TARA: Simple and Efficient Time Aware Retrieval Adaptation of MLLMs for Video Understanding

Piyush Bagad
University of Oxford

Andrew Zisserman
University of Oxford

Abstract

Our objective is to build a general time-aware video-text embedding model for retrieval. To that end, we propose a simple and efficient recipe, dubbed TARA (Time Aware Retrieval Adaptation), to adapt Multimodal LLMs (MLLMs) to a time-aware video-text embedding model without using any video data at all. For evaluating time-awareness in retrieval, we propose a new benchmark with temporally opposite (chiral) actions as hard negatives and curated splits for chiral and non-chiral actions. We show that TARA outperforms all existing video-text models on this chiral benchmark while also achieving strong results on standard benchmarks. Furthermore, we discover additional benefits of TARA beyond time-awareness: (i) TARA embeddings are negation-aware as shown in NegBench benchmark that evaluates negation in video retrieval, (ii) TARA achieves state of the art performance on verb and adverb understanding in videos. Overall, TARA yields a strong, versatile, time-aware video-text embedding model with state of the art zero-shot performance.

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, it is also equally important to gauge how well a retrieval system performs across various aspects of the video such as objects, scene context, motion and temporal dynamics. These aspects can be coarsely categorized as: *static* properties (objects, scene, etc) and *dynamic* properties (motion, visual change, etc). It is well established that most video-text models suffer from *static biases*, *i.e.*, they tend to focus disproportionally on static properties [10, 16, 44, 83, 92] – a problem they inherit from the large-scale video-text training datasets [7, 11, 56]. This static bias is also present in most retrieval benchmarks [3, 13, 80]. In this work, our objective is to develop video-text models that focus equally on dynamic properties,

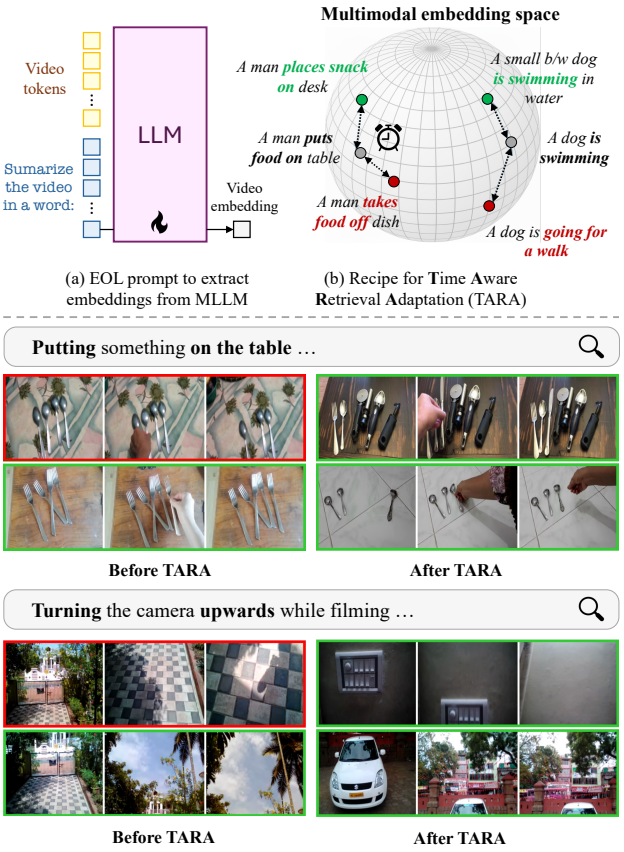

Figure 1. (a) MLLMs ($\mathcal{M}$) can be prompted to output a video embedding using Explicit One-word Limitation prompt [36]. (b) Given that $\mathcal{M}$ projects video/text into a common space, we adapt it contrastively solely on text triplets. By including time-aware triplets (shown with a clock), we achieve strong zero-shot retrieval particularly on time-sensitive queries. Below we show retrieval results for two queries where time ordering is important. Best viewed by zooming in.

in other words, make them more *time-aware*.

But what is time-awareness in the context of video retrieval? First, we are interested in queries that involve some sort of temporal description. Second, 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. For accurate retrieval, a video-text model needs to output embeddings that encode how things change over time in a video (going up/down) rather than only focusing on the spatial context (person, ladder). Recent prior work [5] refers to such temporally antonymous action pairs as *chiral actions*, and we adopt that term here. While [5] studies pure video embedding models that can distinguish such actions, we extend this notion to study *chiral text-to-video retrieval* in a zero-shot setting.

Multimodal Large Language Models (MLLMs) are now dominating other methods on visual question-answering [17] and captioning [69], recently MLLMs have also been adapted for retrieval by carefully extracting embeddings from the last layer [36, 55, 81]. Since MLLMs ingest tokens across multiple frames together, we expect the LLM to be able to model fine-grained temporal dependencies and possibly encode the visual change we care about. Thus, we follow this line of work to build a time-aware retrieval model. Jiang et al. [36] demonstrated that by using a prompt that encourages the model to summarize an image into a single embedding and training contrastively on text alone achieves strong performance on image-text retrieval tasks. We build on this idea and adapt a strong base model (*e.g.*, Tarsier-7B [69]) for video-text retrieval by *training on text alone*. We devise a simple automatic procedure to generate time-aware hard negative sentences. These are composed with standard text samples and the model is trained with contrastive loss where two similar sentences are pulled closer and (temporally) dissimilar ones are pushed apart. We call this recipe Time-Aware Retrieval Adaptation (TARA). TARA is simple, intuitive and efficient – on 8 RTX A6000 GPUs, it takes less than an hour to train.

Does this result in time-sensitivity? Through evaluation on multiple benchmarks (chiral actions [5], RTime [24], verb-adverb recognition [22, 57]), we establish that TARA results in strong time-aware embeddings outperforming all competing models *without training on video data*. Furthermore, TARA also demonstrates remarkable performance gains on several tasks beyond time-awareness. It shows superior understanding of *negation* in queries even beating models fine-tuned for negation. It also shows strong zero-shot ability to recognize temporal parts of speech like verbs and adverbs. Finally, we also check TARA's performance on the standard 10 video retrieval and classification datasets that are part of MMEB-v2 benchmark [55]. It not only retains the performance of its base model, in fact it boosts it to beat all competing zero-shot models.

2. Related Work

Time awareness in video benchmarks. Early video understanding focused on action recognition with datasets like UCF [65], HMDB [42] and retrieval with MSRVTT [80], DiDeMo [2]. The dominance of MLLMs has prompted a suite of benchmarks for question-answering (QA) [47, 58, 78] and captioning [12, 69, 81]. However, the community has repeatedly discovered that most of these do not actually test for time; a single frame or an orderless set of frames can solve them [10, 14, 16, 32, 45, 83, 93]. Most de-facto video retrieval datasets like MSRVTT [80] or MSVD [13] also face this issue. Meta-benchmarks like MMEB-v1/v2 [37, 55] also likely inherit this issue as they are comprised of the same datasets. Recent efforts [16, 63, 83] aim to address this issue for video QA tasks. Similar time-aware benchmarks for retrieval are rare [24, 76]. We build on a recent dataset based on *chiral actions* (temporally opposite actions) [5] and repurpose it for retrieval using text descriptions of actions. Unlike Du et al. [24], this dataset is not built artificially by reversing the arrow of time of videos but instead mines existing datasets (SSv2 [29], EPIC [19], Charades [64]) for chiral actions. Our benchmark helps quantify time-awareness in video retrieval models.

Time-awareness in video retrieval models. Time has been creatively used as a source of self-supervision: space-time jigsaw [41], time arrow [77], time order [28, 84], speed [8], tracking [33, 66], contrasting temporal views [20, 59, 62], cycle consistency in time [25] or explicitly modeling temporal dynamics [15, 34, 86]. But these are usually pure video models without attachment to text. For video retrieval, early methods [7, 49, 56, 79] explored dual encoder models trained contrastively on large-scale datasets [7, 30, 56]. The generalizability of CLIP [60] prompted a deluge of work on adapting it for videos [54]. However, most of these methods do not explicitly model time [82]. Even if there is explicit temporal modeling [53, 70, 73], the resulting embeddings are not necessarily time-aware (as we shall show), perhaps, due to (i) training objectives [85], (ii) deficient text encoders [39, 40] or (iii) static-biased training datasets [7, 11]. This has pushed the community towards exploring highly performant MLLMs for retrieval tasks.

Adapting MLLMs for retrieval. We have witnessed a staggering rise in the abilities of open MLLMs on image [6, 21, 90] and video tasks [6, 69, 75]. 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 [36, 50, 55, 87, 88]. However, as also reflected

in benchmarks for MLLMs like MMEB-v1/v2 [55], the focus is still on images and static-biased video understanding. There is some work on video retrieval, *e.g.* [51, 55, 81] to fine-tune MLLMs on a combination of video-text datasets and text-only datasets. However, much like video-only datasets [10, 44], these training datasets also suffer from focusing on static-biases over temporal awareness. In contrast to prior work, we achieve stronger time-sensitivity by text-only fine-tuning with augmented time-aware samples in the train dataset [27]. We include a thorough review of work on extracting embeddings from MLLMs in Sec. 3.1.

3. TARA: Time Aware Retrieval Adaptation

Our goal is to build a video-text model $\mathbf{F}(\mathbf{v}, \mathbf{t})$ that computes a similarity score between video $\mathbf{v}$ and text description $\mathbf{t}$. We can use $\mathbf{F}$ for retrieval (or classification) by ranking similarity scores between query $\mathbf{q}$ (*e.g.*, text query) and candidates $\mathbf{c}_n, \forall n$ (*e.g.*, gallery of videos).

Instead of using a separate encoder for each modality as in CLIP, we follow a recent line of work [36, 55, 81, 87] by embedding both the video and text under the same model, an MLLM $\mathcal{M}$. Let $f_{\mathcal{M}}(.)$ denote a function to extract an embedding out of $\mathcal{M}$. Then,

$$\mathbf{F}(\mathbf{v}, \mathbf{t}) := f_{\mathcal{M}}(\mathbf{v})^T . f_{\mathcal{M}}(\mathbf{t}). \quad (1)$$

While $\mathcal{M}$ can take any combination of video/text, it is only trained to generate text. Hence, the challenge is two-fold: (i) design $f_{\mathcal{M}}$, and (ii) fine-tune $\mathcal{M}$ such that $f_{\mathcal{M}}$ outputs a time-aware embedding.

In the following, we first review background literature on training (M)LLMs on text alone for retrieval across modalities in Sec. 3.1. Then, in Sec. 3.2, we describe how we adapt this idea to obtain time-aware video embeddings given a carefully chosen training set, and in Sec. 3.3 we describe how we construct such a training dataset with time-aware text samples.

3.1. Review: extracting embeddings from LLMs

Jiang et al. [35] showed that by passing a careful prompt to an LLM (*e.g.*, "`This sentence: [text] means in one word:`", where `[text]` is a placeholder for the sentence) one can extract single token sentence embeddings. The idea is to encourage the LLM to condense the semantic meaning of the sentence into the hidden state of the next token, which is used as the sentence embedding. This is termed an 'Explicit One-word Limitation' (EOL) prompt. In our notation, this EOL extraction process represents $f_{\mathcal{M}}$, and $\mathcal{M}$ is the LLM in this case. To achieve this single token sentence embedding capability, $\mathcal{M}$ is fine-tuned with a Direct Preference Optimization (DPO) [61] inspired objective where similar sentence pairs are preferred over dissimilar sentence pairs.

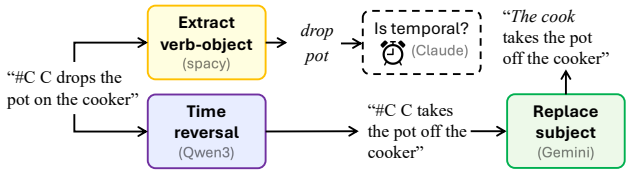

Figure 2. **Pipeline to extract time-aware hard negatives.** Given a caption from Ego4D, we extract verb-object to verify if it is *chiral*. If so, we prompt an LLM to generate a time-aware hard negative and replace the anonymized subject with a realistic one.

More recently, E5-V by Jiang et al. [36] extended this idea to embed images, texts or their combination using a separate EOL prompt for each modality. They show that such EOL prompts dissolve the modality gap [48] between image and text embedding spaces. This enables them to obtain one token embedding for images, *i.e.* $f_{\mathcal{M}}(\text{image})$, while training $\mathcal{M}$ solely on text samples. Unlike [35] that used RL to train the LLM on text pairs, [36] used simpler contrastive learning on text triplets.

Formally, for a triplet $(\mathbf{t}_i, \mathbf{t}_i^+, \mathbf{t}_i^-)$ where $\mathbf{t}_i^+$ is a positive match for the anchor $\mathbf{t}_i$, and $\mathbf{t}_i^-$ a negative match.

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

where $\langle ., . \rangle$ denotes cosine similarity. In [36], the model is trained on the NLI dataset [27] with 275K sentence triplets with "entailment" pairs as positives and "contradiction" pairs as hard negatives.

3.2. Adapting MLLMs for videos

We extend this idea by [36] to obtain time-aware *video-text* embeddings. Our key insight is that the kinds of positives and hard-negatives used during training will reflect the capabilities of the resulting embeddings. For example, if the the positive sentence shares synonymous nouns with the anchor while the hard negative has antonymous nouns, then the resulting embeddings would be good at distinguishing samples by the noun objects they contain. Thus, we investigate the following question: **Can we engineer triplets such that the resulting embeddings are time-sensitive?** So, for example, they are able to distinguish chiral action pairs, as in [5].

To this end, we construct a small text training dataset such that the positives share the actions with the anchor, whilst the negatives have temporally opposite actions. Note, recent and concurrent work by Xu et al. [81] and Liu et al. [51] also use the EOL prompt idea for fine-tuning with text-only data for video tasks. We differ in using a much smaller dataset augmented with time-aware triplets.

3.3. Time aware text dataset construction

Text triplets in NLI [27] include an anchor, a positive (by entailment) and a hard negative (by contradiction). Consider the first example shown in Tab. 1. The positive shares the part "child walking on concrete ledge" with the anchor while the negative differs on "crawling on sandy ledge". Imagine these were captions of three videos. A single frame showing the depicted situations would suffice in recognizing the positive as being closer to the anchor with no need for temporal understanding.

To address this static bias, we design the following process to generate triplets with *temporal* hard-negatives. For this, we use the captions in the Ego4D dataset [30] as our starting point, since it has a large, diverse collection of actions. It includes everyday actions, *e.g.* dropping something or picking something up, that feature the chiral verbs that we care about. In total, Ego4D has 5.3M clips with 1.6M unique short text captions. We proceed in two steps: first, mining the chiral verbs from the Ego4D captions to obtain anchors and positives, and then generating the time-aware hard negatives.

1. Mining chiral verbs in Ego4D. First, we generate a list of chiral action verbs by prompting the Claude-4 LLM [4] to generate as many chiral pairs as possible. This generates a set of 532 chiral pairs. Then, we mine the Ego4D captions for any that contain these chiral verbs: we use the `spacy` library to extract the main verb and object in each caption sentence. Then, we select a subset of captions that feature a chiral verb (*e.g.*, *opening/closing*, *pick-up/put-down*, *etc.*) based on the generated set. For each anchor caption that consists of verb-object pair (v, o), all other captions that feature a similar verb-object pair serve as the positives. Below, we explain how we automatically generate hard-negatives.

2. Generating time-aware hard negatives. An overview of how time-sensitive hard negatives are constructed is shown in Fig. 2. For each anchor caption with a chiral verb, because there exists an antonym verb, we can use the linguistic skills of LLMs to generate the corresponding antonym caption. Specifically, we prompt Qwen3-1.7B to generate a temporal antonym caption for the given caption while retaining the rest of the context. This generated sentence serves as our time-aware hard negative.

3. Subject replacement. Since Ego4D captions are anonymized, *e.g.* "#C C" denotes camera wearer, we replace it with a plausible subject (*e.g.*, "A man") by prompting an LLM. Note that we make sure that the subject generally remains the same for a given triplet.

A few samples are given Tab. 1. See Supplemental for more details and examples. We sample n=9000 static-biased triplets from NLI and n=1000 triplets from our time-aware Ego4D subset. The time-aware subset comprises v=35 chiral verb pairs.

$\mathbf{t}$ (Anchor)	$\mathbf{t}^+$ (Positive)	$\mathbf{t}^-$ (Hard negative)
NLI: 9,000 triplets (90%)		
A little child walking on a concrete ledge.	A small child takes steps along the concrete.	A little girl crawls on her hands and knees along a sandy ledge.
A small black and white dog is swimming in water.	A dog is swimming.	A dog is going for a walk.
A young couple kissing by a bike rake.	There are people showing affection.	The people are sleeping.
Ego4D time-aware samples: 1,000 triplets (10%)		
A man puts the food on the dish	A man puts a snack in the pan	A man takes the food off the dish
The lady closes the container with its cover.	The lady closes the container on the table.	The lady opens the container with its cover.
The bartender puts the bottle down	The bartender puts bottle on sink top	The bartender picks up the bottle
The student removes her left hand from the book on the table.	The woman removes her hand from the wool.	The student places her left hand on the book on the table.

Table 1. **Samples of text triplets used in training.** While samples from NLI focus on static bias (corresponding videos can be distinguished by single frame), Ego4D samples focus on temporal actions. Words marked in blue represent synonymous parts while those in red represent the antonymous parts of the sentence. More samples are given in the supplementary.

Composition of static and temporal triplets. For training we use triplets from the NLI [27] dataset together with those constructed from Ego4D. We intend the triplets from the NLI to provide the model with static understanding while those from our Ego4D time-aware triplets to provide better temporal understanding. Sample triplets from NLI and our time-aware ones are shown in Tab. 1. Notice how the hard negatives in NLI are visually so dissimilar that one can distinguish them with a single frame or without any temporal modeling. In contrast, by using chiral verbs in Ego4D, our time-aware triplets need the model to ignore the spatial context and focus on the temporal context. We find that using only 9K samples from NLI augmented with 1K samples from Ego4D produces a strong time-aware model.

4. Chirality in Action: Retrieval Benchmark

Recently, Bagad and Zisserman [5] proposed a benchmark, *Chirality in Action* (CiA), to probe time-sensitivity of video embeddings using chiral actions, *i.e.*, actions that are temporally opposite in nature such as *"opening" vs. "closing door"* or *"folding"* vs *"unfolding paper"*. Distinguishing between such actions requires the video to encode temporal change in a video. However, the CiA evaluation is based on *classification* by training linear probes for each chiral

action pair. In contrast, we repurpose the dataset for zero-shot video-to-text and text-to-video *retrieval*. We call this the *CiA-Retreival* benchmark. While it is similar in spirit to the "RTime" [24], we do not artificially reverse the arrow of time. Thus, all our videos are physically plausible.

Datasets. We use the same three datasets: SSv2, EPIC and Charades from [5]. SSv2 has 1430 videos (16 chiral pairs), EPIC has 3108 videos (66 pairs) and Charades has 5498 videos (28 pairs). Examples retrieval scenarios are shown in Fig. 1 with samples sourced from SSv2. Note the datasets cover both ego- and exo-centric camera viewpoints.

Splits. As usual, we consider both $t \rightarrow v$ and $v \rightarrow t$ settings. For each, we have three splits: (i) *chiral*: the gallery consists of only temporally opposite samples, (ii) *Non-chiral*: consists of all samples except the chiral opposites, and (iii) *All*: consists of all samples. (i) is the most time-sensitive setting while (ii) is least time-sensitive and (iii) balances both.

Metrics. For $v \rightarrow t$ setting, we consider $R@1$ as the primary metric since we need to measure if the best matching text is selected for the given video. For $t \rightarrow v$, an example in the chiral setup is: given "opening door", one needs to rank videos in the gallery consisting of videos of opening or closing door. In such a case, we care about the overall ranking and not just the top matched video. Thus, for $t \rightarrow v$, we consider mAP as the primary metric.

5. Experiments

We evaluate the TARA model on a diverse set of retrieval and classification tasks in a zero-shot manner, *i.e.*, the model is trained once on our NLI dataset augmented with time-aware samples and then evaluated on all downstream tasks (without fine-tuning). First, in line with our main objective, we test the time-awareness of TARA in Sec. 5.1. We also include thorough ablations on our design choices. Second, in Sec. 5.2, we demonstrate that TARA shows a strong understanding of negation in queries outperforming models fine-tuned for this task. Third, in Sec. 5.3, we show that TARA also exhibits impressive understanding of verbs and adverbs in videos. Finally, in Sec. 5.4, we also test TARA on some standard benchmarks (video tasks in MMEB-v2, standard retrieval and classification tasks).

Implementation details. By default, we use Tarsier-7B [69] as the base model and fine-tune it on the dataset detailed in Sec. 3.3. In the ablation study, we also experiment with other base models and show TARA's generality. 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, if not stated otherwise, we use $F{=}16$ uniformly spaced frames in the video.

5.1. Time-awareness

Brief takeaway: TARA produces time-sensitive video-text embeddings as demonstrated by strong zero-shot performance on two benchmarks: our proposed *CiA-Retrieval* [5] and *Reversed in Time* [24].

CiA-Retrieval. We evaluate TARA's time-awareness on our benchmark introduced in Sec. 4. In Tab. 2, we report results on the *Chiral* split and *All* split which includes both chiral and non-chiral samples in the gallery. We do not show results on *Non-chiral* split to save space but it is included in the Supplemental. We compare with (a) dual-encoder models like CLIP, (b) directly extracting the last token embeddings out of MLLMs, (c) other retrieval adaptation recipes like CaRe [81]. Across all datasets and splits, TARA comprehensively outperforms all models, especially in the most time-sensitive setting of *Chiral* retrieval. On average, TARA beats the second best model (CaRe) on 'chiral' split by **+17.2** points and on 'all' split by **+13.0** points while being trained on just **4%** of the data used by CaRe.

Reversed in Time. On the recent RTime [24] benchmark, TARA also outperforms all competing models (even those fine-tuned on RTime) in retrieving forward/reverse videos for a given caption. Results are tabulated in Tab. 3.

Ablation study. For ablation experiments, we report in Tab. 4 the metrics for chiral splits in SSv2 as the base model and training data are varied. We make the following remarks: (i) Tarsier-7B is the strongest base model but TARA improves all base models substantially. (ii) With Tarsier fixed, we vary the dataset to show that including time-aware (🕓) Ego4D samples helps (see rows 2, 4). Here, for fair comparison in the NLI-only setting, we replace the 1K rows from Ego4D with 1K samples from NLI. (iii) Replacing anonymized subjects with realistic subjects in Ego4D captions also helps (see rows 3, 4).

Video embeddings on CiA. We also compare the video embeddings from TARA with those from strong video encoders of the probing benchmark proposed in [5]. TARA video embeddings achieve new state-of-the-art performance beating the LiFT embedding (developed in [5]) by +7.1 points.

Method	🗁 (M)	🎞	SSv2 $t\to v$ Chiral	SSv2 $t\to v$ All	SSv2 $v\to t$ Chiral	SSv2 $v\to t$ All	EPIC $t\to v$ Chiral	EPIC $t\to v$ All	EPIC $v\to t$ Chiral	EPIC $v\to t$ All	Charades $t\to v$ Chiral	Charades $t\to v$ All	Charades $v\to t$ Chiral	Charades $v\to t$ All
Chance	-	-	50.0	3.1	50.0	3.1	50.0	3.1	50.0	3.1	50.0	3.1	50.0	3.1
Dual encoder models														
CLIP (avg.) [60]	-	-	52.0	12.7	52.1	5.9	51.0	7.0	54.1	5.0	48.4	6.5	51.5	5.5
DINO.txt [38]	-	-	52.1	13.1	52.3	5.5	50.6	6.0	53.5	8.3	50.7	10.1	51.5	7.6
XCLIP [53]	-	-	54.7	16.8	52.4	5.3	49.1	4.5	53.0	2.4	49.0	7.6	51.2	3.9
ViCLIP [73]	-	-	50.8	16.2	51.4	6.2	51.4	7.9	54.0	5.1	49.5	8.8	51.2	6.8
Perception Enc. [9]	-	-	50.1	17.2	51.8	7.4	48.5	1.8	53.9	2.8	51.3	7.2	52.0	4.8
InternVideo 2 [74]	-	-	52.5	20.6	51.6	10.9	48.3	8.8	53.9	9.6	50.7	11.9	51.8	10.0
MLLMs zero-shot														
Qwen2VL-7B [71]	-	-	60.2	17.3	58.4	9.0	53.7	9.6	55.3	7.7	55.9	8.1	53.7	6.5
Qwen2.5VL-7B [6]	-	-	67.6	20.6	63.7	12.6	55.4	9.5	56.9	6.4	55.8	11.1	54.5	7.4
MLLMs with fine-tuning														
VLM2Vec-V2 [55]	1.7	✓	58.8	15.9	55.8	8.9	49.4	12.9	53.8	8.7	53.5	10.5	53.2	8.1
LAMRA [51]	1.4	✗	55.3	7.8	54.2	10.3	53.7	9.0	13.2	7.9	52.1	11.3	53.0	9.1
GVE-7B [31]	13.0	✓	53.4	4.0	52.5	7.3	54.7	7.3	53.8	4.6	54.2	10.2	51.9	4.1
E5-V [36]	0.3	✗	52.6	14.7	51.2	5.3	57.1	6.5	53.8	3.1	48.9	7.1	50.9	4.4
ArrowRL [83]	0.02	✓	67.5	22.5	66.4	14.3	55.7	9.6	57.5	6.3	57.1	12.2	56.0	8.1
CaRe [81]	0.3	✗	66.4	23.7	63.9	23.8	62.3	16.9	58.3	14.3	56.1	12.9	52.8	9.8
TARA (Ours)	0.01	✗	**85.1**	**47.8**	**84.0**	**32.8**	**77.3**	**30.6**	**76.8**	**19.7**	**71.8**	**29.9**	**68.3**	**18.8**

Table 2. **Results on CiA-Retrieval.** Across all datasets, TARA outperforms all MLLMs fine-tuned for retrieval. For $t\to v$, we report mAP and for $v\to t$, we report R@1. 'Chiral' denotes retrieving from a gallery of temporally antonymous samples while 'All' denotes a gallery that includes chiral and non-chiral samples. 🗁 denotes size of the training dataset (in millions) and 🎞 denotes whether or not videos were used during retrieval adaptation.

Method	RTime (Binary) Accuracy T2V	RTime (Binary) Accuracy V2T
Zero-shot		
Singularity [44]	48.7	49.9
Internvideo2-1B	50.0	51.0
Qwen2VL	56.3	62.3
Qwen2.5VL	53.4	66.6
Tarsier	64.9	65.0
Tarsier + TARA	**71.6**	**71.3**
Fine-tuned on RTime		
CLIP4Clip [52]	49.8	49.8
UMT [46]	51.2	51.3
UMT-Neg [24]	54.5	54.2
ArrowR-Qwen2 [83]	57.1	68.8
ArrowRL-Qwen2.5 [83]	55.6	69.6

Table 3. **Evaluation on ReversedInTime** proposed by Du et al. [24]. This is loosely similar to our 'Chiral' retrieval setting but a negative video is obtained by reversing the arrow of time of a positive video. TARA outperforms even models fine-tuned on RTime.

5.2. Negation understanding

Brief takeaway: TARA shows strong negation understanding on *NegBench* even outperforming fine-tuned models.

Base model	Fine-tuning dataset	Chiral $v\to t$	Chiral $t\to v$
Tariser-7B	-	76.4	73.9
Tarsier-7B	NLI	81.8	78.5
Tarsier-7B	NLI + 🕓 Ego4D w/o subj.	84.1	82.9
Tarsier-7B	NLI + 🕓 Ego4D	**85.1**	**84.0**
Qwen2VL-7B	-	60.2	58.4
Qwen2VL-7B	NLI + 🕓 Ego4D	70.1	72.7
InternVL2-8B	-	58.5	61.1
InternVL2-8B	NLI + 🕓 Ego4D	66.9	67.6
CaRe-S1	-	52.2	53.0
CaRe-S2	NLI-275K	66.4	63.9
CaRe-S1	NLI + 🕓 Ego4D	71.8	72.7

Table 4. **Ablation study.** We report accuracy on the chiral retrieval split of SSv2 as we vary the base MLLM and training data. Here, NLI denotes the subset of 9K samples that we use. 🕓-Ego4D denotes 1K time-aware samples drawn from Ego4D. In row 3, "w/o subj." denotes Ego4D samples where we do not replace the anonymized subjects with realistic subject values.

Recently, Alhamoud et al. [1] showed that de-facto vision-language models do not understand negation in queries while retrieving images/videos. On their proposed

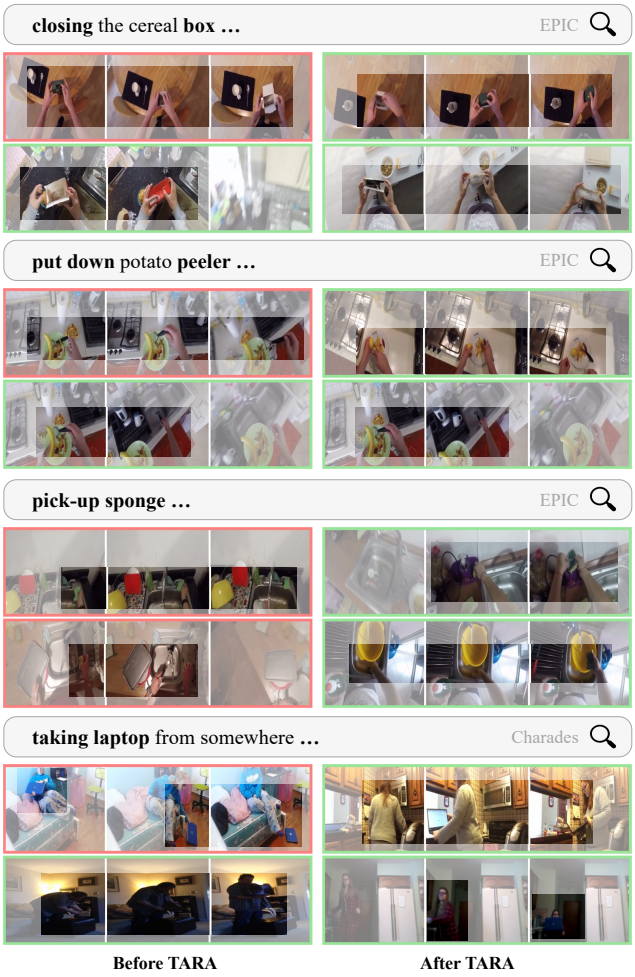

Figure 3. **Qualitative results.** We show qualitative retrieval results for various queries with base MLLM (Tarsier-7B) **before** (left) and **after** (right) TARA fine-tuning. Since it is hard to see key details, we highlight the part of the video that depicts the desired action. TARA improves understanding of chiral actions where one needs to distinguish between similar looking temporally opposite action videos. Kindly zoom in to view details clearly.

NegBench benchmark, we evaluate TARA on both text-to-image (COCO) and text-to-video (MSRVTT) retrieval with-/without negation in queries. As shown in Tab. 6, on both datasets, TARA zero-shot outperforms even models specifically fine-tuned for negation understanding.

5.3. Verb and adverb sensitivity

Brief takeaway: TARA shows strong a zero-shot ability to recognize verbs and adverbs which often require time-sensitive understanding (*e.g.*, in distinguishing *"walking slowly/quickly"*).

To test verb understanding, we evaluate TARA on the verb-focused subset of Kinetics proposed by Momeni et al. [57] and the verb-noun annotations in EPIC-Kitchens [19]. Given a video, the task is to choose the correct verb phrase from a given set of choices. Results are given in Tab. 7. TARA zero-shot outperforms all its competitors, particularly VFC [57] which is trained with hard-negative verbs. For adverb understanding, we use the benchmark proposed in Doughty and Snoek [22]. Since adverbs co-occur with verbs, the task is, given a video and an action verb, select the correct adverb between two choices (the correct adverb

	Chiral actions			
Video embedding	**SSv2**	**EPIC**	**Charades**	**Avg.**
DINOv2 (concat.)	79.7	74.1	65.8	73.2
SigLIP2 (concat.)	76.8	74.7	67.8	73.1
VideoMAE	80.3	70.5	59.1	70.0
InternVideo 2.5	80.0	70.9	62.8	71.2
Video JEPA	85.4	70.8	57.1	71.1
Video JEPA 2	78.4	66.1	57.0	67.2
CaRe	85.7	76.7	64.2	75.5
LiFT	86.6	75.5	69.5	77.2
Tarsier 7B + TARA (Ours)	**90.8**	**85.1**	**76.9**	**84.3**

Table 5. **Time-awareness in video embeddings on CiA.** We probe video embeddings from TARA on the CiA benchmark [5]. It beats the previous state of the art: LiFT as well as large video encoders like VJEPA.

Model	Fine-tune data	R@5 (↑)	R-Neg@5 (↑)
(a) **COCO**			
CLIP	None	54.8	48.0
	CC12M	58.8	54.5
	CC12M-NegCap	58.5	57.8
	CC12M-NegFull	54.2	51.9
NegCLIP [1]	None	68.7	64.4
	CC12M	70.2	66.0
	CC12M-NegCap	68.6	67.5
	CC12M-NegFull	69.0	67.0
Tarsier	None	57.4	45.6
Tarsier + TARA	Ours	**72.6**	**68.7**
(b) **MSR-VTT**			
CLIP	None	50.6	45.8
	CC12M	53.7	49.9
	CC12M-NegCap	54.1	53.5
	CC12M-NegFull	46.9	43.9
NegCLIP [1]	None	53.7	51.0
	CC12M	56.4	52.6
	CC12M-NegCap	56.5	54.6
	CC12M-NegFull	54	51.5
Tarsier	None	55.7	49.7
Tarsier + TARA	Ours	**69.0**	**68.7**

Table 6. **NegBench Evaluation** checks understanding of negation in queries. TARA (zero-shot) beats strong baselines in [1].

Method	Kinetics-Verbs	EPIC (Verb + Noun)
Chance	1.0	0.03
CLIP [60]	67.4	1.8
DINO.txt [38]	65.8	1.4
VFC [57]	57.1	-
CaRe [81]	72.2	3.6
TARA (Ours)	**73.0**	**6.1**

Table 7. **Verb recognition.** TARA outperforms all competing methods on recognizing verb phrases zero-shot on Kinetics-Verbs [57] and EPIC [19].

and its antonym). For example, given a video of a person "walking", one needs to select if the walking is "slow/fast". We encode the video together with the action verb in a single prompt. We provide the full prompt used in the supplemental. Likewise, for text, we represent the verb-adverb by including them together in a sentence. See supplemental for details. Results are shown in Tab. 8. TARA zero-shot exceeds the semi-supervised model in [22].

Method	Fine-tuned on	VATEX	MSRVTT
Chance	-	50.0	50.0
Action Modifiers [23]	H2M+VTX(20%)	64.2	-
AM w/ pseudo-labels [22]	H2M+VTX(20%)	67.5	65.0
AM w/ pseudo-labels [22]	. . . + MSRVTT	67.5	70.5
TARA (Ours)	-	**73.2**	**75.3**

Table 8. **Adverb recognition.** TARA exceeds semi-supervised baselines trained specially on adverb recognition. H2M denotes adverb subset of HowTo100M, VTX that of VATEX. Here, 20% is the % of labelled data used while the rest of the dataset uses pseudo-labels [22].

5.4. Standard benchmarks

Brief takeaway: On the video classification and retrieval subsets of the MMEB-v2 benchmark, TARA outperforms all competing models.

We also evaluate on standard video benchmarks for classification and retrieval. We use the Massive Multimodal Embedding Benchmark (MMEB-v2) [55]. For retrieval, MMEB-v2 is made up of subsets of 5 datasets: MSRVTT [80], MSVD [13], DiDeMo [3], YouCook2 [89] and VATEX [72] with 9,421 videos in total. For classification, subsets of Kinetics-700 [11], SSv2 [29], HMDB [42], UCF [65] and Breakfast [43] with a total video count of 4,433 are used. The results are tabulated in Tab. 9. We compare with top multimodal embedding models [55, 87] including recently proposed thinking-augmented retrieval [18] and training-free prompting based retrieval [91]. TARA outperforms all competing methods. [18, 91] are complementary approaches to TARA and can be combined with TARA to improve performance at test time.

Model # of Datasets →	Classification 5	Retrieval 5
ColPali v1.3 (3B) [26]	26.7	21.6
GME (7B) [87]	37.4	28.4
LamRA-Qwen2.5 (7B) [51]	32.9	23.2
VLM2Vec-Qwen2VL (7B) [37]	39.1	29.0
VLM2Vec-V2 (7B) [55]	45.9	27.6
Think-Then-Embed-7B [18]	57.5	38.0
FreeRet (Qwen2VL-7B) [91]	63.2	39.3
TARA (Tarsier-7B)	**63.7**	**43.1**

Table 9. **Evaluation on the video tasks in MMEB-v2 [55].** TARA beats all competing zero-shot methods on both tasks.

6. Conclusion and Discussion

In this paper, we propose a simple and efficient recipe, TARA, to adapt a Multimodal LLM for time-sensitive video retrieval. Building on prior work on image-text retrieval, TARA is trained on text triplets alone with a contrastive objective. The composition of the text triplets is engineered to have time-aware samples as well as static-biased samples. Time-aware text triplets are automatically generated based on a subset of captions in Ego4D. TARA is trained on only 10K samples in less than an hour on 8 RTX A6000 GPUs. As a benchmark for time-sensitivity in video retrieval, we repurpose the *Chirality in Action* (CiA) dataset by [5].

TARA achieves state-of-the-art results on time-aware video retrieval on CiA as well as on an existing benchmark, *Reversed in Time* [24]. Beyond time-awareness, TARA shows strong understanding of negation in queries as measured on NegBench [1] and also demonstrates solid understanding of verbs-adverbs [22, 57]. TARA also outperforms all competing models on standard video retrieval and classification tasks as measured on MMEB-v2 [55].

It is an intriguing finding of this paper that instilling time-awareness into a model benefits multiple other retrieval tasks that are seemingly time-unaware. It is evidence that training sets and training methods would benefit from more time-sensitive data, and that 'a little time-awareness goes a long way'. Indeed, 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 including 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 Research. We thank Ashish Thandavan for support with infrastructure.

References

[1] Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29612–29622, 2025. 6, 7, 8, 4

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 2

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing Moments in Video with Natural Language. In *ICCV*, 2017. 1, 8

[4] Anthropic. Claude 4 system card: Claude opus 4 and claude sonnet 4. https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf, 2025. Accessed: 2025-11-13. 4

[5] Piyush Bagad and Andrew Zisserman. Chirality in action: Time-aware video representation learning by latent straightening. *arXiv preprint arXiv:2509.08502*, 2025. 2, 3, 4, 5, 7, 8

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 6

[7] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-end Retrieval. In *ICCV*, 2021. 1, 2

[8] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the Speediness in Videos. In *CVPR*, 2020. 2

[9] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. 6

[10] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" Video" in Video-Language Understanding. In *CVPR*, 2022. 1, 2, 3

[11] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, 2017. 1, 2, 8

[12] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 2

[13] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 1, 2, 8

[14] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024. 2

[15] Siyi Chen, Minkyu Choi, Zesen Zhao, Kuan Han, Qing Qu, and Zhongming Liu. Unfolding Videos Dynamics Via Taylor Expansion. *arXiv preprint arXiv:2409.02371*, 2024. 2

[16] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Lost in time: A new temporal benchmark for videollms. *arXiv preprint arXiv:2410.07752*, 2024. 1, 2

[17] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Tvbench: Redesigning video-language evaluation. *Arxiv*, 2024. 2

[18] Xuanming Cui, Jianpeng Cheng, Hong-you Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Kumar Mishra, Qi Guo, Ser-Nam Lim, et al. Think then embed: Generative context improves multimodal embedding. *arXiv preprint arXiv:2510.05014*, 2025. 8

[19] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling Egocentric Vision: The EPIC-Kitchens Dataset. In *ECCV*, 2018. 2, 7, 8, 5

[20] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. TCLR: Temporal Contrastive Learning for Video Representation. *Computer Vision and Image Understanding*, 2022. 2

[21] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104, 2025. 2

[22] Hazel Doughty and Cees GM Snoek. How do you do it? fine-grained action understanding with pseudo-adverbs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13832–13842, 2022. 2, 7, 8, 5

[23] Hazel Doughty, Ivan Laptev, Walterio Mayol-Cuevas, and Dima Damen. Action modifiers: Learning from adverbs in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 868–878, 2020. 8

[24] Yang Du, Yuqi Liu, and Qin Jin. Reversed in time: A novel temporal-emphasized benchmark for cross-modal video-text retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5260–5269, 2024. 2, 5, 6, 8

[25] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal Cycle-Consistency Learning. In *CVPR*, 2019. 2

[26] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024. 8

[27] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021. 3, 4

[28] Amir Ghodrati, Efstratios Gavves, and Cees GM Snoek. Video Time: Properties, Encoders and Evaluation. *arXiv preprint arXiv:1807.06980*, 2018. 2

[29] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *ICCV*, 2017. 2, 8, 5

[30] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *CVPR*, 2022. 2, 4

[31] Zhuoning Guo, Mingxin Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Xiaowen Chu. Towards universal video retrieval: Generalizing video embedding via synthesized multimodal pyramid curriculum. *arXiv preprint arXiv:2510.27571*, 2025. 6

[32] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7366–7375, 2018. 2

[33] Allan Jabri, Andrew Owens, and Alexei Efros. Space-Time Correspondence as a Contrastive Random Walk. *NeurIPS*, 2020. 2

[34] Dinesh Jayaraman and Kristen Grauman. Slow and Steady Feature Analysis: Higher Order Temporal Coherence in Video. In *CVPR*, 2016. 2

[35] Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196, 2024. 3

[36] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024. 1, 2, 3, 6, 5

[37] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024. 2, 8

[38] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24905–24916, 2025. 6, 8

[39] Kumara Kahatapitiya, Anurag Arnab, Arsha Nagrani, and Michael S Ryoo. Victr: Video-conditioned text representations for activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18547–18558, 2024. 2

[40] Amita Kamath, Jack Hessel, and Kai-Wei Chang. Text encoders bottleneck compositionality in contrastive vision-language models. *arXiv preprint arXiv:2305.14897*, 2023. 2

[41] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised Video Representation Learning with Space-Time Cubic Puzzles. In *AAAI*, 2019. 2

[42] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: A Large Video Database for Human Motion Recognition. In *ICCV*, 2011. 2, 8

[43] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014. 8

[44] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing Single Frame Bias for Video-and-Language Learning. *arXiv:2206.03428*, 2022. 1, 3, 6

[45] Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–507, 2023. 2

[46] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19948–19960, 2023. 6

[47] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A Comprehensive Multi-modal Video Understanding Benchmark. In *CVPR*, 2024. 2

[48] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 3, 5

[49] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 2

[50] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*, 2024. 2

[51] Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4015–4025, 2025. 3, 6, 8

[52] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 6

[53] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pages 638–647, 2022. 2, 6

[54] Neelu Madan, Andreas Møgelmose, Rajat Modi, Yogesh S Rawat, and Thomas B Moeslund. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*, 2024. 2

[55] Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025. 2, 3, 6, 8

[56] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 1, 2

[57] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in Action: Improving Verb Understanding in Video-Language Models. In *ICCV*, 2023. 2, 7, 8, 4

[58] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023. 2

[59] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal Contrastive Video Representation Learning. In *CVPR*, 2021. 2

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, 2021. 2, 6, 8

[61] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 3

[62] Adria Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Pătrăucean, Florent Altché, Michal Valko, et al. Broaden Your Views for Self-supervised Video Learning. In *ICCV*, 2021. 2

[63] Darshana Saravanan, Varun Gupta, Darshan Singh, Zeeshan Khan, Vineet Gandhi, and Makarand Tapaswi. Velociti: Benchmarking video-language compositional reasoning with strict entailment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18914–18924, 2025. 2

[64] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *ECCV*, 2016. 2, 5

[65] K Soomro. UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. *arXiv:1212.0402*, 2012. 2, 8

[66] Shashanka Venkataramanan, Mamshad Nayeem Rizve, João Carreira, Yuki M Asano, and Yannis Avrithis. Is ImageNet Worth 1 Video? Learning Strong Image Encoders from 1 Long Unlabelled Video. *arXiv preprint arXiv:2310.08584*, 2023. 2

[67] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5270–5279, 2024. 2, 3

[68] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr-2: Automatic data construction for composed video retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[69] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024. 2, 5

[70] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2

[71] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6

[72] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591, 2019. 8

[73] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 2, 6

[74] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling Foundation Models for Multimodal Video Understanding. In *ECCV*, 2024. 6

[75] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025. 2

[76] Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. Paxion: Patching action knowledge in video-language foundation models. *Advances in Neural Information Processing Systems*, 36:20729–20749, 2023. 2

[77] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and Using the Arrow of Time. In *CVPR*, 2018. 2

[78] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 2

[79] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 2

[80] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 1, 2, 8

[81] Yifan Xu, Xinhao Li, Yichun Yang, Rui Huang, and Limin Wang. Fine-grained video-text retrieval: A new benchmark and method. *arXiv e-prints*, pages arXiv–2501, 2024. 2, 3, 5, 6, 8

[82] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. CLIP-VIP: Adapting Pre-trained Image-Text Model to Video-Language Representation Alignment. *arXiv preprint arXiv:2209.06430*, 2022. 2

[83] Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the Arrow of Time in Large Multimodal Models. *arXiv preprint arXiv:2506.03340*, 2025. 1, 2, 6

[84] Charig Yang, Weidi Xie, and Andrew Zisserman. Made to Order: Discovering Monotonic Temporal Changes Via Self-supervised Video Ordering. *arXiv preprint arXiv:2404.16828*, 2024. 2

[85] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022. 2

[86] Heng Zhang, Daqing Liu, Qi Zheng, and Bing Su. Modeling Video as Stochastic Processes for Fine-grained Video Representation Learning. In *CVPR*, 2023. 2

[87] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024. 2, 3, 8

[88] Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. Vista: Visualized text embedding for universal multi-modal retrieval. *arXiv preprint arXiv:2406.04292*, 2024. 2

[89] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 8

[90] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 2

[91] Yuhan Zhu, Xiangyu Zeng, Chenting Wang, Xinhao Li, Yicheng Xu, Ziang Yan, Yi Wang, and Limin Wang. Freeret: Mllms as training-free retrievers. *arXiv preprint arXiv:2509.24621*, 2025. 8

[92] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An Exploration of Video Understanding in Large Multimodal Models. *arXiv preprint arXiv:2412.10360*, 2024. 1

[93] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18891–18901, 2025. 2

TARA: Simple and Efficient Time Aware Retrieval Adaptation of MLLMs for Video Understanding

Supplementary Material

Table of Contents

A. Data Processing

In this section, we describe in more detail the processing used to obtain the training dataset, *i.e.*, generating temporal antonym sentences (Sec. A.1), replacing anonymous subjects with realistic subjects in Ego4D captions (Sec. A.2) and some example triplets in the training dataset composed on NLI and Ego4D (Sec. A.3).

A.1. Generating (temporal) antonym sentence.

We prompt the Qwen3-1.7B LLM model to generate temporally antonymous sentences for Ego4D captions. We use some in-context samples to guide the model. The exact prompt with few in-context samples is provided below. Based on our initial filtering of chiral verbs, we use this stage to get antonyms for 425K sentences.

An alternative to using an LLM for this stage is to simply detect and replace the chiral verb phrase. But we observe that in a lot of cases the antonym involves more than just replacing a chiral verb, *e.g.*, "pushing something from left to right" should become "pushing something from right to left". Using an LLM instead of hard-coded rules ensures against errors in such cases.

Prompt for temporal antonym generation.

You are a helpful assistant expert at natural language understanding and grammatical nuance.

You are given a caption.

Your task is to generate a temporally antonymous version of the caption.
You should retain the broader context of the caption but only change the action described in the caption as if the video is temporally reversed.

In case where the verb phrase in the caption may not have a temporal antonym, you should return the None as the output. Never return the same caption as the output.

Here are some examples:

(1) Caption: #C C unrolls the yarn from her left index finger
Output: #C C rolls the yarn onto her left index finger

(2) Caption: #C C folds the cloth
Output: #C C unfolds the cloth

(3) Caption: #C C puts the pan on the stove
Output: #C C takes the pan off the stove

(4) Caption: Someone is walking on the street
Output: None

(5) Caption: #C C checks the cloth
Output: None

Output in a JSON format 'caption_forward': ..., 'caption_reverse': ... where caption_forward is the original caption and caption_reverse is the temporally reversed caption.

A.2. Subject replacement

Since Ego4D captions are anonymized, *e.g.* "#C C" denotes camera wearer, to make the sentences more realistic, we replace it with a plausible subject (*e.g.*, "A man") by prompting Gemini 2.5 Flash Lite. We found that using Qwen3-VL does not produce sufficient diversity and it often copies the subjects from the provided in-context samples. We provide

one in-context example for reference. The detailed prompt is provided below.

Prompt for subject replacement.

You are an expert in English comprehension and writing.

Given three sentences where the subjects may be anonymized, your task is to fill the placeholders for subjects with realistic subjects.

For example, given these sentences,

S1: #C C Puts down a serving spoon and chop sticks on a cooking pot
S2: #C C puts a spoon in a bowl.
S3: #C C Picks up a serving spoon and chop sticks from a cooking pot

a valid response could be something like:

S1: The chef puts down a serving spoon and chop sticks on a cooking pot
S2: The chef puts a spoon in a bowl.
S3: The chef picks up a serving spoon and chop sticks from a cooking pot

This is only an example, think logically what subject would best fit the given description and situation. For example, you will not find a cook doing carpentary. In case you are not sure, you can use generic subject pronouns like 'The man' or 'The person' or 'The lady', or use proper nouns like name of a person etc. Do not just use the template examples, you can be slightly creative. Make sure it is the same subject in all three sentences.

Test input:

A.3. More example triplets

More example triplets from NLI and time-aware triplets from Ego4D are provided in Tab. 10. The Ego4D subset has 35 unique chiral verbs (*e.g.*, 'opening') and 203 verb phrases (*e.g.*, 'opening box').

$\mathbf{t}$ (Anchor)	$\mathbf{t}^+$ (Positive)	$\mathbf{t}^-$ (Hard negative)
NLI (90%)		
A group of performers sing a song.	The performers are singing.	The performers are painting pictures.
Man falling off a blue surfboard in the ocean.	Man is outside.	The man is at church.
A man doing a back flip while another takes a picture.	A man doing a back flip	A man is laying on a tarp
Hillary cannot be there!	Hillary is not allowed there	Hillary must be there.
A man in a green shirt is on the computer . . .	A man on a computer.	A man watching tv.
Ego4D time-aware samples (10%)		
The mechanic closes the tool box	The mechanic closes the box	The mechanic opens the tool box
The doorman opens door	The doorman opens the door.	The doorman closes door
The cook lifts the bowl of ingredient . . .	The cook lifts the bowl of fruits.	The cook places the bowl of ingredient . . .
The gardener uproots the weeds with her hand	The gardener uproots weeds	The gardener plants the weeds with her hand
The carpenter places her left hand on the plank	The carpenter places his hand on the table	The carpenter removes her left hand from the plank
The person turns off the tap	The person turns off the tap with her right hand	The person turns on the tap

Table 10. **More samples of text triplets used in training.** While samples from NLI focus on static bias (corresponding videos can be distinguished by a single frame), Ego4D samples focus on temporal actions. Words marked in blue represent synonymous parts while those in red represent the antonymous parts of the sentence.

B. Additional Experiments

In this section, we present results on some additional experiments. First, we present results on all splits across all three datasets in CiA [5] for completeness. Second, we discuss composed video retrieval using TARA in a zero-shot manner. Then, we present ablation on the size and composition of the fine-tuning dataset used for TARA. Finally, we specify additional details of the evaluation tasks.

B.1. Full results on CiA-Retrieval

In the main paper, we omit the *Non-chiral* split results to save space. In Tab. 11, we provide full results across all three datasets. The rows marked in green denote TARA fine-tuning with different settings. TARA achieves best results across all datasets including on *Non-chiral* splits.

B.2. Composed video retrieval

WebVid-CoVR [67]. Since TARA is based on extracting embeddings out of an MLLM, it inherits the flexibility of an MLLM to take as input composition of video and text together. We leverage this and evaluate on the task of composed video retrieval introduced by Ventura et al. [67]. 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 and construct an EOL prompt for joint video-text inputs:

Model	$v \rightarrow t$ (R@1)			$t \rightarrow v$ (mAP)		
	Chiral	*Non-chiral*	*All*	*Chiral*	*Non-chiral*	*All*
Chance	50.0	6.7	3.1	50.0	6.7	3.1
SSv2						
Perception Enc.	50.1	32.7	17.2	51.8	14.6	7.4
InternVideo2	52.5	35.7	20.6	51.6	21.8	10.9
Qwen2VL-7B	60.2	28.0	17.3	58.4	14.2	9.0
Qwen2.5VL-7B	67.6	31.5	20.6	63.7	18.3	12.6
CaRe (Stage 1)	52.2	31.3	17.7	53.0	14.5	7.7
CaRe (Stage 2)	66.4	46.2	28.7	63.9	37.7	23.8
Qwen2.5VL-ArrowRL	67.5	33.8	22.5	66.4	19.5	14.3
Qwen2VL-7B + TARA	70.1	39.6	27.4	72.7	26.8	20.5
Tarsier-7B + TARA (✗ Ego4D)	81.8	**60.0**	46.7	78.5	38.7	30.0
Tarsier-7B + TARA (✗ subj.)	84.3	58.5	47.3	82.9	39.1	31.2
Tarsier-7B + TARA	**85.1**	58.9	**47.8**	**84.0**	**41.0**	**32.8**
EPIC						
Perception Enc.	51.3	12.8	7.2	52.0	9.4	4.8
InternVideo2	50.7	23.7	11.9	51.8	19.2	10.0
Qwen2-VL-7B	53.1	11.2	7.6	57.6	12.2	8.3
Qwen2.5-VL-7B	55.4	12.4	9.5	56.9	10.5	6.4
CaRe (Stage 1)	49.7	11.7	5.1	51.2	5.0	2.5
CaRe (Stage 2)	62.3	25.0	16.9	58.3	22.0	14.3
Qwen2.5VL-ArrowRL	55.7	12.4	9.6	57.5	9.7	6.3
Qwen2VL-7B + TARA	65.0	27.9	20.8	65.1	20.6	16.0
Tarsier-7B + TARA (✗ Ego4D)	69.1	33.1	26.4	66.0	24.9	18.6
Tarsier-7B + TARA (✗ subj.)	76.9	**38.6**	**32.0**	72.4	**28.1**	**22.5**
Tarsier-7B + TARA	**77.3**	37.6	30.6	**76.8**	27.5	20.7
Charades						
Perception Enc.	48.5	4.2	1.8	53.9	5.1	2.8
InternVideo2	48.3	22.1	8.8	53.9	16.3	9.6
Qwen2VL-7B	61.5	16.7	10.7	55.0	13.2	7.8
Qwen2.5VL-7B	55.8	17.0	11.1	54.5	12.0	7.4
CaRe (Stage 1)	54.8	12.7	7.0	56.3	2.9	1.6
CaRe (Stage 2)	56.1	25.2	12.9	52.8	17.9	9.8
Qwen2.5VL-ArrowRL	57.1	18.6	12.2	56.0	12.8	8.1
Qwen2VL-7B + TARA	65.5	31.8	22.7	63.9	22.7	16.2
Tarsier-7B + TARA (✗ Ego4D)	68.5	39.5	27.0	61.4	27.8	18.1
Tarsier-7B + TARA (✗ subj.)	71.4	**41.3**	29.8	65.9	**29.1**	**20.6**
Tarsier-7B + TARA	**71.8**	40.9	**29.9**	**68.3**	28.7	19.8

Table 11. **Full results on CiA-Retrieval.** In the main paper, we omit the *Non-chiral* split results to save space. Here, we provide full results across all three datasets. Rows marked with green denote models fine-tuned with our TARA recipe. Here, (✗ Ego4D) indicates that no Ego4D text triplets were used in training and (✗ subj.) indicates that '#C C' is not replaced by a realistic subject in the Ego4D caption.

```
USER: [video] Edit instruction: [sent]
Imagine the given text edit instruction
applied on the given video. Summarize
the resulting video in one word.
ASSISTANT:
```

We evaluate on WebVid-CoVR [67] test set consisting of 2,556 query-video samples. TARA is evaluated in a zero-shot manner. As shown in Tab. 12, while TARA falls short of beating specialized models fine-tuned on WebVid-CoVR, it does beat all zero-shot competing methods. This is promising and needs further investigation to improve beyond fine-tuned models.

B.3. Ablation on data size and composition

Data size. NLI [27] has 275K text triplets. As shown in the main paper, with only 10K samples, we achieve strong performance on temporally sensitive chiral actions as well as other benchmarks. However, it is natural to ask: does performance improve with more data? To test this, we fix the composition of NLI:Ego4D to be 0.9:0.1 and vary the total number of samples from n=5000, . . . , 200000. We plot the avg. accuracy on the *Chiral* split on SSv2 in Fig. 4. We also plot the total GPU hours (Quadro RTX 8000) alongside the performance. The performance does not change significantly as data size increases. Thus, we pick n=10, 000 that can be trained within an hour with 8 GPUs. Another reason for using low number of samples is that since we are fine-tuning the LLM weights with contrastive loss, we want to ensure that the model does not drift too far away from the base model. Training on a larger dataset reduces loss on this dataset but suffers generalization on other benchmarks.

				WebVid-CoVR-Test			
	Modalities	Backbone	F	R@1	R@5	R@10	R@50
Chance	-	-	-	0.1	0.2	0.4	1.8
Zero-shot							
	T	BLIP	-	19.7	37.1	45.9	65.1
	V	BLIP	15	34.9	59.2	68.0	86.0
	V+T	CLIP	15	44.4	69.1	77.6	93.0
	V+T	BLIP	15	45.5	70.5	79.5	93.3
TARA	V+T	Tarsier	8	50.8	75.6	83.0	96.0
TARA w/ caps.	V+T	Tarsier	15	**53.1**	**78.6**	**86.4**	**97.3**
Fine-tuned							
	T	BLIP	-	23.7	45.9	55.1	77.0
	V	BLIP	15	38.9	65.0	74.0	92.1
	V+T	CLIP	1	50.6	77.1	85.1	96.6
	V+T	BLIP	1	50.6	74.8	83.4	95.5
	V+T	BLIP	1	51.8	78.3	85.8	97.1
Ventura et al. [67]	V+T	BLIP	15	53.1	79.9	86.9	97.7
Ventura et al. [68]	V+T	BLIP2	15	**59.8**	**83.8**	**91.3**	**98.2**

Table 12. **Composed video retrieval on WebVid-CoVR.** TARA beats all other zero-shot methods but there is room for improvement when compared to fine-tuned models. Here, 'TARA w/ caps.' model uses the original image caption as part of the prompt'.

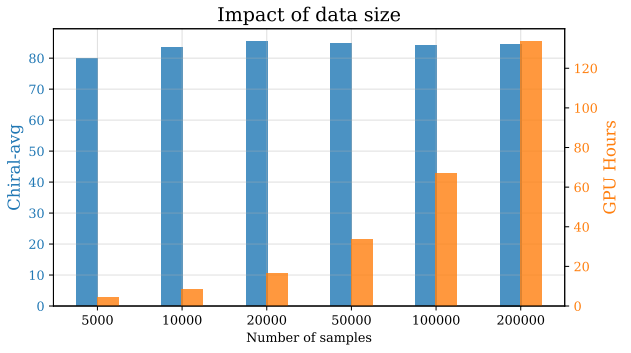

Figure 4. **Ablation on data size.** The data composition of NLI:Ego4D is fixed to 0.9:0.1 and total number of samples is varied. Beyond n=10, 000, the increment in accuracy is not substantial compared to the number of GPUs hours that increase. Left scale corresponds to blue bars showing accuracy, right scale corresponds to orange bars showing GPU hours.

Data composition. For data composition, we fix the total number of samples to n=10,000 but vary the proportion of text samples from Ego4D that is composed with those from NLI. In Fig. 5, we plot the avg. (across $t \rightarrow v, v \rightarrow t$) performance on the chiral split (y-axis) vs the non-chiral split (x-axis) on SSv2. Let $\alpha \in [0, 1]$ be the fraction of Ego4D data used during TARA fine-tuning. α is shown beside each scatter point in Fig. 5. We find that using $0.1 \leq \alpha \leq 0.6$ achieves best trade-off. The models corresponding to $\alpha \in \{0.1, 0.2, \ldots, 0.6\}$ perform similarly but note that as α increases, the additional cost (*e.g.*, of replacing subjects with an LLM) increases. Furthermore, using α=0.1 shows slightly better generalization on other datasets (Charades, EPIC as shown in Fig. 6). Thus, we pick α=0.1 to build our fine-tuning dataset.

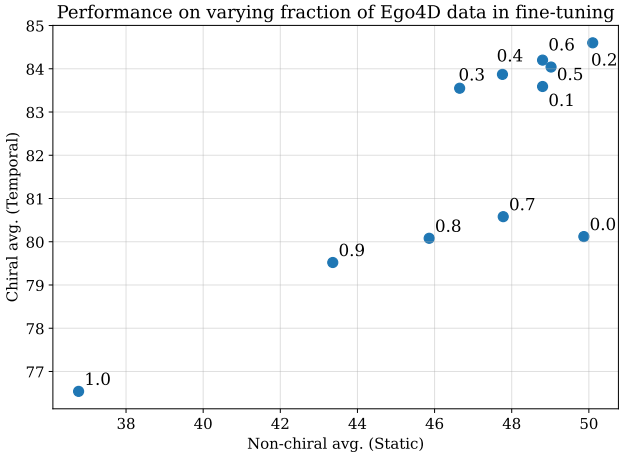

Figure 5. **Ablation on data composition.** This figure plots the chiral (temporal) accuracy (y-axis) *vs*. non-chiral (static) accuracy (x-axis) for different composition data composition. Let $\alpha \in [0, 1]$ be the fraction of Ego4D data used during TARA fine-tuning. α is shown beside each scatter point. We find that using $0.1 \leq \alpha \leq 0.6$ achieves best trade-off. Models corresponding to $\alpha \in \{0.1, 0.2, \ldots, 0.6\}$ vary slightly in performance, but we pick α=0.1 since (i) it generalizes better to other datasets (see Fig. 6), (ii) it incurs low cost (*e.g.*, in using LLM for subject replacement for only n_{Ego4D}=1000 samples).

Statistical significance over multiple runs. To establish statistical significance, we fine-tune with n=10,000 samples with ratio of NLI:Ego4D as 0.9:0.1 multiple times (k=5) with different random seeds. Note that the key change is the training dataset which is sampled randomly each time. We evaluate the trained models on SSv2 and report numbers in Tab. 13. The performance is robust across random seeds with low standard deviation.

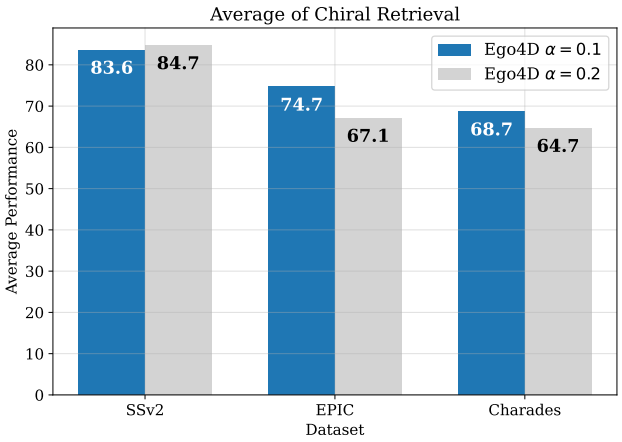

Figure 6. **Comparing Ego4D fraction across datasets.** We compare the avg. chiral accuracy with α=0.1, 0.2 across all three datasets. While α=0.2 outperforms α=0.1 on SSv2, the latter generalizes better to EPIC and Charades.

	$v \rightarrow t$ (R@1)			$t \rightarrow v$ (mAP)		
Model	*Ch.*	*NC*	*All*	*Ch.*	*NC*	*All*
TARA	**84.5**$_{\pm 0.6}$	**59.0**$_{\pm 0.6}$	**47.0**$_{\pm 1.1}$	**83.9**$_{\pm 0.5}$	**39.9**$_{\pm 0.7}$	**32.0**$_{\pm 0.5}$

Table 13. **Statistical significance.** We train TARA with $k = 5$ different seeds and report the avg. and standard deviation on SSv2. *Ch.* denotes chiral and *NC* non-chiral.

B.4. Details of downstream tasks

NegBench evaluates negation in text queries while retrieving images/videos. Alhamoud et al. [1] modify standard captions by including negations, evaluating how models handle queries that specify both present and absent elements. For example, for text-to-image retrieval, if the original caption for an image in COCO dataset is `[Original Caption]`, then it is modified to include negation with non-existing objects: `There is no [X] in the image. [Original Caption]` or `[Original Caption]. There is no [X] in the image`. To introduce linguistic diversity, LLaMA 3.1 is used to paraphrase these captions. COCO has 5,000 images and MSRVTT has 1,000 videos to be retrieved.

Verb understanding. The subset of Kinetics proposed by [57] includes 97 classes that share a common noun with another class, but have a different verb (and therefore action). For example, for noun *hair*, it has classes like: 'braiding hair', 'brushing hair', 'curling hair', *etc*. This includes 4,619 videos in total.

Adverb understanding. For adverb recognition, given a video and an action verb, the task is to choose between an adverb and its antonym. For example, given a video

of "stirring soup", the model needs to decide if it is stirring "slowly" or "quickly". First, we embed the video and the action verb jointly. We do that by slightly modifying the EOL prompt to summarize a vieeo and the action together (see prompt below). Then, we embed the sentence description of the action with either of the two adverbs: `The action [action] is performed [adverb]`. Finally, the similarity is computed in the common embedding space.

```
USER: [video]
Action: This video shows the action [sent]
Look at the video carefully. Summarize
the action in the video in one word:
ASSISTANT:
```

We follow the test splits for MSRVTT and VATEX in Doughty and Snoek [22]. MSRVTT has 1,824 clips with 18 adverb pairs while VATEX has 2,835 clips with 34 adverb pairs.

B.5. Modality gap

Jiang et al. [36] showed that using the EOL prompt dissolves the modality gap [48] between text and *image* representation spaces obtained from MLLMs. We analyze the analogous phenomenon for *video*-text data on 1,000 samples from MSRVTT. We observe that with the base Tarsier [69] model (no fine-tuning), the EOL prompt does reduce the modality gap but it still does not dissolve entirely (see top row in Fig. 7). However, with TARA fine-tuning, we confirm that the EOL prompt does indeed remove the gap (see bottom row in Fig. 7).

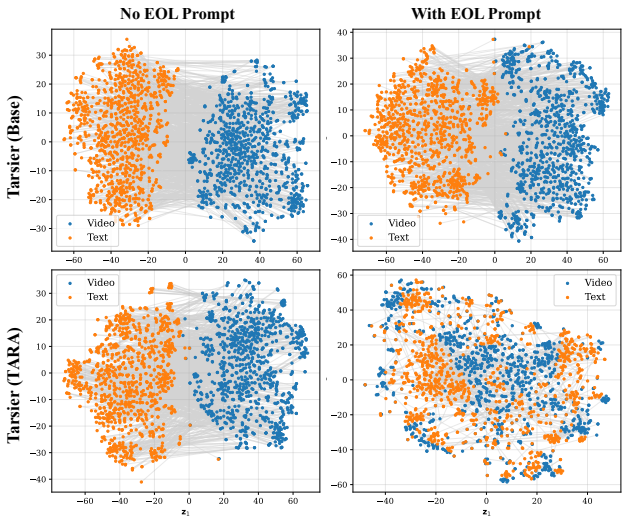

Figure 7. **Modality gap.** We analyse video and text embeddings from the test set of MSRVTT (n=1000 video-caption pairs).

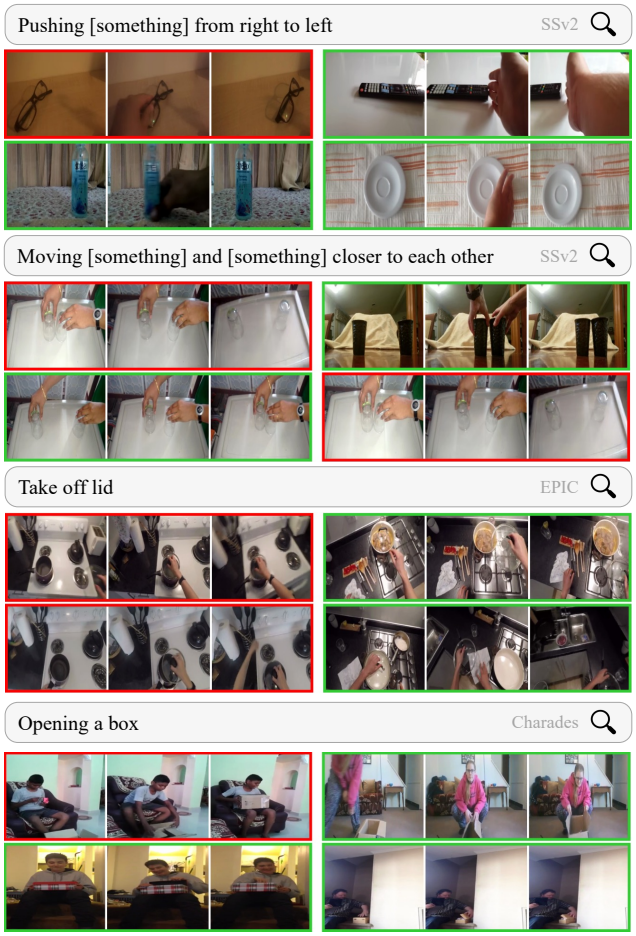

Figure 8. **More qualitative results** from the CiA-Retrieval splits. Left shows top-2 videos retrieved from the base Tarsier model and right shows those from Tarsier adapted with TARA.

C. Additional Qualitative Results

C.1. Retrieval results.

In Fig. 8, we present more retrieval results with the Tarsier model before (left) and after (right) TARA fine-tuning. TARA fine-tuning improves results across all three datasets in CiA.

The top two rows show examples from SSv2 [29] followed by one example each from EPIC [19] and Charades [64]. While the base model often gets confused between similar looking, temporally distinct actions (*e.g.*, "take off lid" *vs*. "put lid down"), TARA retrieves videos temporally consistent with the query. The queries includes various kinds of visual change: change in object positions ("pulling something from right to left"), change in object state ("opening/closing box") or spatial distances ("moving two things closer to each other").

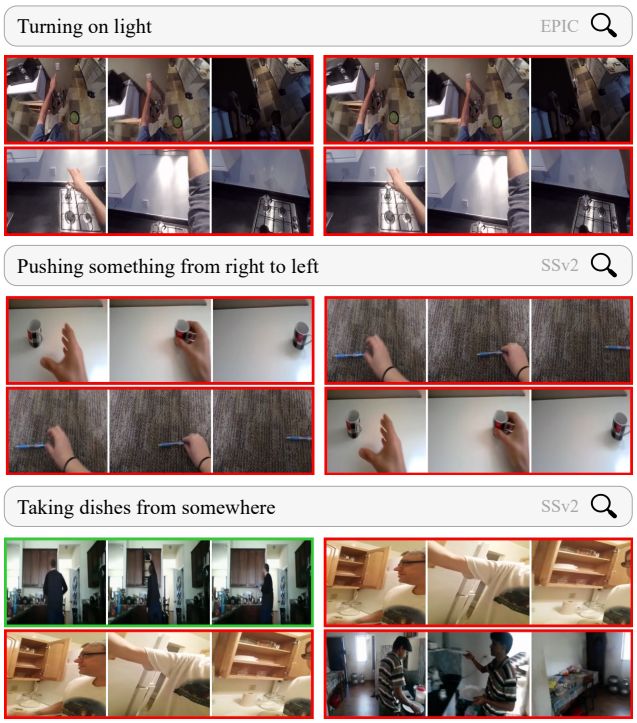

Figure 9. **Some failure cases** of TARA. Left shows top-2 videos retrieved from the base Tarsier model and right shows those from Tarsier adapted with TARA.

C.2. Some failure cases.

There are some cases where TARA fine-tuning does not improve the base model's abilities for time-sensitive retrieval shown in Fig. 9. In case of some samples from EPIC and Charades, for example, see the first and last rows in Fig. 9, the key action goes out of view (*e.g.*, in "turning on light", "taking dishes from somewhere") which contributes to embeddings that are unable to distinguish between such actions.