATEX	6603	3317	3293	319	168	316
MSR-VTT	987	306	454	225	114	225
ActivityNet	1490	634	848	635	537	543

Table C.1: Statistics of our dataset splits for the retrieval of unseen adverb-action compositions on the MSR-VTT Adverbs and ActivityNet Adverbs datasets. Statistics are also provided for the VATEX Adverbs dataset for features from [Mol+23].

C.2 Exploring the Use of Different Word Embeddings for Unseen Adverb-Action Compositions

Our ReGAdA framework composes adverb and action text embeddings in a shared embedding space. Specifically, we used a text model that was jointly trained with the S3D video model. In this section, we show results for different choices of word embeddings. Existing and widely-adopted word embeddings like GloVe [PSM14], word2vec [Mik+13], and fastText [Boj+17] rely on unsupervised learning techniques to generate vector representations of words based on their co-occurrence statistics in a large corpus of text. Specifically, word2vec and GloVe focus on co-occurrences of words, whereas fastText uses co-occurrences of n-gram characters, which can be useful when dealing with rare words.

Prior works on video-adverb retrieval leveraged GloVe embeddings of class labels [Dou+20; DS22], while approaches in zero-shot learning commonly use word2vec or fastText embeddings as side information [Man+21; Mer+22a; Mer+22b; Nae+21; Xia+16].

However, recent advances in language modelling have shown impressive progress on a variety of natural language processing tasks. For instance, large language models incorporate contextual information at the sentence level and beyond, which could result in more informative and accurate embeddings. To investigate their usefulness for our retrieval task, we extract word embeddings with GPT-3 [Bro+20] using the OpenAI API for the `text-embedding-ada-002` model. While word2vec, fastText, and GloVe provide 300-dimensional embeddings, GPT-3 embeddings have a much larger dimension of 1536. All text embeddings are projected to 400-dimensional vectors before being input into the text encoder. For CLIP [Rad+21], we extract visual CLIP features for each second of the video and CLIP text embeddings from the action-adverb labels (e.g. *cut slowly*). We then use the cosine similarity between temporally-averaged frame features and text embeddings for retrieval.

Table C.2 shows that the choice of the text embedding results in significant performance changes, measured by the binary antonym classification accuracy. ReGAdA uses text embeddings jointly trained with the S3D video model like the other baselines (referred to as S3D embeddings in the following), and it is able to outperform all the baselines, as shown in the main paper. However, from Table C.2 it can be observed that ReGAdA with S3D embeddings is outperformed by ReGAdA with GPT-3 embeddings on VATEX,

leading to a performance of 63.3 compared to 61.7 for S3D embeddings. GPT-3 embeddings contain more contextual and fine-grained semantic information but suffer from a significant reduction in dimensions in the projection. We find that higher-dimensional text embeddings perform worse when training data is scarce (e.g. 53.5/60.3 for GPT-3 vs. 58.4/61.0 for S3D on ActivityNet/MSR-VTT), likely caused by a lack of training data to learn the down-projection. Overall, word2vec, fastText, and GloVe embeddings yield slightly worse results than S3D embeddings across datasets.

C.3 Training Without Antonyms

In Table C.3, we present the video-to-adverb and adverb-to-video retrieval performance when training without antonyms. This task was introduced in [Mol+23]. For the results in the main paper, ReGAdA is trained with antonyms as negative examples in its triplet loss. As it might not always be feasible to require adverb-action samples that are additionally annotated with an adverb-antonym, this scenario inspects the generalisation capabilities of ReGAdA to dataset settings with fewer constraints.

When training without adverb-antonyms, ReGAdA randomly samples an adverb as a negative sample which is not identical to the positive adverb sample. As there is no access to information about the adverb-antonym during evaluation, the Acc-A metric cannot be used in this context.

In Table C.3 we can observe that ReGAdA outperforms all prior methods for this task across all datasets and metrics. For example, on VATEX ReGAdA obtains a mAP W score of 0.292 compared to 0.283 for AC_{CLS}. Moreover, ReGAdA obtains a mAP M score of 0.136 which significantly outperforms AC_{CLS} with a score of 0.108.

Model	VATEX	ActivityNet	MSR-VTT
CLIP [Rad+21]	54.5	55.1	57.0
Act. Mod. [DS22]	53.8	57.0	56.0
AC_{CLS} [Mol+23]	54.3	55.1	53.7
AC_{REG} [Mol+23]	54.9	53.9	59.0
ReGAdA	61.7	**58.4**	**61.0**
ReGAdA w2v	60.5	53.1	60.0
ReGAdA fastText	60.8	53.5	57.3
ReGAdA GloVe	58.0	54.0	57.7
ReGAdA GPT-3	**63.3**	53.5	60.3

Table C.2: Effect of using different types of word embeddings in our ReGAdA framework on the performance for retrieving unseen action-adverb compositions on the VATEX, ActivityNet and MSR-VTT benchmarks. [DS22] uses pseudo-labelling.

	HowTo100M [Dou+20]		Adverbs in Recipes [Mol+23]		ActivityNet [DS22]		MSR-VTT [DS22]		VATEX [DS22]	
	mAP W	mAP M	mAP W	mAP M	mAP W	mAP M	mAP W	mAP M	mAP W	mAP M
Priors	0.446	0.354	0.491	0.263	0.217	0.159	0.308	0.152	0.216	0.086
S3D pre-trained	0.339	0.238	0.389	0.173	0.118	0.071	0.194	0.075	0.122	0.038
TIRG [Vo+19]	0.441	0.476	0.485	0.228	0.186	0.111	0.297	0.113	0.195	0.065
Act. Mod. [Dou+20]	0.408	0.352	0.508	0.249	0.187	0.127	0.233	0.134	0.144	0.060
AC_{CLS} † [Mol+23]	0.562	0.420	0.606	0.289	0.130	0.096	0.305	0.131	0.283	0.108
AC_{REG} † [Mol+23]	0.573	0.481	0.667	0.319	0.143	0.093	0.287	0.121	0.282	0.100
ReGAdA	**0.580**	**0.536**	**0.668**	**0.466**	**0.282**	**0.211**	**0.401**	**0.252**	**0.292**	**0.136**

Table C.3: Results *without* antonyms during training for adverb-to-video retrieval (mAP W/M). Higher is better for all metrics. † refers to updated results provided by the authors of [Mol+23].

	ActivityNet			MSR-VTT			VATEX		
	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A
S3D pre-tr.	0.118	0.070	0.560	0.194	0.075	0.603	0.122	0.038	0.586
CLIP [Rad+21]	0.120	0.067	0.611	0.206	0.084	0.677	0.129	0.039	0.644
ReGAdA$_{CLIP}$	0.201	0.151	**0.781**	0.352	0.142	0.784	0.247	0.098	**0.837**
ReGAdA	**0.239**	**0.175**	0.771	**0.378**	**0.228**	**0.786**	**0.290**	**0.113**	0.817

Table C.4: Comparing ReGAdA with CLIP as a baseline, and when replacing ReGAdA's S3D video/text embeddings with CLIP embeddings (ReGAdA$_{CLIP}$).

C.4 Comparing ReGAdA with CLIP

In this section, we present additional video-adverb retrieval results with CLIP [Rad+21] in addition to the retrieval results for unseen compositions (see Table C.2).

Similar to the experiment on unseen compositions (see Appendix C.2), we use the cosine similarity between temporally-averaged CLIP frame features and text embeddings for the retrieval with CLIP. Additionally, we examine the impact of replacing the S3D video/text embeddings of ReGAdA with CLIP embeddings (ReGAdA$_{CLIP}$).

In Table C.4, we can observe that CLIP performs marginally better than the S3D pre-trained baseline. Using CLIP features in ReGAdA improves adverb retrieval (Acc-A) slightly on ActivityNet and VATEX. However, ReGAdA$_{CLIP}$ is worse than ReGAdA for video retrieval, likely caused by inferior visual features when extracting those only from a few video frames.

C.5 Seed Experiments

In Table C.5, we provide experimental results that test the robustness of our model with regard to the seeds used, as done in [Mol+23]. To compute these numbers, we use four seeds and compute the mean and the standard deviation over these runs. It can be observed that ReGAdA achieves a higher mean than the other baselines. Furthermore, the standard deviation with our model is relatively low.

	Adverbs in Recipes [Mol+23]		
	mAP W	mAP M	Acc-A
Act. Mod.	0.394 ± 0.023	0.140 ± 0.026	0.843 ± 0.013
MLP+Act. Mod.	0.407 ± 0.044	0.151 ± 0.033	0.842 ± 0.012
AC_{CLS} †	0.605 ± 0.001	0.287 ± 0.001	0.841 ± 0.000
AC_{REG} †	0.611 ± 0.002	0.239 ± 0.007	0.845 ± 0.001
REGADA	**0.699 ± 0.004**	**0.419 ± 0.012**	**0.876 ± 0.001**

Table C.5: Performance of our REGADA framework on the Adverbs in Recipes dataset when using multiple random seeds. † refers to updated results provided by the authors of [Mol+23].

D

Supplementary Material: EgoCVR: An Egocentric Benchmark for Fine-Grained Composed Video Retrieval

In this appendix, we report results when applying re-ranking to other methods in Appendix D.1, show failure cases and future directions of TFR-CVR in Appendix D.2, and report results with TFR-CVR on WebVid-CoVR-Test [Ven+24] in Appendix D.3. Furthermore, we provide more details on EgoCVR's diversity (Appendix D.4.1), present the instruction prompts given to the LLM models to create the text modification (Appendix D.4.2), to create the target captions for TF-CVR (Appendix D.4.3) and to perform the temporal event detection analysis on the CVR benchmarks (Appendix D.4.4), as well as provide additional qualitative examples (Appendix D.5).

D.1 Re-Ranking Applied to Other Methods

We show the results for applying re-ranking using the same LanguageBind encoder for all the methods below. While the re-ranking improves $\text{BLIP}_{\text{CoVR}}$, the overall performance and improvement are much larger for TFR-CVR.

Method	w/o re-ranking			w/ re-ranking			Δ R@{1,5,10}
	R@1	R@5	R@10	R@1	R@5	R@10	
CLIP [Rad+21]	7.5	33.6	**55.6**	7.5	33.5	54.4	0.4 ↓
BLIP [Li+22]	8.7	32.9	52.8	8.7	33.8	54.2	0.8 ↑
$\text{BLIP}_{\text{CoVR}}$ [Ven+24]	5.4	15.2	24.3	10.4	31.9	52.2	16.5 ↑
TFR-CVR	4.4	12.9	18.3	14.1	39.5	54.4	**24.1** ↑

Table D.1: Results on EgoCVR in terms of R@1, R@5 and R@10 on the global setting with and without applying re-ranking. We also report the mean recall change when applying the re-ranking.

D.2 Failure Cases and Future Directions

Due to the modular approach of TFR-CVR, we can break down the failure cases. As shown in Table 5.4 in the main paper, employing the ground-truth (GT) captions achieves only an R@1 of 51.7% with a gallery of 7 candidates (Local). Therefore, the biggest source of improvement on EgoCVR would be from applying stronger text-to-video retrieval models. We can also trace back errors to erroneous video captions. In Figure D.1, we show an example of the wrong retrieval caused by the text-to-video retrieval method at the bottom. In addition, we highlight an error caused by the video captioning method on the top. Here, the *"pot"* was mistaken for a *"bowl"* during the captioning, leading to the retrieval of a video that mainly depicts a bowl.

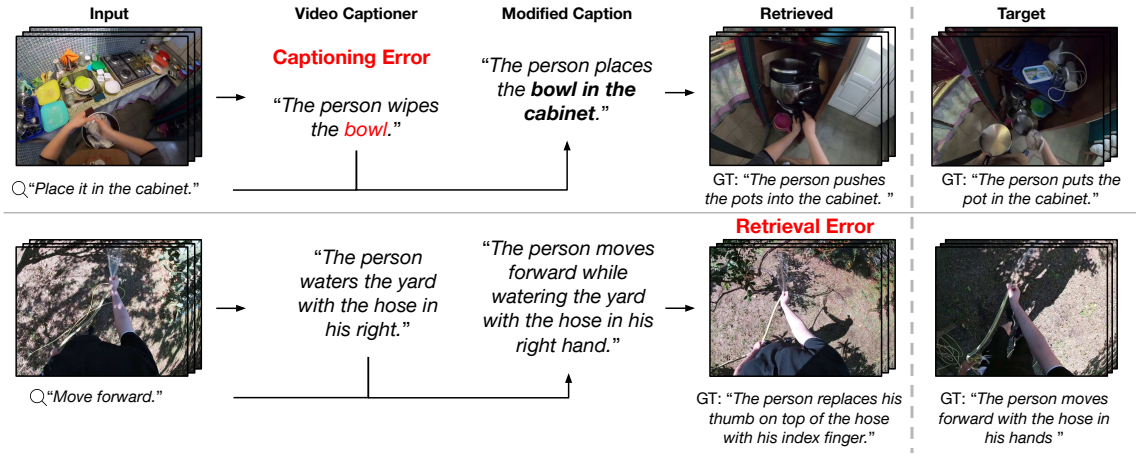

Figure D.1: Qualitative depiction of failure cases of TFR-CVR. The modular approach of TFR-CVR allows us to trace back failure cases mostly to two main sources: video *captioning errors* (top) and text-to-video *retrieval errors* (bottom).

D.3 Results on WebVid-CoVR Benchmark

We additionally show results of TFR-CVR on the WebVid-CoVR-Test [Ven+24] benchmark in Table D.2. For TFR-CVR, we employ LanuguageBind [Zhu+24] as the visual and textual encoder, since it was trained on different types of videos, unlike EgoVLPv2 [Pra+23] which is restricted to egocentric videos. We observe that TFR-CVR is able to achieve the best results among all methods that do not explicitly train on the WebVid-CoVR training set. The performance (R@1 of 51.7%) is also quite competitive with the state-of-the-art result obtained by $\text{BLIP}_{\text{CoVR}}$ [Ven+24] (R@1 of 53.1%). We also continue to notice a constant benefit from applying the re-ranking strategy. After applying the visual-only filtering of LanguageBind (which in itself is a weak model scoring an R@1 value of 43.2%), the performance of the text-based TF-CVR method improves from R@1 of 48.4% to 51.7%.

Method	Zero-Shot	Fusion Strategy	R@1	R@5	R@10
Random	✓	-	0.08	0.23	0.35
CLIP [Rad+21]	✓	Avg	44.4	69.1	77.7
BLIP [Li+22]	✓	Avg	45.5	70.5	79.5
LanguageBind [Zhu+24] (stage 1; visual)	✓	-	43.2	66.3	75.2
TF-CVR	✓	Captioning	48.4	73.7	**81.9**
TFR-CVR (LanguageBind → TF-CVR) (n_c = 20)	✓	Captioning	**51.7**	**75.3**	80.7
$\text{BLIP}_{\text{CoVR}}$ [Ven+24]	✗	Cross-Attention	53.1	79.9	86.9

Table D.2: Results on WebVid-CoVR-Test [Ven+24] in terms of Recall@1, Recall@5 and Recall@5. We report results with models evaluated in the Zero-Shot setting [Li+22; Rad+21; Zhu+24] (including our TF-CVR and TFR-CVR) and with models specifically trained for this data set [Ven+24].

D.4 Additional Details to EgoCVR

D.4.1 Dataset Diversity

EgoCVR entails 179 different actions across 47 distinct scenarios (top-6: *crafting, cooking, cleaning, construction, carpenter, car mechanic*). In Figure D.2, we show (i) a word cloud illustrating the various actions included in EgoCVR, and (ii) exemplary of diversity in action environments ("*pick*" appears in 30 and "*turn*" in 9 scenarios). EgoCVR offers wide-ranging environments, objects and actions.

(a) Diversity in actions (b) Diverse environments

Figure D.2: Diversity of actions and environments in EgoCVR.

D.4.2 Creating Video Modification Instructions

In this section, we give more details on how we create video modification instructions. We automatically generate instructions from query and target clip narrations. More specifically, given the narrations of the source and the target video, the goal is to generate the instruction (modification) text. To achieve this, we use 15 in-context examples showing how this could be done effectively, focusing only on the modification and not the source and target captions itself.

Dataset Generation (Instruction) Prompt

I have 2 videos. Given a brief description of the source and the target video, write an instruction that describes the transformation from the source to the target. The caption you generate should only talk about the necessary modifications. Keep the instruction as short as possible, and focus always on the action. Mention objects only when absolutely necessary. You should not describe objects common to both descriptions, instead use pronouns. Describe only the transformation required. Use the examples below for reference.

Source Narration: #C C picks up the jug.
Target Narration: #C C cleans the jug.
Instruction: The person is cleaning.

Source Narration: #C C picks a spanner from the table with his right hand.
Target Narration: #C C picks a gasket from a table with his right hand.
Instruction: Gasket being picked up.

Source Narration: #C C picks up the wood from the shelf with his left hand
Target Narration: #C C detaches a wood from the wooden structure with his right hand
Instruction: Person uses the other hand and detaches.

Source Narration: #C C fixes the bolt on the motorbike with his right hand.
Target Narration: #C C holds the part of the motorbike with his left hand.
Instruction: Person holds it with the other hand.

Source Narration: #C c climbs up the steps.
Target Narration: #C c climbs down the steps
Instruction: The person climbs down.

Source Narration: #C C Wipes as paint brush with a paper towel.
Target Narration: #C C dips brush in water.
Instruction: Dip the object in water.

Source Narration: #C C pours the water in the shoe.
Target Narration: #C C rinses the shoe.
Instruction: Rinse it instead.

Source Narration: #C C puts electric shoe cleaner on the sink
Target Narration: #C C puts shoe in the sink
Instruction: Same action with a shoe.

Source Narration: #C C puts down penetrant oil
Target Narration: #C C sprays the oil
Instruction: Spray it.

Source Narration: #C C moves the scissors aside
Target Narration: #C C moves the coins aside
Instruction: Change it to coins.

Source Narration: #C C fixes the lawn mower basket
Target Narration: #C C holds the lawn mower basket
Instruction: Hold it instead.

Source Narration: #c c sits on the mat
Target Narration: #C C kneels on the carpet
Instruction: Kneels instead.

Source Narration: #C C drops the plate of food on the sink slap.

Target Narration: #C C picks a plate from the sink slap.
Instruction: Pick it up.

Source Narration: #C C holds the basket of flowers on the floor with her left hand.
Target Narration: #C C puts the white flowers on the tray with her right hand.
Instruction: Transfer it to the tray.

Source Narration: #C C scrapes carrot remains from the grater into the brown bowl
Target Narration: #C C picks carrot from the brown bowl with her right hand
Instruction: Pick it up from the bowl.

D.4.3 Creating Target Caption

We present the prompt utilized to obtain a valid target caption from the query video caption and the instruction text. Our goal is to take a video caption along with an instruction specifying some changes and then generate a valid target caption. Similar to the dataset generation process, this also uses a few in-context examples to improve the quality of the generated captions.

TF-CVR Prompt

I have a video. Given a brief description of the source video and a instruction that modifies it, write a description of the target video. Keep the modified description as short as possible, while being complete. Mention objects only when absolutely necessary. Use the examples below for reference.

Source Narration: #C C picks up the jug.
Instruction: The person is cleaning.
Target Narration: #C C cleans the jug.

Source Narration: #C C picks a spanner from the table with his right hand.
Instruction: Gasket being picked up.
Target Narration: #C C picks a gasket from a table with his right hand.

Source Narration: #C C picks up the wood from the shelf with his left hand.
Instruction: Person uses the other hand and detaches.
Target Narration: #C C detaches a wood from the wooden structure with his right hand.

Source Narration: #C C fixes the bolt on the motorbike with his right hand.
Instruction: Person holds it with the other hand.
Target Narration: #C C holds the part of the motorbike with his left hand.

Source Narration: #C c climbs up the steps.
Instruction: The person climbs down.
Target Narration: #C c climbs down the steps.

Source Narration: #C C Wipes as paint brush with a paper towel.
Instruction: Dip the object in water.
Target Narration: #C C dips brush in water.

Source Narration: #C C pours the water in the shoe.
Instruction: Rinse it instead.
Target Narration: #C C rinses the shoe.

Source Narration: #C C puts electric shoe cleaner on the sink.
Instruction: Same action with a shoe.
Target Narration: #C C puts shoe in the sink.

Source Narration: #C C puts down penetrant oil.
Instruction: Spray it.
Target Narration: #C C sprays the oil.

Source Narration: #C C moves the scissors aside.
Instruction: Change it to coins.
Target Narration: #C C moves the coins aside.

Source Narration: #C C fixes the lawn mower basket.
Instruction: Hold it instead.
Target Narration: #C C holds the lawn mower basket.

Source Narration: #C C sits on the mat.
Instruction: Kneels instead.
Target Narration: #C C kneels on the carpet.

Source Narration: #C C drops the plate of food on the sink slap.
Instruction: Pick it up.
Target Narration: #C C picks a plate from the sink slap.

Source Narration: #C C holds the basket of flowers on the floor with her left hand.
Instruction: Transfer it to the tray.
Target Narration: #C C puts the white flowers on the tray with her right hand.

Source Narration: #C C scrapes carrot remains from the grater into the brown bowl.
Instruction: Pick it up from the bowl.
Target Narration: #C C picks carrot from the brown bowl with her right hand.

D.4.4 Analysing Modification Instructions for Temporal vs. Object Events

We analyse video modification instructions for both WebVid-CoVR [Ven+24] and our EgoCVR benchmark to study the type of modifications existing in the data sets. To categorize modification instructions as temporal or object-centric, we employ GPT-4. In the prompt, apart from a description of the task itself, we also provide a few in-context examples shown below.

Temporal Event Detection Prompt

I have an instruction to modify a video. Looking at just the instruction, you should decide whether the instruction is focused on temporal events such as actions, or if it is just focused on objects. The answer you generate should only be "yes" or "no". Use the examples below for reference.

Instruction: have him fishing
Answer: yes

Instruction: turn it red on a watercolor stain.
Answer: no

Instruction: make the sax player into a drummer
Answer: no

Instruction: the girl is crying
Answer: yes

Instruction: change the meat to prawns
Answer: no

Instruction: remove the man.
Answer: no

Instruction: dip it into the paint.
Answer: yes

Instruction: cut the carrot instead.
Answer: no

Instruction: insert it into the roof.
Answer: yes

Instruction: pick it up instead.
Answer: yes

D.5 Additional Qualitative Examples

We illustrate a few more examples extracted from `EgoCVR` in Figure D.3. It is clearly visible that the data set contains a wide variety of activities in different environments, while the examples typically focus on action focused changes.

We also illustrate resulting order of the videos obtained after re-ranking performed by `TFR-CVR` in Figure D.4 and Figure D.5. We observe that the correct video is not in the first position in the first stage, however, after performing the second stage, the correct video is in the top position.

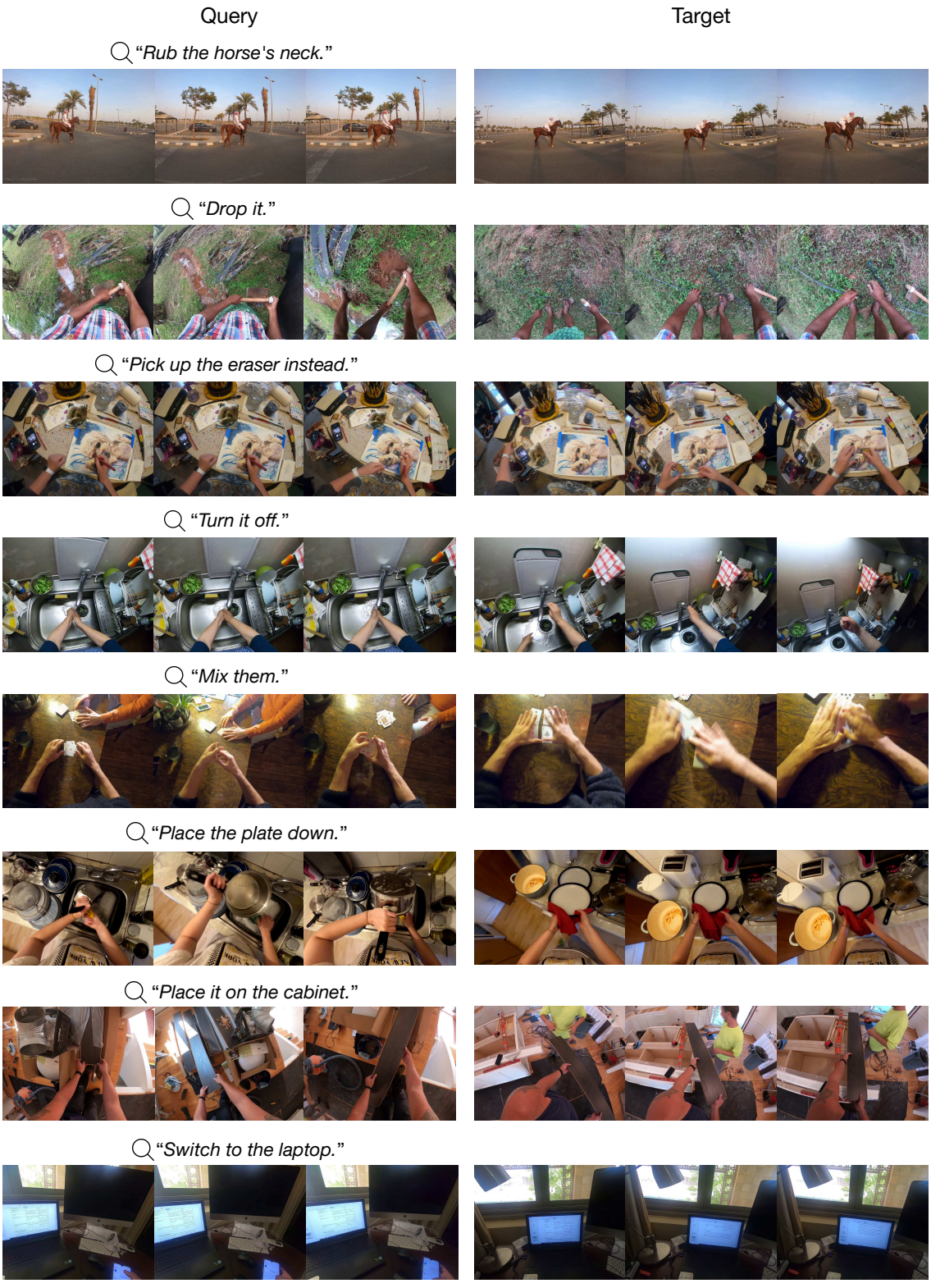

Figure D.3: Additional examples from our EgoCVR benchmark. We present the input video, the text modification and the target video. The dataset is diverse, covering various types of scenes and activities.

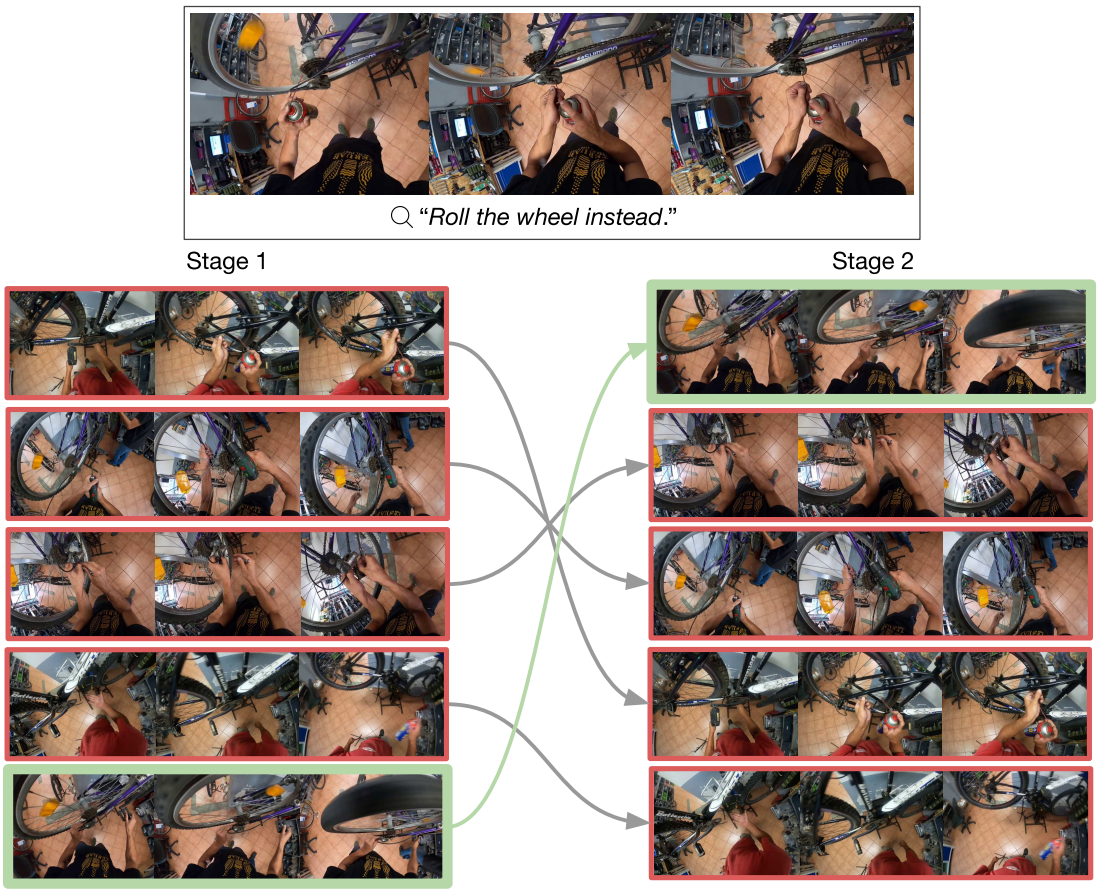

Figure D.4: Qualitative example for the first and the second stage ranking results of our `TFR-CVR` method for the query instruction "`Roll the wheel instead.`". The arrows indicate how the ranking was changed after re-ranking. The correct video is showcased in green.

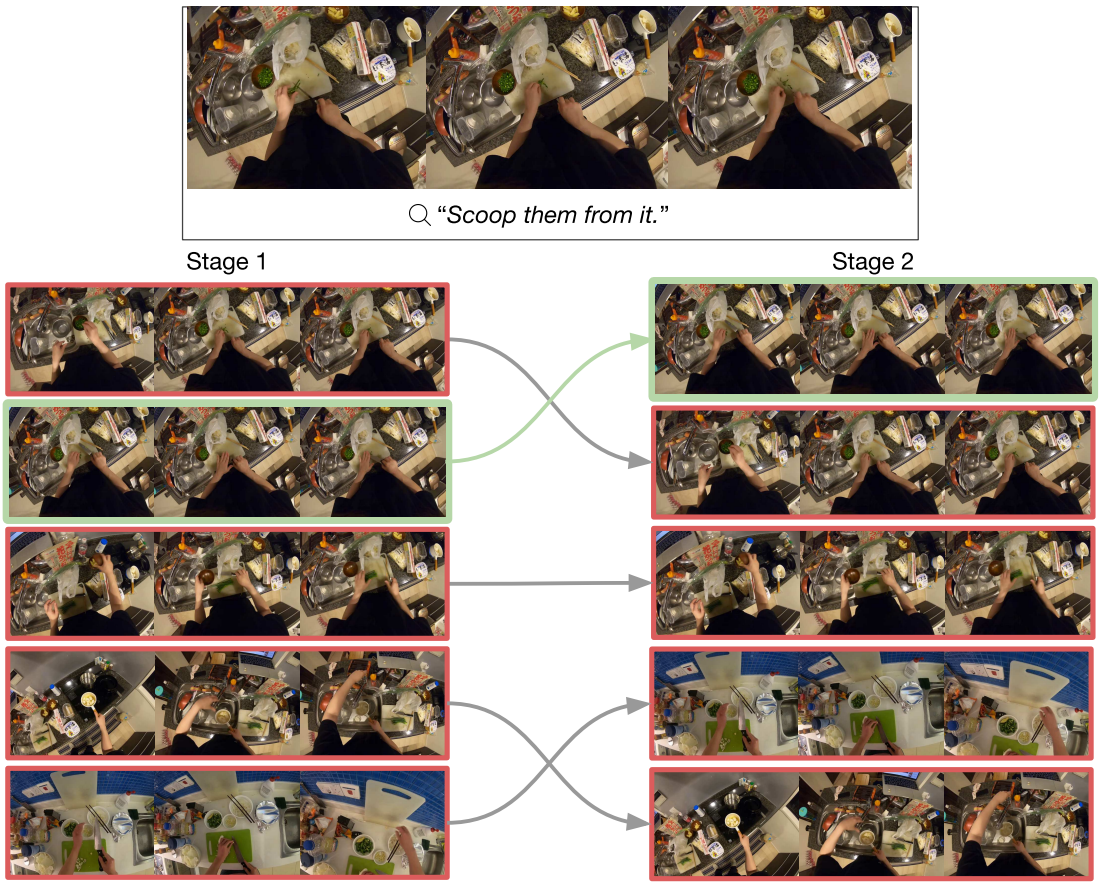

Figure D.5: Qualitative example for the first and the second stage ranking results of our `TFR-CVR` method for the query instruction "`Scoop them from it.`". The arrows indicate how the ranking was changed after re-ranking. The correct video is showcased in green.

E

Publications and Contributions

E.1 List of Publications

This section lists the publications that are part of this thesis. Shared first-author publications due to equal contributions of authors are indicated with an asterisk (*). An overview of the individual contributions is given in appendix E.2.

O.-B. Mercea*, **T. Hummel***, A. S. Koepke and Z. Akata. "Temporal and cross-modal attention for audio-visual zero-shot learning". In *European Conference on Computer Vision (ECCV)*. 2022.

O.-B. Mercea, **T. Hummel**, A. S. Koepke and Z. Akata. "Text-to-feature diffusion for audio-visual few-shot learning". In *DAGM German Conference on Pattern Recognition (DAGM GCPR)*. 2023.

T. Hummel, O.-B. Mercea, A. S. Koepke and Z. Akata. "Video-adverb retrieval with compositional adverb-action embeddings". In *British Machine Vision Conference (BMVC)*. ***Oral presentation***. 2023.

T. Hummel*, S. Karthik*, M.-I. Georgescu and Z. Akata. "EgoCVR: An Egocentric Benchmark for Fine-Grained Composed Video Retrieval". In *European Conference on Computer Vision (ECCV)*. 2024.

While completed during the PhD, the publication below is excluded from this thesis due to a significant divergence from the tasks investigated in this thesis.

S. Alaniz*, **T. Hummel*** and Z. Akata. "Semantic Image Synthesis with Semantically Coupled VQ-Model". In *ICLR Workshop on Deep Generative Models for Highly Structured Data*. 2022.

E.2 Contributions

This section outlines the individual contributions of the authors of the publications presented in this thesis and listed in appendix E.1.

Chapter 2: Temporal and Cross-Modal Attention for Audio-Visual Zero-Shot Learning.
This work was done in collaboration with Otniel-Bogdan Mercea, A. Sophia Koepke, and Zeynep Akata. Otniel-Bogdan Mercea and Thomas Hummel contributed equally as shared first-authors. Otniel-Bogdan Mercea contributed more to the task formulation and losses, and Thomas Hummel contributed more to the model and attention design. A. Sophia Koepke and Zeynep Akata provided oversight through weekly meetings to discuss the project and future milestones. All authors contributed to writing the paper.

Chapter 3: Text-to-Feature Diffusion for Audio-Visual Few-Shot Learning.
This work was done in collaboration with Otniel-Bogdan Mercea, A. Sophia Koepke, and Zeynep Akata. Otniel-Bogdan Mercea served as the first author and contributed by creating the benchmarks, implementing the baselines, developing the model framework, and running most of the experiments. Thomas Hummel was the second author and contributed ideas, implemented some of the ablations, and helped run some experiments. A. Sophia Koepke and Zeynep Akata provided oversight through weekly meetings to discuss the project and future milestones. All authors contributed to writing the paper.

Chapter 4: Video-Adverb Retrieval with Compositional Adverb-Action Embeddings.
This work was done with Otniel-Bogdan Mercea, A. Sophia Koepke, and Zeynep Akata. Thomas Hummel served as the first author of the paper. Thomas Hummel proposed the project's original idea, developed the state-of-the-art system, ran most of the experiments, and implemented most of the ablations. Otniel-Bogdan Mercea contributed to implementing some of the ablations, helped run some of the experiments, and created the zero-shot adverb-action splits. A. Sophia Koepke and Zeynep Akata provided oversight through weekly meetings to discuss the project and future milestones. All authors contributed to writing the paper.

Chapter 5: EgoCVR: An Egocentric Benchmark for Fine-Grained Composed Video Retrieval.
This work was done with Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. Thomas Hummel and Shyamgopal Karthik were both first authors and contributed equally. Thomas Hummel contributed more to the benchmark creation, and Shyamgopal Karthik to the proposed video framework. Thomas Hummel, Shyamgopal Karthik and Mariana-Iuliana Georgescu all participated in the annotation process for the benchmark. In addition, Mariana-Iuliana Georgescu contributed her domain expertise to both video modelling and benchmark creation. Zeynep Akata provided oversight through weekly meetings to discuss the project and future milestones. All authors contributed to writing the paper.