# Adapting MLLMs for Nuanced Video Retrieval

Piyush Bagad and Andrew Zisserman

VGG, Department of Engineering Science
University of Oxford
https://www.robots.ox.ac.uk/~vgg/research/tara

**Abstract.** Our objective is to build an embedding model that captures the nuanced relationship between a search query and candidate videos. We cover three aspects of nuanced retrieval: (i) temporal, (ii) negation, and (iii) multimodal. For temporal nuance, we consider *chiral actions* that need distinguishing between temporally opposite actions like "opening a door" *vs*. "closing a door". For negation, we consider queries with negators such as "not", "none" that allow a user to specify what they do not want. For multimodal nuance, we consider the task of *composed retrieval* where the query comprises a video along with a text edit instruction. The goal is to develop a unified embedding model that handles such nuances effectively. To that end, we repurpose a Multimodal Large Language Model (MLLM) trained to generate text into an embedding model. We fine-tune it with a contrastive loss on *text alone* with carefully sampled hard negatives that instill the desired nuances in the learned embedding space. Despite the text-only training, our method achieves state of the art performance on *all* benchmarks for nuanced video retrieval. We also analyze how this improvement is achieved, and show that text-only training reduces the *modality gap* between text and video embeddings leading to better organization of the embedding space.

## 1 Introduction

The amount of video content on the Internet continues to grow rapidly with over 3M videos uploaded on YouTube every single day. Efficiently analyzing, organizing and searching through such scale is a necessity. Text provides a concise and efficient interaction layer between users and large-scale video content. Thus, developing reliable and performant video-text models for *retrieval* is crucial. Furthermore, user search queries are often very specific and require *nuanced* video understanding. For example, consider the query "Harry closes the door slowly, with nobody watching". First, to perceive the door closing, the video model must take into account the frame-order, else it cannot distinguish between closing the door *vs*. opening the door. Likewise, it must also understand the pace of action, *i.e*., the door closes *slowly*. Then, it should also understand negation. Finally, the user can augment the text query with an image/video of what Harry looks like. Each of these scenarios demands increasingly specific retrieved videos among a larger pool of related videos. We unify such challenges into the common theme of *nuanced video retrieval*.

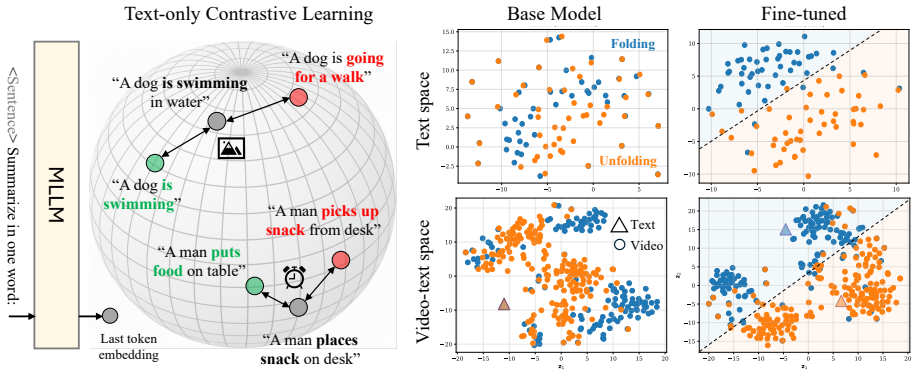



**Fig. 1:** We repurpose an MLLM as a joint video-text embedding model. (Left) We fine-tune it on text samples with contrastive objective but with hard negatives chosen carefully to instill nuance in retrieval, *e.g.*, temporal nuance (shown in the figure with a clock), to distinguish between *chiral actions* that are temporally opposite in nature. (Right) We visualize the tSNE projections of embeddings before and after fine-tuning for a unseen action pair: *folding vs. unfolding* something. In the top row, our fine-tuned model better separates sentence embeddings specifying such pairs. The bottom row shows the joint video-text embedding space is adeptly organized for strong retrieval without training on videos. In the Base Model, before fine-tuning, the text embeddings (shown in △) for the two actions overlap in this visualization. After finetuning, the text embeddings are better separated and aligned with their video clusters.

We consider three aspects of nuanced video retrieval. First, we consider time-sensitive retrieval where queries involve some sort of temporal description and we want to retrieve videos that are *temporally consistent* with the query among a set of videos that share similar spatial contexts but differ exactly in how they vary over time. Consider an example query, "climbing up a ladder". We want to retrieve videos that show a person climbing up a ladder and not those of someone climbing down a ladder. We measure this time-sensitivity on recently proposed benchmarks [4, 18]. Second, we consider retrieval where the query is linguistically subtly modified by negation [1]. The model must understand how such *negation* modifiers change semantics. Third, we consider *multimodal* queries with the task of composed video retrieval (CoVR) [51]. Here the model must understand how the text modifies (edits) the video example in forming the query for retrieval.

In this paper, we develop an encoder for the query and video such that they are mapped to a common embedding space but, unlike a dual encoder (*e.g.*, CLIP), we use a unified encoder to produce both the query and the video embeddings by repurposing an MLLM. These embeddings are obtained by prompting an MLLM to summarize a given query/video and using the last token's final layer hidden representation as the embedding [25], as illustrated in Fig. 1. Prior work has adapted MLLMs to generate embeddings suitable for retrieval, since the original MLLMs are typically trained for text generation on massive multimodal datasets, rather than for retrieval. To adapt them for retrieval, a two-stage

post-training approach is generally used [33, 38, 40]: (i) first, the model is trained contrastively with positive and negative pairs from the text triplets of NLI [19] – this adapts it for retrieval on text; (ii) then, it is further trained on varying amounts of visual-text data to improve cross-modal retrieval.

The key insight of this paper is that by carefully choosing hard negatives to instill each of the three desired nuances: time, negation and multimodal, the MLLM can be fine-tuned using *only text triplets*. For example in Fig. 1, for temporal nuance (🕒), the anchor-positive share similar verb phrase while the hard-negative differs by using a temporally opposite *chiral* verb phrase. Once trained, the model meaningfully encodes novel temporally opposite actions like "folding / unfolding something" as shown in the tSNE projections in Fig. 1. We show that by fine-tuning with this carefully curated training set of text triplets with hard negatives, we are able to achieve state of the art over multiple benchmarks for nuanced video retrieval – this is quite surprising because the method surpasses others who have used far more training data, including a combination of text and visual. We show that a partial reason for this success is that training with this curated set of text triplets closes the *modality gap* [36, 47] between the query and video embeddings.

In summary we make the following contributions: (i) We introduce a method for *Text Adapted Retrieval Alignment* (`TARA`) that involves curating text-triplets with hard negatives for fine-tuning MLLMs to instill strong retrieval capabilities; (ii) We show that applying `TARA` to any base MLLM *always* improves its performance on nuanced retrieval, often substantially, and indeed we achieve state of the art over multiple video retrieval benchmarks (including CiA [4], RTime [18], NegBench [1], CoVR [51]); (iii) We show that text-only fine-tuning also improves performance on the video tasks of standard benchmarks like MMEB-V2 [40]; finally, (iv) We explain why text-only fine-tuning is so effective through the lens of the modality gap.

It is worth noting that since the curated dataset is small, it only takes an hour to fine-tune a 7B model on 8 RTX A6000 GPUs. All code, datasets and state of the art models are released on the project page.

## 2 Related Work

**Video evaluation.** Early work focused on action recognition with datasets like UCF [49], HMDB [28] and retrieval with MSRVTT [61], DiDeMo [2]. The dominance of MLLMs has prompted a suite of benchmarks for question-answering [32, 43, 60] and captioning [10, 52, 62]. However, the community has repeatedly discovered that most of these do not actually test for nuanced aspects, *e.g.* time: a single frame or an orderless set of frames can solve them [8, 11, 13, 23, 29, 63, 77]. Most de-facto video retrieval datasets like MSRVTT [61] also face this issue. Meta-benchmarks like MMEB [26, 40] also inherit this issue as they are comprised of the same datasets. Recent efforts [13, 46, 63] aim to address this issue for video QA tasks. [9] proposed fine-grained evaluation with subtle single-word variations across nouns, verbs, adjectives, *etc*. In this work, we measure nuanced

aspects in video retrieval through specialized benchmarks: [4, 18] for time, [1] for negation, and [51] for composed video retrieval.

**Fine-grained video retrieval.** Contrastively trained dual-encoder models like CLIP tend to focus on global, coarse-level vision-language matching [9, 69]. Early work on fine-grained retrieval explicitly modeled parts-of-speech into a joint multimodal space [59]. [12] proposed a Hierarchical Graph Reasoning model, which decomposes video-text matching into global-to-local levels by splitting text into embeddings for events, actions, entities. Recent work also adapts CLIP for fine-grained retrieval with densely annotated videos [39, 48]. Other work adapt specialized models for specific aspects such as arrow of time [18, 63], negation [57], adverb understanding [16, 17], verb understanding [41], *etc*. In this work, we develop a unified recipe for fine-tuning an MLLM to instill some of these fine-grained qualities in video retrieval.

**Adapting MLLMs for retrieval.** We have witnessed a staggering rise in the abilities of open MLLMs on image [6, 15, 76] and video tasks [6, 52, 56]. A key benefit of open models is that we can analyze and use the hidden representations within the MLLM for retrieval. This has led to a new exciting area of adapting MLLMs as universal encoders [25, 35, 37, 40, 70, 75]. However, as also reflected in benchmarks for MLLMs like MMEB [40], the focus is still on images and static-biased video understanding. Most contemporary work on adapting MLLMs for retrieval fine-tune on large-scale multimodal data using standard contrastive objective [21, 33, 40, 58]. However, much like video-only datasets [8, 30], these training datasets also suffer from focusing on static-biases over the nuances we care about. Some early attempts at text-only fine-tuning [22, 34] map modality-specific encoders (*e.g.*, CLIP for video) to a pre-trained LLM which makes it data-hungry (500K-4.2M training samples). In contrast, we directly adapt an MLLM post-trained for text generation into a retrieval model for temporal, negation and multimodal nuance with just 20K text triplets.

**Modality gap**, a systematic offset between image and text embeddings, was first identified in contrastive dual-encoder models like CLIP [36]. Follow-up work [47, 71] showed that this gap is constant across instances and classes, orthogonal to both modality subspaces. These insights led to methods showing that closing the gap (e.g., via mean centering [22, 31, 66]) enables text-only training for tasks that typically need paired multimodal data [67, 72]. Jiang et al. [25] argue that using the careful prompt while extracting embedding from an MLLM dissolves the modality gap which enables training on text alone. They train on the NLI dataset [19] that has 275K text triplets. Related work on videos also involves training on text but is complemented with training on videos either before [62] or after [38] text training. More recently, attention has turned to the modality gap in multimodal large language models (MLLMs). Methods based on preference/embedding alignment [65, 74] show that reducing the gap in MLLMs yields clear benefits for retrieval, suggesting it remains a key bottleneck even in unified architectures. We show that a simple text-only fine-tuning of an MLLM reduces the modality gap and better organizes the multimodal embedding spaces.

## 3 Method

Our goal is to adapt an MLLM $\mathcal{M}$ for video retrieval. Given a query $\mathbf{q}$ and video candidates $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N\}$, we want to rank the candidates by relevance to the query using $\mathcal{M}$. Let $f_{\mathcal{M}}(.) : \mathcal{X} \rightarrow \mathbb{R}^d$ denote a function to extract an embedding out of $\mathcal{M}$ where $\mathcal{X}$ is the space of all inputs and $d$ is the embedding dimension. Then, we can compute similarity $s$ between $\mathbf{q}$ and $\mathbf{c}$ as:

$$s(\mathbf{q}, \mathbf{c}) = f_{\mathcal{M}}(\mathbf{q}).f_{\mathcal{M}}(\mathbf{c})^T \tag{1}$$

Since $\mathcal{M}$ is only trained to generate text, the challenge is two-fold: (i) design $f_{\mathcal{M}}$, and (ii) fine-tune $\mathcal{M}$ such that $f_{\mathcal{M}}$ encodes desired semantic nuances.

### 3.1 Designing $f_{\mathcal{M}}$: Extracting embeddings from MLLMs

Following the work of Jiang et al. [24, 25] we use prompts to encourage the MLLM to condense the semantic meaning of a sentence or video into the hidden state of the next token, which is used as the sentence or video embedding. This is termed an 'Explicit One-word Limitation' (EOL) prompt. For example, an EOL prompt for a video is constructed as: "`<video>: Summarize the video in one word:` ". Then, in generating the next token, the final layer hidden representation is used as the video embedding. This extends over the E5-V work of [25] where only images, text, and their combination were used. For a multimodal input (a video with a text edit instruction), we use a slightly different EOL prompt: "`Source video: <video>; Edit instruction: <text>; Imagine this edit instruction being applied to the source video. Summarize the resulting edited video in one word:`". In our notation, this EOL-based embedding extraction process represents $f_{\mathcal{M}}$.

### 3.2 Training $\mathcal{M}$: Text-only training enables cross-modal retrieval

Text-only fine-tuning is straightforward: for a triplet $(\mathbf{t}_i, \mathbf{t}_i^+, \mathbf{t}_i^-)$ where $\mathbf{t}_i^+$ is the embedding of a positive match sentence for the anchor $\mathbf{t}_i$, and $\mathbf{t}_i^-$ a hard-negative, the contrastive loss is given by:

$$\mathcal{L}_{\text{con.}}(\mathcal{M}) = -\log\left(\frac{e^{\text{s}(\mathbf{t}_i, \mathbf{t}_i^+)/\tau}}{\sum_j e^{\text{s}(\mathbf{t}_i, \mathbf{t}_j^+)/\tau} + e^{\text{s}(\mathbf{t}_i, \mathbf{t}_j^-)/\tau}}\right), \tag{2}$$

where $\text{s}(\mathbf{x}, \mathbf{y})$ denotes cosine similarity between $\mathbf{x}, \mathbf{y}$. As will be shown in Sec. 5, this text-only contrastive training reduces the modality gap between video and text embedding spaces. This leads to a meaningful re-organization of the embedding space which in-turn results in improved cross-modal retrieval. We refer to this text-only fine-tuning recipe as `TARA`: Text Adapted Retrieval Alignment.

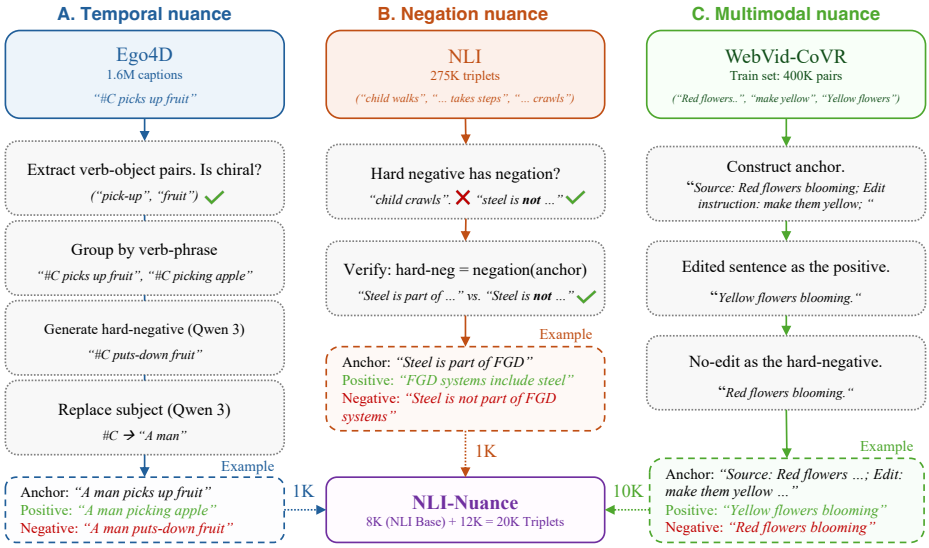



**Fig. 2: Text dataset construction.** We augment samples in NLI with text triplets designed specifically to instill desired nuances in the embedding model. We call the resulting text dataset as NLI-Nuance comprising of 8K triplets from NLI and 12K triplets covering temporal, negation and multimodal nuances. More examples in arXiv.

### 3.3 Instilling nuance in retrieval with hard negatives

Previous work [25, 38, 62] has used text triplets from NLI [19] where "entailment" pairs are treated as positives and "contradiction" pairs as hard negatives. We argue that this helps with image understanding but is insufficient to instill nuance in retrieval. Consider the triplet ("child walking on concrete ledge", "... takes steps along the concrete", "... crawls along sandy ledge"). The positive and anchor describe a child walking on concrete while the negative differs in *crawling on sandy edge*. If these were captions of videos, a single frame would suffice to recognize the positive to be closer to the anchor, with no need for nuanced temporal understanding. Thus, we augment NLI with text triplets specifically designed to instill nuanced understanding in retrieval for (i) temporal, (ii) negation and (iii) multimodal. We sketch the steps to generate hard-negatives in Fig. 2.

A. **Temporal nuance.** We extract verb-object pairs in Ego4D [20] and pick only those with a *chiral* verb [4] based on a pre-compiled list of chiral verbs generated by Claude. Positives are chosen s.t. they share the chiral verb with the anchor while hard-negatives are generated by prompting Qwen 3 as shown with the example in Fig. 2. In the arXiv version, we detail verification of the hard-negatives via LLM-as-a-Judge followed by a manual review.
B. **Negation nuance.** We filter those triplets from NLI where the hard-negative has a negation operator (*e.g.*, "not"). Example is shown in Fig. 2.
C. **Multimodal nuance.** Building on the idea of converting a video task to a text-only task, we translate Composed Video Retrieval to a Composed Text Retrieval. Starting with train set of WebVid-CoVR [51], we replace the source video and to-be-retrieved video with their corresponding captions as shown in Fig. 2.

In order to retain the static understanding of the model, we also retain text triplets from the original NLI dataset. In summary, we build a text dataset called **NLI-Nuance** consisting of a total of $N$=20,000 triplets. We fine-tune a given MLLM on this dataset. Note that this dataset consists of text samples sourced from Ego4D train set, WebVid-CoVR train set and the original NLI dataset. *It does not use any text captions from the evaluation datasets.*

## 4 Experimental Results

Our experiments are designed to measure if `TARA` generalizes to nuanced video retrieval. In the following Sections, we introduce evaluation datasets for each aspect of nuanced video retrieval. As each dataset may have a different evaluation metric, we introduce these as we describe the dataset. In Sec. 4.1, we evaluate temporal nuance on time-sensitive benchmarks for video retrieval. In Sec. 4.2, we probe two kinds of linguistic nuance: negation and adverbs in text queries. In Sec. 4.3, we test multimodal nuance through composed video retrieval. In each case, we show that `TARA` improves performance, sometimes significantly, and this applies across multiple open-source MLLMs. In Sec. 4.4, we show that `TARA` does *not* harm retrieval on standard evaluation datasets. For each evaluation dataset, we provide the number of query and number of candidates in Tab. 1. Ablations on training data composition and size are provided in the arXiv version.

**Implementation details.** While we experiment with 5 MLLMs (Qwen2VL [53], Qwen3VL [33], InternVL3 [76], Tarsier 1 [52] and 2 [68]), we obtain best results with fine-tuning Tarsier 2 which we treat as our default model. We only fine-tune the LLM weights and freeze the vision and projection networks. We train for 2 epochs with batch size of 768 and base learning rate $2e{-}5$. On 8 Nvidia RTX A6000 GPUs, it takes less than an hour to train `TARA` . During inference, we use $F$=16 uniformly spaced frames in the video.

**Table 1: Evaluation Dataset statistics.**

| | **CiA** (All) | | | **NegBench** | | **WebVid-CoVR** |
|---|---|---|---|---|---|---|
| | SSv2 | EPIC | Charades | MS-COCO | MSRVTT | |
| Number of Queries | 32 | 128 | 56 | 5915 | 995 | 2556 |
| Average number of Candidates | 1430 | 3108 | 5498 | 5915 | 1000 | 2556 |

### 4.1 Temporal Nuance

**CiA evaluation dataset.** First, we use the Chirality in Action (CiA) dataset proposed by Bagad et al. [3]. It is made up of temporal actions from three datasets: SSv2, EPIC and Charades. We use the corresponding text captions as queries and measure mAP over the retrieved videos. CiA has three settings: (i) *chiral*: the video gallery $\mathcal{G}$ consists of only the correct action and its temporal opposite (*e.g.*, for query 'opening door', $\mathcal{G}$ has videos of opening and closing doors.), (ii) *static*: $\mathcal{G}$ consists of the correct action and all other temporally irrelevant actions (*e.g.*, for query 'opening door', $\mathcal{G}$ has distractors 'folding paper', 'wrapping gift', *etc.*), (iii) *all*: $\mathcal{G}$ consists of the correct action and all other videos, *i.e.*, it has the temporal opposite action as well as temporally irrelevant

actions. Results are presented in Tab. 2. `TARA` comprehensively achieves state of the art performance on CiA while being fine-tuned on text alone. Note, TARA always improves over the base model for chiral action retrieval. The degree of improvement depends on the base model: Q3VL-Emb. [33] is already trained contrastively on multimodal data, so TARA shows smaller benefits. But with other base models trained autoregressively, TARA shows much larger benefits.

**Table 2: Results on CiA-Retrieval.** We report mAP for T2V retrieval on CiA [4]. *Chiral* denotes retrieving from a video gallery $\mathcal{G}$ of only temporally antonymous (chiral) samples , *Static*: $\mathcal{G}$ has non-chiral actions and *All*: $\mathcal{G}$ includes chiral/non-chiral. `TARA` always improves over base model; achieves state-of-the-art performance with Tarsier-2.

| Method | Fine-tuning data | | SSv2 | | | EPIC | | | Charades | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Size (K) | Modalities | *Chiral* | *Static* | *All* | *Chiral* | *Static* | *All* | *Chiral* | *Static* | *All* |
| Chance | - | - | 50.0 | 6.3 | 3.1 | 50.0 | 1.5 | 0.8 | 50.0 | 3.6 | 1.8 |
| **Dual encoder models** | | | | | | | | | | | |
| CLIP (avg.) [45] | - | - | 52.0 | 18.3 | 12.7 | 51.0 | 12.0 | 7.0 | 48.4 | 13.2 | 6.5 |
| DINO.txt [27] | - | - | 52.1 | 19.9 | 13.1 | 50.6 | 14.6 | 6.0 | 50.7 | 19.0 | 10.1 |
| Perception Enc. [7] | - | - | 50.1 | 32.7 | 17.2 | 48.5 | 4.2 | 1.8 | 51.3 | 12.8 | 7.2 |
| InternVideo 2 [55] | - | - | 52.5 | 35.7 | 20.6 | 48.3 | 22.1 | 8.8 | 50.7 | 11.9 | 11.9 |
| **MLLMs** | | | | | | | | | | | |
| VLM2Vec-V2 [40] | 1700 | 🎞🖼✎📄 | 58.8 | 27.8 | 15.9 | 49.4 | 25.4 | 12.9 | 53.5 | 18.8 | 10.5 |
| LAMRA [38] | 1400 | 🖼✎ | 55.3 | 15.2 | 7.8 | 53.7 | 11.3 | 9.0 | 52.1 | 21.2 | 11.3 |
| GVE-7B [21] | 13000 | 🎞🖼✎ | 53.4 | 17.1 | 4.0 | 54.7 | 22.4 | 7.3 | 54.2 | 21.0 | 10.2 |
| E5-V [25] | 300 | ✎ | 52.6 | 18.9 | 14.7 | 57.1 | 16.4 | 6.5 | 48.9 | 19.7 | 7.1 |
| ArrowRL [63] | 25 | 🎞✎ | 67.5 | 33.8 | 22.5 | 55.7 | 12.4 | 9.6 | 57.1 | 18.6 | 12.2 |
| CaRe [62] | 275 | 🎞✎ | 66.4 | 46.2 | 23.7 | 62.3 | 25.0 | 16.9 | 56.1 | 25.2 | 12.9 |
| Qwen2VL-7B [53] | - | - | 60.2 | 28.0 | 17.3 | 53.7 | 11.2 | 9.6 | 55.9 | 16.7 | 8.1 |
| Qwen2VL + `TARA` | 20 | ✎ | 70.1 | 39.6 | 27.4 | 65.0 | 27.9 | 20.8 | 65.5 | 31.8 | 22.7 |
| Qwen2.5VL-7B [6] | - | - | 67.6 | 31.5 | 20.6 | 55.4 | 12.4 | 9.5 | 55.8 | 17.0 | 11.1 |
| Qwen3VL-7B [5] | - | - | 56.8 | 16.4 | 11.0 | 57.8 | 2.9 | 1.8 | 52.1 | 9.1 | 6.6 |
| Qwen3VL + `TARA` | 20 | ✎ | 63.6 | 23.5 | 13.6 | 64.0 | 25.1 | 20.4 | 59.4 | 17.5 | 12.4 |
| Qwen3VL-Emb. [33] | NA | 🎞🖼✎📄 | 72.0 | 43.4 | 31.8 | 62.1 | 28.6 | 20.6 | 65.3 | 37.3 | 26.1 |
| Qwen3VL-Emb + `TARA` | 20 | ✎ | 73.0 | 51.5 | 37.8 | 64.0 | 25.1 | 20.1 | 68.3 | 36.4 | 28.7 |
| InternVL3-8B [76] | - | - | 56.8 | 12.5 | 6.5 | 57.5 | 6.2 | 4.3 | 54.0 | 16.1 | 9.4 |
| InternVL3 + `TARA` | 20 | ✎ | 73.1 | 36.0 | 27.6 | 68.9 | 33.9 | 24.3 | 64.1 | 27.5 | 19.8 |
| Tarsier [52] | - | - | 76.4 | 51.8 | 35.0 | 62.7 | 22.1 | 17.4 | 64.7 | 25.9 | 17.7 |
| Tarsier + `TARA` | 20 | ✎ | 84.9 | 61.4 | 49.5 | 80.5 | 37.3 | 31.7 | 71.7 | 38.7 | 27.8 |
| Tarsier2-7B [68] | - | - | 77.7 | 26.9 | 24.0 | 67.4 | 22.0 | 15.3 | 60.5 | 13.4 | 9.2 |
| Tarsier 2 + `TARA` | 20 | ✎ | **88.9** | **66.7** | **58.6** | **81.1** | **45.6** | **38.9** | **71.4** | **38.6** | **29.0** |

**RTime.** We also evaluate on the Reversed in Time (RTime) benchmark by Du et al. [18] that uses the arrow of time to generate distractor videos: a video with caption 'a woman opening a laptop' is time-reversed with a new caption 'a woman closing a laptop'. For V2T, the task is to choose between the correct and reversed caption. Likewise, for T2V, the task is to choose between the actual and reversed video. Results (R@1) are shown in Tab. 3. `TARA` outperforms all competing methods, including those from [18] that are fine-tuned on the RTime training set.

**Table 3:** R@1 on RTime.

| Method | T2V | V2T |
|---|---|---|
| **Zero-shot** | | |
| Singularity | 48.7 | 49.9 |
| Internvideo2-1B | 50.0 | 51.0 |
| Qwen2.5VL | 53.4 | 66.6 |
| Tarsier 2 | 58.8 | 59.5 |
| Tarsier 2 + `TARA` | **67.2** | **77.9** |
| **Fine-tuned on RTime** | | |
| CLIP4Clip | 49.8 | 49.8 |
| UMT | 51.2 | 51.3 |
| UMT-Neg | 54.5 | 54.2 |
| ArrowR-Qwen2 | 57.1 | 68.8 |
| ArrowRL-Qwen2.5 | 55.6 | 69.6 |

### 4.2 Negation Nuance

We use the NegBench [1] benchmark to evaluate text-to-image ( COCO ) and text-to-video ( MSRVTT ) retrieval where text queries involve negation. We evaluate on standard queries (*e.g.*, 'a picture of a dog') and on negated queries (*e.g.*, 'a picture of a dog but *not on grass*'). We compare methods from Alhamoud et al. [1] fine-tuned on variations of Conceptual Captions (CC) and report R@5. As shown in Tab. 4, TARA outperforms all competing methods on both datasets. Relatedly, we also show TARA's *adverb* understanding in the arXiv, *e.g.*, given a video of a person walking, distinguishing if the walking is "slow"/"quick".

**Table 4: NegBench Evaluation** checks understanding of negation in queries. TARA (zero-shot) beats strong baselines in [1].

| Method | Fine-tuning data | COCO | | MSR-VTT | |
|---|---|---|---|---|---|
| | | R@5 | R-Neg@5 | R@5 | R-Neg@5 |
| CLIP | None | 54.8 | 48.0 | 50.6 | 45.8 |
| | CC (🖼, ✎) | 58.8 | 54.5 | 53.7 | 49.9 |
| | CC-NegCap (🖼, ✎) | 58.5 | 57.8 | 54.1 | 53.5 |
| | CC-NegFull (🖼, ✎) | 54.2 | 51.9 | 46.9 | 43.9 |
| NegCLIP [1] | None | 68.7 | 64.4 | 53.7 | 51.0 |
| | CC (🖼, ✎) | 70.2 | 66.0 | 56.4 | 52.6 |
| | CC-NegCap (🖼, ✎) | 68.6 | 67.5 | 56.5 | 54.6 |
| | CC-NegFull (🖼, ✎) | 69.0 | 67.0 | 54.0 | 51.5 |
| Tarsier 2 | None | 33.3 | 21.5 | 25.6 | 18.9 |
| Tarsier 2 + TARA | NLI-Nuance (✎) | **76.7** | **73.6** | **65.1** | **65.0** |

### 4.3 Multimodal Nuance

Since TARA is based on extracting embeddings out of an MLLM, it inherits the flexibility of an MLLM to take as input the composition of video-text together. We leverage this and evaluate on the task of Composed Video Retrieval introduced by Ventura et al. [51]. The queries are composed of a video and a text edit instruction. Similar to the EOL prompt for each modality, we modify the EOL prompt slightly for joint video-text inputs. We provide the full prompts used in the arXiv version. We evaluate on the WebVid-CoVR [51] test set consisting of 2,556 query-video samples. TARA is evaluated in a zero-shot manner. As shown in Tab. 5, TARA outperforms even methods fine-tuned on WebVid-CoVR.

**Table 5: Composed video retrieval on WebVid-CoVR.** TARA beats all competing methods while being fine-tuned only with text data.

| Method | R@1 | R@5 | R@10 |
|---|---|---|---|
| Chance | 0.1 | 0.2 | 0.4 |
| **Zero-shot** | | | |
| BLIP (T) | 19.7 | 37.1 | 45.9 |
| BLIP (V) | 34.9 | 59.2 | 68.0 |
| CLIP (V+T) | 44.4 | 69.1 | 77.6 |
| BLIP (V+T) | 45.5 | 70.5 | 79.5 |
| Tarsier 2+TARA | **66.3** | **86.7** | **91.5** |
| **Fine-tuned on CoVR** | | | |
| BLIP (T) | 23.7 | 45.9 | 55.1 |
| BLIP (V) | 38.9 | 65.0 | 74.0 |
| CLIP (V+T) | 50.6 | 77.1 | 85.1 |
| BLIP (V+T) | 50.6 | 74.8 | 83.4 |
| BLIP (V+T) | 51.8 | 78.3 | 85.8 |
| Ventura et al. [51] | 53.1 | 79.9 | 86.9 |
| Ventura et al. [50] | 59.8 | 83.8 | 91.3 |

### 4.4 Standard Benchmarks

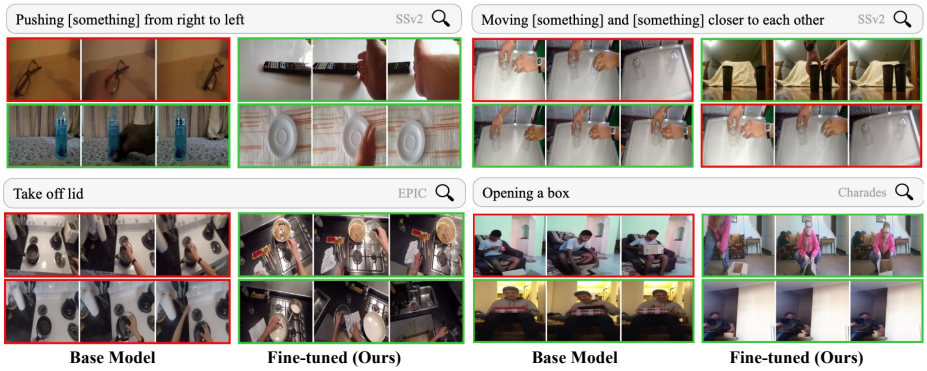
Does text-only fine-tuning on our NLI-Nuance dataset deteriorate performance on standard benchmarks for video recognition/retrieval? To test this, we evaluate `TARA` on video tasks from the MMEB-V2 benchmark [40] across 10 standard datasets for classification and retrieval. The results are tabulated in Tab. 6. We compare with the top multimodal embedding models [40, 70] including recently proposed thinking-augmented retrieval [14]. First, we note that `TARA` comprehensively improves upon its base model (Tarsier 2). Second, despite being fine-tuned only on text samples, it is competitive with other methods trained on massive amounts of multimodal data. In terms of average classification/retrieval performance, it is second only to Qwen3-VL-Embedding [33]. Furthermore, we also explore a a simple ensemble (Concatenation of Embeddings) of TARA Qwen3-VL-Embeddings and show that it outperforms the latter, demonstrating the complementary abilities that TARA possesses.

**Table 6: Results on subset of tasks in MMEB-V2.** Despite fine-tuning only on text samples, `TARA` comprehensively improves upon the base model and is only second to Qwen3VL-Embedding trained on massive amount of multimodal data. TARA $\oplus$ Q3VLE denotes a concatenation of TARA and Qwen3VL-Embedding. It outperforms Qwen3VL on average highlighting the complementary abilities that TARA possesses.

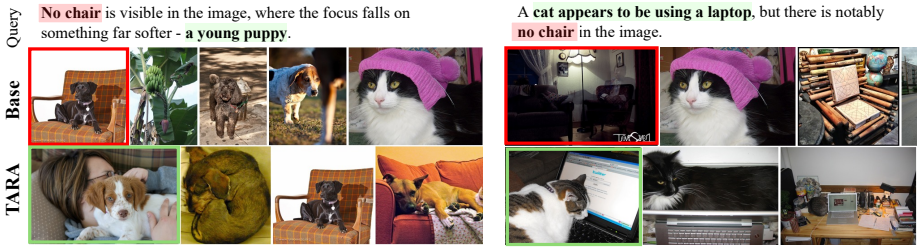
| Method | Video classification | | | | | | Video retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | UCF | HMdb | K700 | BF | SSv2 | Avg. | MSR | MSVD | DDMo | YC2 | VTX | Avg. |
| Chance | 1.0 | 2.0 | 0.1 | 2.2 | 0.6 | - | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | - |
| ColPali v1.3 | 49.4 | 24.8 | 23.4 | 10.9 | 25.1 | 26.7 | 17.6 | 45.4 | 22.8 | 5.3 | 16.7 | 21.6 |
| GME-2B | 52.4 | 43.4 | 35.2 | 13.6 | 29.9 | 34.9 | 27.3 | 47.6 | 22.0 | 7.9 | 23.0 | 25.6 |
| GME-7B | 54.7 | 47.9 | 39.7 | 14.3 | 30.6 | 37.4 | 31.8 | 49.7 | 26.4 | 9.1 | 24.9 | 28.4 |
| LamRA-Qwen2 | 60.4 | 40.5 | 42.3 | 16.9 | 36.3 | 39.3 | 22.1 | 46.1 | 24.8 | 9.2 | 19.1 | 24.3 |
| LamRA-Qwen2.5 [38] | 53.0 | 33.8 | 32.1 | 20.1 | 25.3 | 32.9 | 25.0 | 41.9 | 22.8 | 7.5 | 18.7 | 23.2 |
| VLM2Vec-2B | 57.5 | 33.8 | 31.4 | 13.4 | 30.9 | 33.4 | 25.2 | 38.2 | 19.4 | 4.1 | 16.2 | 20.6 |
| VLM2Vec-7B | 61.8 | 42.2 | 35.5 | 23.8 | 32.1 | 39.1 | 34.5 | 46.7 | 29.3 | 9.0 | 25.5 | 29.0 |
| VLM2Vec-V2.0 [40] | 60.0 | 40.9 | 38.0 | 14.8 | 42.8 | 39.3 | 28.3 | 48.1 | 30.4 | 10.6 | 26.5 | 28.8 |
| TTE-7B [14] | 78.6 | 63.9 | 55.6 | 34.2 | 55.3 | 57.5 | 39.5 | 59.4 | 36.3 | 20.3 | 32.6 | 37.6 |
| Tarsier 2 | 37.9 | 17.4 | 29.6 | 36.1 | 15.9 | 27.4 | 9.5 | 39.8 | 12.2 | 3.9 | 16.6 | 16.4 |
| Tarsier 2 + `TARA` (Ours) | 80.3 | 69.0 | 59.4 | 45.6 | 76.4 | 66.1 | 40.7 | 82.2 | 36.8 | 16.7 | 53.2 | 45.9 |
| Qwen3-VL-Embedding [33] | **94.6** | 77.5 | **71.2** | 67.2 | 76.9 | 77.5 | 53.8 | 87.2 | **56.1** | **32.8** | 64.8 | 58.9 |
| TARA $\oplus$ Q3VLE | 94.3 | **78.3** | 70.0 | **68.6** | **81.4** | **78.5** | **54.5** | **88.4** | **56.1** | 32.1 | **66.2** | **59.5** |

**Qualitative results.** We visualize qualitative retrieval results for each of three nuances: temporal nuance in Fig. 3a, negation nuance in Fig. 3b and multimodal nuance in Fig. 3c. In each case, `TARA` fine-tuned only on text, retrieves results accurately compared to the base model.

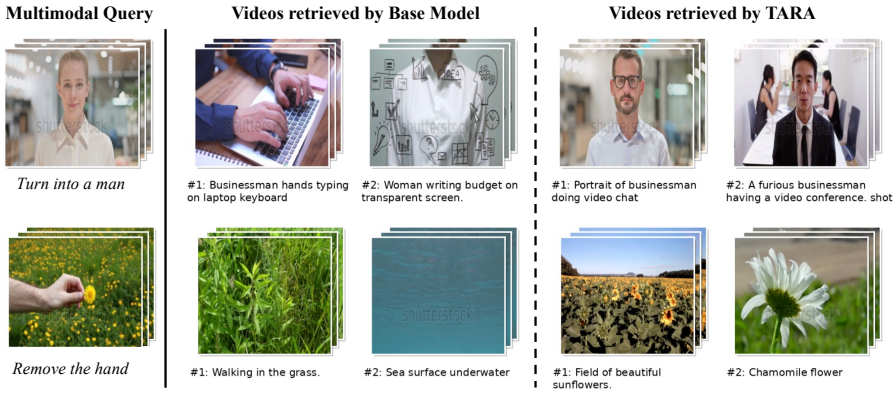
**Controlled evaluation with fixed base model.** To reinforce the effectiveness of `TARA` , we compare with methods that fine-tune the same base model (Qwen2VL) with multimodal data and various objectives. Results are shown in Tab. 7. On avg. across our benchmarks for nuanced retrieval, `TARA` is +10 points better than the second best method despite being fine-tuned only on text.




**(a) Chiral actions.** Queries require *temporal nuance*. For each query, we show top-2 retrieved videos (green border: correct result). `TARA` retrieves accurately particularly for chiral queries.




**(b) Negation in query.** Base model is distracted by negation, while `TARA` fine-tuned model retrieves accurately. Top 1 results is marked by red (incorrect) or green (correct).




**(c) Composed video retrieval.** (Left) Input query composed of a video and edit instruction. Top-2 results from base model (middle) and `TARA` (right). Captions of candidates are not used.

**Fig. 3:** Qualitative retrieval results under different types of nuance.

# 5 Analysis

## 5.1 Why is text-only training effective? A study on modality gap

**Experimental setting.** The modality gap in MLLMs is relatively understudied, particularly in the context of video-text retrieval. We use the evaluation set of MSRVTT to visualize and measure the modality gap. We pick $N$=1000 video-

**Table 7: Comparison with the same base model.** We compare methods that use the same base model but fine-tune it with various data. `TARA` fine-tuned on text alone almost always outperforms existing work that uses multimodal data.

| Method | Fine-tuning modalities | CiA (SSv2) | | | RTime | | WebVid-CoVR | | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| | | Chiral | Static | All | T2V | V2T | R@1 | R@5 | - |
| Base (Qwen2VL-7B) | - | 60.2 | 28.0 | 17.3 | 59.9 | 64.7 | 15.5 | 34.4 | 40.0 |
| ArrowRL [44] | 🎞✎ | 67.5 | 33.8 | 22.5 | 57.1 | 68.8 | 41.8 | 67.9 | 51.3 |
| CaRe (Stage 2) [62] | 🎞✎ | 66.4 | **46.2** | 23.7 | 59.8 | 69.9 | 35.6 | 61.8 | 51.9 |
| LAMRA [38] | 🖼✎ | 55.3 | 15.2 | 7.8 | 57.9 | 63.9 | 31.5 | 56.1 | 41.1 |
| VLM2Vec-2.0 [40] | 🎞🖼✎ | 58.8 | 27.8 | 15.9 | 54.3 | 61.8 | 42.9 | 67.1 | 46.9 |
| Base + `TARA` (Ours) | ✎ | **72.7** | 39.6 | **28.3** | **65.9** | **73.8** | **44.8** | **72.6** | **56.8** |

text pairs, embed them through an MLLM using a standard EOL prompt [25] and measure the modality gap in the shared embedding space based on definitions in [36]. It is denoted by $\Delta_{\text{gap}}$ defined as the avg. of difference between embeddings of paired samples from two modalities and we report norm of this vector. tSNE projections and $||\Delta_{\text{gap}}||_2$ are shown in Fig. 4.

**MLLMs exhibit modality gap despite shared backbone.** Unlike dual-encoder models like CLIP, MLLMs share the LLM backbone that ingests token embeddings either from video or text. Despite that, we observe clear separation between modalities for embeddings from Qwen2VL as also measured quantitatively based on the metric from [36] (similar results for InternVL3, Tarsier, Qwen3VL included in the arXiv version). We hypothesize that this happens because video and text tokens come from two different pathways. Video is tokenized via a pre-trained vision encoder and then projected in the LLM space by an MLP while text is tokenized via learned embedding matrices. Modality gap hinders retrieval capabilities in several ways: (i) many components of the embeddings end up encoding the modality wasting representational capacity to embed fine-grained semantics, (ii) cross-modal cosine similarities are skewed by this gap and show smaller variation which affects discrimination between fine-grained pairs.

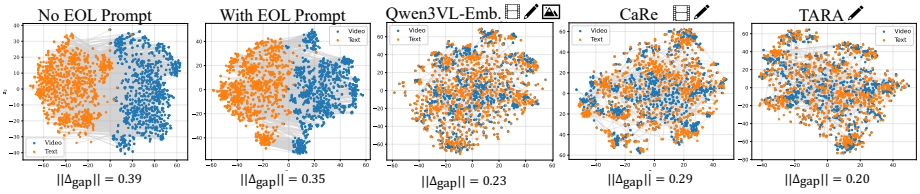



**Fig. 4: Modality Gap in MLLMs** as measured on video-text pairs from MSRVTT. (i) Using EOL prompt with Qwen2VL model alone does not fix modality gap. (ii) While methods like Qwen3VL-Embedding [33], CaRe [62] train on video 🎞/ image 🖼/ text ✎data, we show that text-only ✎fine-tuning is sufficient to reduce the modality gap.

**EOL prompt does not reduce modality gap; text-only training does!** Jiang et al. [25] using LLaVA-Next-8B as the base model report that for image-text data, using the EOL prompt dissolves the modality gap. However, we find this is untrue for video-text data using Qwen2VL as shown in the first two images

in Fig. 4 (as well as for InternVL3, Qwen3VL, Tarsier as reported in the arXiv version). Contemporary methods such as CaRe [62], Qwen3VL-Embedding [33] trained on large amount of video-text data show reduced modality gap. Other prior work uses various hand-crafted heuristics to reduce this gap [31, 64, 66, 72]. In contrast, we discover that simple text-only fine-tuning is sufficient to reduce the modality gap as shown in the last image.

But why does this happen? We hypothesize two reasons for this: (i) unimodal contrastive learning leads to *uniformity* pressure, i.e., it tends to spread apart the embeddings to occupy a larger subspace of the hypersphere [42, 54]. This uniformity makes the text embeddings more centered around the origin. Since the LLM shares the weights while ingesting a video, the video embeddings also spread around the origin. This means the centroids of both text/video spaces move closer to the origin and thus the reduction in modality gap. Additionally, the MLLM is typically pre- and post-trained to map video to text whereas here, we use text-only contrastive training which makes a difference to the MLLM's embedding space (ii) Since the negatives in the contrastive setup are also text samples, it encourages the embedding to ignore the modality information and focus more on semantics. This leaves more room for the embeddings to encode semantics thus improving retrieval capabilities. In the arXiv version, we also empirically show (1) modality gap ($\|\Delta_{\text{gap}}\|$ from [71]) reduces in tandem with the training loss; the downstream retrieval (SSv2) improves as $\|\Delta_{\text{gap}}\|$ decreases. (2) We also measure the **uniformity pressure loss** ($\mathcal{L}_{\text{uniform}}$ defined in [54]) in each modality. In both modalities, TARA reduces uniformity loss compared to the base model, *i.e.*, the embeddings are more spread apart in TARA.

**Analyzing learned embeddings.** To interpret `TARA`'s embeddings, we feed them through the LM head and visualize the top 20 tokens by logits for a video and its caption independently (Fig. 5). Unlike the base model which generates text only and produces a semantically meaningless first token, `TARA`'s embeddings reflect the relevant action (*e.g.*, "closing the box"), as fine-tuning trains them to discriminate between nuanced captions. More examples in the arXiv version. We also visualize embeddings for chiral action pairs (*e.g.*, "put-down" *vs*. "pick-up") [4] via 2D tSNE in Fig. 6. Fine-tuning the shared LLM weights to distinguish chiral text pairs also improves video embedding separation, while the reduced $\|\Delta_{\text{gap}}\|$ aligns text-video embeddings by class, enabling strong retrieval.

### 5.2 Theoretical analysis

To answer why text-only training is so effective in cross-modal retrieval, we also theoretically analyze the two losses: $\mathcal{L}_{\text{text-only}}$ (text-only contrastive loss) and $\mathcal{L}_{\text{t2v}}$ (text-to-video contrastive loss). By using the characterization of modality gap in prior work [71, 73] (MGA), we prove that decreasing $\mathcal{L}_{\text{text-only}}$ also decreases $\mathcal{L}_{\text{t2v}}$ up to an irreducible alignment noise (Proposition 1).

**Assumption 1 (Modality Gap Assumption (MGA) [71, 73])** *A video-text pair* $(v, t)$ *with embeddings* $\mathbf{e}_v, \mathbf{e}_t$ *satisfy* $\mathbf{e}_v - \mathbf{e}_t = \mathbf{c}_\perp + \boldsymbol{\epsilon}$, *where* $\mathbf{c}_\perp$ *is a constant*

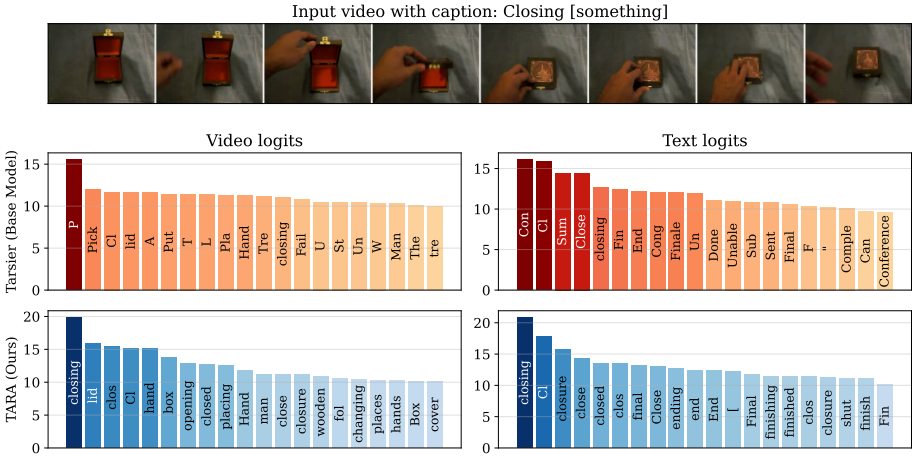



**Fig. 5: Token distribution for video/text embedding.** Compared to base model, `TARA` top tokens for the video/text correspond to semantically relevant action in the video, *i.e.*, "closing the box".

*vector representing the modality gap, orthogonal to both embedding spans, and* $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \nu^2\mathbf{I})$ *is alignment noise. See Fig. 2 of [73].*

**Proposition 1 (Ours).** *Given a text triplet* $(t^q, t^+, t^-)$ *with* $t^q$ *as the query,* $t^+$ *as the positive and* $t^-$ *as the negative, the text-only contrastive triplet loss reduces to* $\mathcal{L}_{text\text{-}only} = \mathbb{E}\big[-\log\sigma\big(\mathbf{e}_{t^q}^\top(\mathbf{e}_{t^+}-\mathbf{e}_{t^-})\big)\big]$ *(*$\sigma$ *denotes sigmoid) while the text-to-video contrastive loss is given by* $\mathcal{L}_{t2v} = \mathbb{E}\big[-\log\sigma\big(\mathbf{e}_{t^q}^\top(\mathbf{e}_{v^+}-\mathbf{e}_{v^-})\big)\big]$*. Under the MGA,* $\mathcal{L}_{t2v} \geq \mathcal{L}_{text\text{-}only}$*, and decreasing* $\mathcal{L}_{text\text{-}only}$ *decreases* $\mathcal{L}_{t2v}$ *up to a factor of alignment noise.*

*Proof (Proof sketch).* By the MGA assumption, the modality gap vector $\boldsymbol{c}_\perp$ cancels exactly in $\boldsymbol{e}_{v^+}-\boldsymbol{e}_{v^-}$, reducing the difference between the two losses to alignment noise $\boldsymbol{\delta}=(\boldsymbol{\epsilon}_+ - \boldsymbol{\epsilon}_-)\sim\mathcal{N}(\mathbf{0}, 2\nu^2\boldsymbol{I})$. Convexity of $-\log\sigma(\cdot)$ and Jensen's inequality then give $\mathcal{L}_{\text{t2v}} \geq \mathcal{L}_{\text{text-only}}$, with the gap controlled by $\nu^2$. Since both losses are strictly monotonically decreasing in the same scalar $a = \boldsymbol{e}_{t^q}^\top(\boldsymbol{e}_{t^+}-\boldsymbol{e}_{t^-})$, any improvement in $\mathcal{L}_{\text{text-only}}$ implies a corresponding improvement in $\mathcal{L}_{\text{t2v}}$. A similar proof holds for video-to-text loss exploiting $\boldsymbol{c}_\perp \perp \text{span}(t)$., *i.e.*, the modality gap vector $c_\perp$ is perpendicular to the hyperplane spanned by the text embeddings. A complete proof is included in the arXiv version.

## 6 Conclusion and Discussion

In this work, we proposed a recipe, `TARA`, to adapt an MLLM for various kinds of nuances in video retrieval: temporal, negation and multimodal. It is based on fine-tuning with contrastive loss on text triplets where the hard negatives are chosen carefully to inject the desired nuances in the embedding space. Our method achieves state of the art performance on all benchmarks evaluating nuanced retrieval while also improving upon the base model on standard retrieval

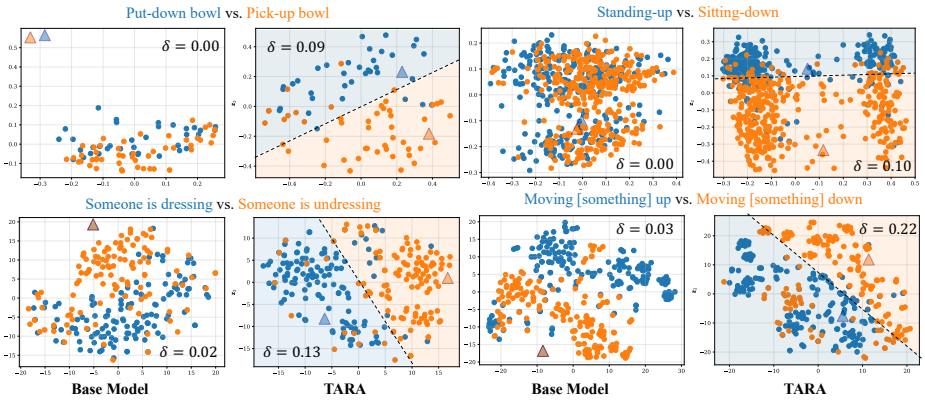



**Fig. 6: Visualizing video-text embeddings.** For each chiral pair, we visualize embeddings before and after `TARA` fine-tuning (text embeddings: △). Post fine-tuning, video embeddings separate better, and text embeddings align with their corresponding videos due to reduced modality gap. We quantify this alignment ($\delta$) as the difference in avg. similarity between text and matched *vs*. mismatched videos. For the base model, $\delta \sim 0$, indicating poor text-video association, whereas `TARA` achieves much higher $\delta$.

datasets. We also present an analysis of why text-only fine-tuning is so effective through the lens of the modality gap. Specifically, contrastive training on text triplets reduces the modality gap as a result of the uniformity pressure [54], *i.e.*, the video and text subspaces expand on the hypersphere and thus get closer to each other in the process. A reduced gap leads to better organization of the multimodal embedding space which in turn leads to stronger retrieval.

Since `TARA` works well across different base MLLMs, it opens up the possibility of incorporating complementary advances across the LLM space (reasoning models, test-time optimization, mixture of experts, *etc.*). While training on text alone is a strength of `TARA`, investigating ways of utilizing videos (and other modalities) in the train set to obtain a truly universal encoder remains an interesting avenue for future research. Likewise, exploring ways of using `TARA` in a two-stage (retrieval-and-reranking) setup should further boost performance.

**Acknowledgments.** This research is funded by the EPSRC Programme Grant VisualAI EP/T028572/1, and a Royal Society Research Professorship RSRP\R \241003. We are also grateful for funding from Toshiba Corporation. We thank Yuta Shirakawa and Stephan Liwicki from Toshiba for comments and suggestions. We thank Ashish Thandavan for support with infrastructure.

# Bibliography


[1] Alhamoud, K., Alshammari, S., Tian, Y., Li, G., Torr, P.H., Kim, Y., Ghassemi, M.: Vision-language models do not understand negation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 2, 3, 4, 9

[2] Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing Moments in Video with Natural Language. In: IEEE/CVF International Conference on Computer Vision (2017) 3

[3] Bagad, P., Tapaswi, M., Snoek, C.G.: Test of Time: Instilling Video-Language Models with a Sense of Time. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 7

[4] Bagad, P., Zisserman, A.: Chirality in Action: Time-Aware Video Representation Learning by Latent Straightening. In: Conference on Neural Information Processing Systems (2025) 2, 3, 4, 6, 8, 13

[5] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 8

[6] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 4, 8

[7] Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., et al.: Perception encoder: The best visual embeddings are not at the output of the network. arXiv preprint arXiv:2504.13181 (2025) 8

[8] Buch, S., Eyzaguirre, C., Gaidon, A., Wu, J., Fei-Fei, L., Niebles, J.C.: Revisiting the" Video" in Video-Language Understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 3, 4

[9] Chen, A., Doughty, H., Li, X., Snoek, C.G.: Beyond coarse-grained matching in video-text retrieval. In: Asian Conference on Computer Vision (2024) 3, 4

[10] Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Annual Meeting of the Association for Computational Linguistics (2011) 3

[11] Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? Conference on Neural Information Processing Systems (2024) 3

[12] Chen, S., Zhao, Y., Jin, Q., Wu, Q.: Fine-grained video-text retrieval with hierarchical graph reasoning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020) 4

[13] Cores, D., Dorkenwald, M., Mucientes, M., Snoek, C.G., Asano, Y.M.: Lost in time: A new temporal benchmark for videollms. arXiv preprint arXiv:2410.07752 (2024) 3

[14] Cui, X., Cheng, J., Chen, H.y., Shukla, S.N., Awasthi, A., Pan, X., Ahuja, C., Mishra, S.K., Guo, Q., Lim, S.N., et al.: Think then embed: Generative context improves multimodal embedding. arXiv preprint arXiv:2510.05014 (2025) 10

[15] Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., et al.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 4

[16] Doughty, H., Laptev, I., Mayol-Cuevas, W., Damen, D.: Action modifiers: Learning from adverbs in instructional videos. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020) 4
[17] Doughty, H., Snoek, C.G.: How do you do it? fine-grained action understanding with pseudo-adverbs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 4
[18] Du, Y., Liu, Y., Jin, Q.: Reversed in time: A novel temporal-emphasized benchmark for cross-modal video-text retrieval. In: ACM International Conference on Multimedia (2024) 2, 3, 4, 8
[19] Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821 (2021) 3, 4, 6
[20] Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the World in 3,000 Hours of Egocentric Video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 6
[21] Guo, Z., Li, M., Zhang, Y., Long, D., Xie, P., Chu, X.: Towards universal video retrieval: Generalizing video embedding via synthesized multimodal pyramid curriculum. arXiv preprint arXiv:2510.27571 (2025) 4, 8
[22] Hong, S., Kim, J., You, J., Choi, S., Kwak, S., Cho, H.: Textme: Bridging unseen modalities through text descriptions. arXiv preprint arXiv:2602.03098 (2026) 4
[23] Huang, D.A., Ramanathan, V., Mahajan, D., Torresani, L., Paluri, M., Fei-Fei, L., Niebles, J.C.: What makes a video a video: Analyzing temporal information in video understanding models and datasets. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 3
[24] Jiang, T., Huang, S., Luan, Z., Wang, D., Zhuang, F.: Scaling sentence embeddings with large language models. In: Conference on Empirical Methods in Natural Language Processing (2024) 5
[25] Jiang, T., Song, M., Zhang, Z., Huang, H., Deng, W., Sun, F., Zhang, Q., Wang, D., Zhuang, F.: E5-v: Universal embeddings with multimodal large language models. arXiv preprint arXiv:2407.12580 (2024) 2, 4, 5, 6, 8, 12
[26] Jiang, Z., Meng, R., Yang, X., Yavuz, S., Zhou, Y., Chen, W.: Vlm2vec: Training vision-language models for massive multimodal embedding tasks. arXiv preprint arXiv:2410.05160 (2024) 3
[27] Jose, C., Moutakanni, T., Kang, D., Baldassarre, F., Darcet, T., Xu, H., Li, D., Szafraniec, M., Ramamonjisoa, M., Oquab, M., et al.: Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 8
[28] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: A Large Video Database for Human Motion Recognition. In: IEEE/CVF International Conference on Computer Vision (2011) 3
[29] Lei, J., Berg, T., Bansal, M.: Revealing single frame bias for video-and-language learning. In: Annual Meeting of the Association for Computational Linguistics (2023) 3
[30] Lei, J., Berg, T.L., Bansal, M.: Revealing Single Frame Bias for Video-and-Language Learning. arXiv:2206.03428 (2022) 4
[31] Li, B., Zhang, Y., Wang, X., Liang, W., Schmidt, L., Yeung-Levy, S.: Closing the modality gap for mixed modality search. arXiv preprint arXiv:2507.19054 (2025) 4, 13
[32] Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A Comprehensive Multi-modal Video Understanding

Benchmark. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 3
[33] Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., et al.: Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720 (2026) 3, 4, 7, 8, 10, 12, 13
[34] Li, W., Fan, H., Wong, Y., Kankanhalli, M., Yang, Y.: Topa: Extending large language models for video understanding via text-only pre-alignment. Conference on Neural Information Processing Systems (2024) 4
[35] Li, W., Fan, H., Wong, Y., Yang, Y., Kankanhalli, M.S.: Improving context understanding in multimodal large language models via multimodal composition learning. In: International Conference on Machine Learning (2024) 4
[36] Liang, V.W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.Y.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Conference on Neural Information Processing Systems (2022) 3, 4, 12
[37] Lin, S.C., Lee, C., Shoeybi, M., Lin, J., Catanzaro, B., Ping, W.: Mm-embed: Universal multimodal retrieval with multimodal llms. arXiv preprint arXiv:2411.02571 (2024) 4
[38] Liu, Y., Zhang, Y., Cai, J., Jiang, X., Hu, Y., Yao, J., Wang, Y., Xie, W.: Lamra: Large multimodal model as your advanced retrieval assistant. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 3, 4, 6, 8, 10, 12
[39] Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., Ji, R.: X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In: ACM International Conference on Multimedia (2022) 4
[40] Meng, R., Jiang, Z., Liu, Y., Su, M., Yang, X., Fu, Y., Qin, C., Chen, Z., Xu, R., Xiong, C., et al.: Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. arXiv preprint arXiv:2507.04590 (2025) 3, 4, 8, 10, 12
[41] Momeni, L., Caron, M., Nagrani, A., Zisserman, A., Schmid, C.: Verbs in Action: Improving Verb Understanding in Video-Language Models. In: IEEE/CVF International Conference on Computer Vision (2023) 4
[42] Papyan, V., Han, X., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. Proceedings of the National Academy of Sciences (2020) 13
[43] Patraucean, V., Smaira, L., Gupta, A., Recasens, A., Markeeva, L., Banarse, D., Koppula, S., Malinowski, M., Yang, Y., Doersch, C., et al.: Perception test: A diagnostic benchmark for multimodal video models. Conference on Neural Information Processing Systems (2023) 3
[44] Pickup, L.C., Pan, Z., Wei, D., Shih, Y., Zhang, C., Zisserman, A., Scholkopf, B., Freeman, W.T.: Seeing the Arrow of Time. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2014) 12
[45] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning Transferable Visual Models from Natural Language Supervision. In: International Conference on Machine Learning (2021) 8
[46] Saravanan, D., Gupta, V., Singh, D., Khan, Z., Gandhi, V., Tapaswi, M.: Velociti: Benchmarking video-language compositional reasoning with strict entailment. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 3
[47] Schrodi, S., Hoffmann, D.T., Argus, M., Fischer, V., Brox, T.: Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive

vision-language models. In: International Conference on Learning Representations (2025) 3, 4

[48] Singh S, D., Khan, Z., Tapaswi, M.: Figclip: Fine-grained clip adaptation via densely annotated videos. arXiv preprint arXiv:2401.07669v1 (2024) 4

[49] Soomro, K., Zamir, A., Shah, M.: UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. arXiv:1212.0402 (2012) 3

[50] Ventura, L., Yang, A., Schmid, C., Varol, G.: Covr-2: Automatic data construction for composed video retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024) 9

[51] Ventura, L., Yang, A., Schmid, C., Varol, G.: Covr: Learning composed video retrieval from web video captions. In: AAAI Conference on Artificial Intelligence (2024) 2, 3, 4, 6, 9

[52] Wang, J., Yuan, L., Zhang, Y., Sun, H.: Tarsier: Recipes for training and evaluating large video description models. arXiv preprint arXiv:2407.00634 (2024) 3, 4, 7, 8

[53] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 7, 8

[54] Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning. PMLR (2020) 13, 15

[55] Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling Foundation Models for Multimodal Video Understanding. In: European Conference on Computer Vision (2024) 8

[56] Wang, Y., Li, X., Yan, Z., He, Y., Yu, J., Zeng, X., Wang, C., Ma, C., Huang, H., Gao, J., et al.: Internvideo2. 5: Empowering video mllms with long and rich context modeling. arXiv preprint arXiv:2501.12386 (2025) 4

[57] Wang, Z., Chen, A., Hu, F., Li, X.: Learn to understand negation in video retrieval. In: ACM International Conference on Multimedia (2022) 4

[58] Wei, C., Chen, Y., Chen, H., Hu, H., Zhang, G., Fu, J., Ritter, A., Chen, W.: Uniir: Training and benchmarking universal multimodal information retrievers. In: European Conference on Computer Vision. Springer (2024) 4

[59] Wray, M., Larlus, D., Csurka, G., Damen, D.: Fine-grained action retrieval through multiple parts-of-speech embeddings. In: IEEE/CVF International Conference on Computer Vision (2019) 4

[60] Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021) 3

[61] Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2016) 3

[62] Xu, Y., Li, X., Yang, Y., Huang, R., Wang, L.: Fine-grained video-text retrieval: A new benchmark and method. arXiv preprint arXiv:2501.00513v1 (2024) 3, 4, 6, 8, 12, 13

[63] Xue, Z., Luo, M., Grauman, K.: Seeing the Arrow of Time in Large Multimodal Models. arXiv preprint arXiv:2506.03340 (2025) 3, 4, 8

[64] Yaras, C., Chen, S., Wang, P., Qu, Q.: Explaining and mitigating the modality gap in contrastive multimodal learning. In: Chen, B., Liu, S., Pilanci, M., Su, W., Sulam, J., Wang, Y., Zhu, Z. (eds.) Conference on Parsimony and Learning. PMLR (2025) 13

[65] Yin, Y., Zhao, Y., Zhang, Y., Zhang, Y., Lin, K., Wang, J., Tao, X., Wan, P., Zhang, W., Zhao, F.: SEA: Supervised embedding alignment for token-level visual-textual integration in MLLMs. In: Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2025) 4
[66] Yu, X., Ding, P., Zhang, W., Huang, S., Gao, S., Qin, C., Wu, K., Fan, Z., Qiao, Z., Wang, D.: Unicorn: Text-only data synthesis for vision language model training. arXiv preprint arXiv:2503.22655 (2025) 4, 13
[67] Yu, X., Xin, Y., Zhang, W., Liu, C., Zhao, H., Hu, X., Yu, X., Qiao, Z., Tang, H., Yang, X., et al.: Modality gap-driven subspace alignment training paradigm for multimodal large language models. arXiv preprint arXiv:2602.07026 (2026) 4
[68] Yuan, L., Wang, J., Sun, H., Zhang, Y., Lin, Y.: Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. arXiv preprint arXiv:2501.07888 (2025) 7, 8
[69] Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: International Conference on Learning Representations (2023) 4
[70] Zhang, X., Zhang, Y., Xie, W., Li, M., Dai, Z., Long, D., Xie, P., Zhang, M., Li, W., Zhang, M.: Gme: Improving universal multimodal retrieval by multimodal llms. arXiv preprint arXiv:2412.16855 (2024) 4, 10
[71] Zhang, Y., HaoChen, J.Z., Huang, S.C., Wang, K.C., Zou, J., Yeung, S.: Diagnosing and rectifying vision models using language. In: International Conference on Learning Representations (2023) 4, 13
[72] Zhang, Y., Sui, E., Yeung-Levy, S.: Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. In: International Conference on Learning Representations (2024) 4, 13
[73] Zhang, Y., Sui, E., Yeung-Levy, S.: Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. In: International Conference on Learning Representations (2024) 13, 14
[74] Zhao, P., Luan, R., Zhang, W., Wu, P., He, S.: Guiding cross-modal representations with mllm priors via preference alignment. In: Conference on Neural Information Processing Systems (2025) 4
[75] Zhou, J., Liu, Z., Xiao, S., Zhao, B., Xiong, Y.: Vista: Visualized text embedding for universal multi-modal retrieval. arXiv preprint arXiv:2406.04292 (2024) 4
[76] Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025) 4, 7, 8
[77] Zohar, O., Wang, X., Dubois, Y., Mehta, N., Xiao, T., Hansen-Estruch, P., Yu, L., Wang, X., Juefei-Xu, F., Zhang, N., et al.: Apollo: An exploration of video understanding in large multimodal models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 3