




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

Thomas Hummel<sup>\*1,2</sup> , Shyamgopal Karthik<sup>\*1,2</sup> , Mariana-Iuliana Georgescu<sup>2</sup>,  
Zeynep Akata<sup>2,3</sup> 

<sup>1</sup>Tübingen AI Center, University of Tübingen

<sup>2</sup>Helmholtz Munich, MCML <sup>3</sup>TU Munich

{thomas.hummel, shyamgopal.karthik}@uni-tuebingen.de

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

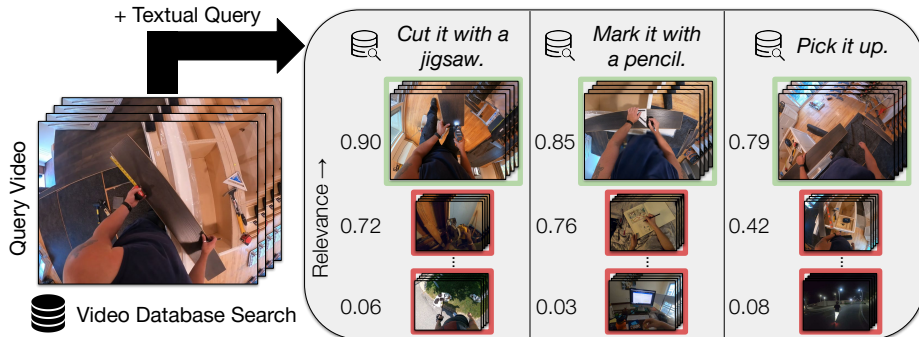
**Keywords:** Video Retrieval · Fine-Grained Video Understanding

## 1 Introduction

Recent advances in Vision-Language Models (VLMs) have enabled video search through free-form textual descriptions. However, expressing complex queries, especially those involving subtle transformations or actions, remains challenging with purely text-based searches. In the image domain, Composed Image Retrieval (CIR) [37, 52, 54] has emerged as a related task where a user provides a reference image and a textual description of the desired modification. In the video domain, the corresponding task is coined as Composed Video Retrieval (CVR) [51] where the aim is to retrieve videos from a database given a reference video and a textual query that describes how the reference video should be modified. For example, a user searching through a long video might provide a short reference clip showing construction work along with a textual description such as “make the person cut with a jigsaw instead” to pinpoint the

---

\* Denotes equal contribution



**Fig. 1:** The goal of the Composed Video Retrieval (CVR) task is to retrieve the correct video using both a query video and a textual video modification instruction that describes the semantic changes required from the query video.

precise video they are looking for (See Figure 1). CVR remains a relatively under-explored area, posing unique challenges due to the added complexity of effectively utilizing the temporal information inherent in videos. CVR is extremely challenging because it requires understanding both the visual and textual inputs and composing them to retrieve the desired video efficiently.

A major step towards tackling the CVR challenge is the introduction of the large-scale WebVid-CoVR training set and a smaller evaluation benchmark. These datasets are automatically collected by using existing video-text datasets looking for pairs that differ only in a single word in the caption, and using a Large Language Model (LLM) [49] to generate the textual instruction. The final training set contains over 1.6M triplets, which is extremely useful for the CVR task. However, the evaluation set quality is quite limited due to the automatic dataset construction. For instance, most of the modifications predominantly focus on the color, shape, and adding/removing objects from the scene that do not require temporal understanding (see Figure 2). Therefore, the task can be tackled with a single image rather than a video, *e.g.* a vision-language model [32] trained on the image level achieves state of the art.

In this work, we propose to create an evaluation set for the Composed Video Retrieval task that requires holistic video understanding to obtain strong performance. To achieve this, we propose **EgoCVR**, a manually curated and high-quality evaluation set with 2,295 videos sourced from the Ego4D [21] dataset. Our **EgoCVR** dataset consists of a query and target clip sourced from the same long video and the textual modifier asking for a subtle change in the action being performed in the clip. As a result, models need to have strong video understanding to be able to achieve strong performance in our evaluation setting.

Furthermore, we evaluate on our new benchmark several methods designed for cross-modal retrieval, consisting of vision-language models such as CLIP [43], BLIP [32], the video-based method LanguageBind [60], as well as the egocentric video model EgoVLP [34, 42], by adapting them to the CVR task. Naively

adapting vision-language models to perform the CVR task, even if the model was finetuned on the large benchmark, does not work well, *e.g.* BLIP<sub>CoVR</sub> finetuned for the CVR task on 1.6M triplets performs poorly on EgoCVR.

To address this shortcoming, we propose to adapt a training-free method proposed for Composed Image Retrieval [28] to the Composed Video Retrieval task. When employed with a generic re-ranking strategy, this approach, which we name TFR-CVR, achieves the best results among all considered methods in various evaluation settings. To summarise, we make the following contributions:

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

## 2 Related Work

**Video-Language Models and Retrieval.** Early work on video retrieval often focused on extending retrieval approaches from images to videos by aggregating image features within a video [14, 41, 48, 57]. However, with the introduction of large-scale video-text datasets [4, 39, 53, 56], and contrastive language-image pre-training [31, 32, 43], there have been several models proposed for the task of video-text retrieval [5, 20, 38, 60]. Due to the growing popularity of egocentric video datasets [21, 22], video-language models have been proposed that specifically focus on this setting [34, 42, 59]. However, while there has been growing interest in developing video-based foundation models [10, 55], these have all been focused on captioning and video-text retrieval. Different from this, we show how existing video-text models can be utilised for fine-grained Composed Video Retrieval.

**Composed Image Retrieval.** The task of Composed Image Retrieval (CIR) has found significant application in conditional search [25, 52, 54], where users perform interactive dialogue to refine a given query image toward retrieving specific items. Classical techniques often employ custom models that project text-image pairs into a common embedding space [1, 7, 11, 12, 29, 52] or use cross-modal attention mechanisms [13]. With the advent of vision-language foundation models [8, 27, 43], interest in CIR has surged, especially in zero-shot settings that avoid the need for task-specific models. Recent works either attempt to train models that avoid the necessity for paired triplets [3, 6, 9, 24, 44, 46] or train models on large datasets that then generalise to a wide variety of scenarios [23, 30, 36, 51]. There have also been several datasets and benchmarks proposed for Composed Image Retrieval including large-scale generic datasets such as CIRR [37], CIRCO [6], as well as fine-grained evaluation benchmarks focusing on fashion [25, 54], fine-grained attributes [50], sketches [19] or birds [17].

In this work, we are inspired by both CIR methods [28, 45] as well as CIR benchmarks [37, 50] in curating a fine-grained Composed Video Retrieval dataset, as well as proposing general methods that can tackle this task.

**Composed Video Retrieval.** To the best of our knowledge, the only existing benchmark available for Composed Video Retrieval is WebVid-CoVR [51]. Further, the only models tailored for it are BLIP models finetuned on the training set of WebVid-CoVR [47, 51]. Concurrent to our work, the task of *video detours* [2] was introduced, which focused on retrieving and localising temporal segments within long videos using free-form textual queries and the query video, specifically for instructional videos. In this work, we propose a fine-grained evaluation benchmark for Composed Video Retrieval with two evaluation settings, along with a training-free method using video-specific models for this task.

### 3 EgoCVR : An Egocentric Benchmark Dataset for Composed Video Retrieval

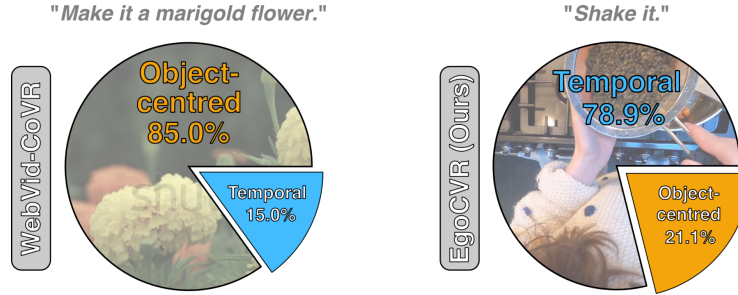
In Section 3.1, we first formally define the task of Composed Video Retrieval, while in Section 3.2, we describe our dataset construction methodology in detail.

#### 3.1 Problem Definition

The Composed Video Retrieval task was first introduced by Ventura *et al.* [51]. Let  $\mathcal{V}$  denote the space of videos and  $\mathcal{T}$  the space of textual instructions. Given a query video  $q_v \in \mathcal{V}$  and textual instruction  $q_t \in \mathcal{T}$ , the goal of composed video retrieval (CVR) is to identify the modified video  $v \in \mathcal{D}$  from a database of videos (gallery)  $\mathcal{D} = \{v_1, \dots, v_n\}$ , where  $n$  is the number of videos in  $\mathcal{D}$ , that most closely represents the semantic modifications described by  $q_t$ . The task can be formalized as a scoring function  $\Phi : \mathcal{V} \times \mathcal{T} \times \mathcal{D} \rightarrow \mathbb{R}$ . This function measures the similarity between the query video  $q_v$ , the modification text  $q_t$ , and each candidate video  $v_i$  in the database,  $0 \leq i \leq n$ . The video with the highest score according to  $\Phi$  is deemed the optimal retrieval result.

The scoring function  $\Phi$  is implemented by representing videos and text within a shared embedding space. We denote the video encoder as  $\Psi_v : \mathcal{V} \rightarrow \mathbb{R}^d$  and the text encoder as  $\Psi_t : \mathcal{T} \rightarrow \mathbb{R}^d$ , where  $d$  is the dimension of the embedding space. The video encoder  $\Psi_v$  processes either single frames (with averaged frame-level embeddings) or frame sequences using a temporal video encoder. The text encoder  $\Psi_t$  embeds the modification instructions into the same space as  $\Psi_v$ . Text and video embeddings are then combined to form a multi-modal video-text embedding  $q_{v,t}$  using a fusion function  $\Psi_q : \{q_v, q_t\} \rightarrow \mathbb{R}^d$ . Candidate videos from the database  $\mathcal{D}$  are also encoded using  $\Psi_v$ . Finally, the cosine similarity is used as a matching score between the query embedding  $q_{v,t}$  and each candidate video embedding  $v_i$ ,  $v_i \in \mathcal{D}$ ,  $0 \leq i \leq n$ .





**Fig. 2:** EgoCVR focuses to a significantly greater extent on temporal and action-related modifications (blue) as opposed to object-centred modifications (orange) when compared to the previously existing WebVid-CoVR-Test benchmark [51].

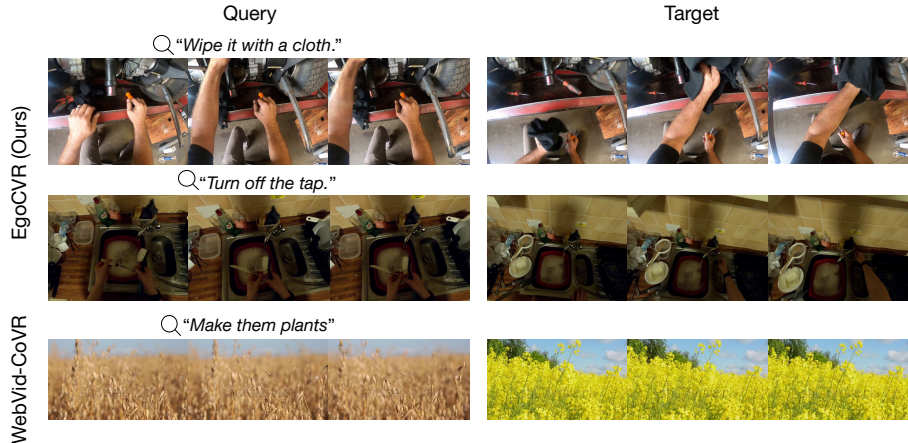
### 3.2 From Egocentric Videos to Composed Video Retrieval

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

The dataset includes 155k narrations, each associated with 2-8 second video clips extracted from 1,250 long-form videos. We reduce the 155k densely annotated clips to 9k distinct clips by automatically filtering out clips with temporal overlap, ensuring a higher likelihood of single, focused actions within each clip. Our dataset annotation process aims to find pairs of videos that have subtle differences. While previous work [51] applied an automated video-matching process by searching for single-word differences in video captions, EgoCVR is created using a careful manual annotation process outlined in the following.

We manually search for possible video pairs within long videos, *i.e.* video pairs originate from the same source video. Creating annotations from the same source video allows for fine-grained comparisons where the primary difference between video pairs is the controlled textual modification. We identify matching pairs through similarity in their narrations, *i.e.* the narrations differ in a single semantic concept like actions (*e.g.* rinsing vs. rubbing) or objects (*e.g.* knife vs. spoon). During annotation, an emphasis was put on creating pairs where temporal modifications are prioritised. We manually disregard annotation pairs through visual inspection when i) the narrations do not accurately describe the clip, ii) the narrated actions or objects are visible only for a fraction of the clip, or iii) the presence of multiple actions would result in ambiguous samples.

<sup>1</sup> The task can be viewed here: [https://ego4d-data.org/docs/tutorials/FHO\\_Overview/](https://ego4d-data.org/docs/tutorials/FHO_Overview/)



**Fig. 3:** Samples consisting of visual and text queries along with the target video from our test set EgoCVR (top two rows) and WebVid-CoVR-Test set [51] (bottom row).

When multiple videos with the same narration are present (*i.e.* “#C C puts down the piece of cloth.” and “#C C puts down the cloth.”), we group the clips together. This allows us to create samples in EgoCVR with multiple ground truth targets, even for narrations that do not perfectly match with an exact textual search. This annotation process resulted in a total of 2,295 queries with an average of 1.2 ground-truth targets per query.

**Creating Textual Video Modification Instructions.** We create textual video modification instructions from the video narrations of paired clips. Ideal modification instructions clearly describe the most prominent change that needs to be applied to the query video to get to the desired target video. We design these modifications to be as concise as possible while still conveying all the relevant information. Instructions in EgoCVR provide only the minimum necessary semantic difference between the query and the target videos. For instance, the instruction “Rinse it instead.” does not provide information on which object to rinse. To create the text modifications, we utilise the reasoning capabilities of LLMs to generate concise instructions that describe the transformation from the provided query clip narration to the target clip narration. As LLM, we employ GPT-4 [40] and provide the LLM with a list of 15 in-context examples [15] together with a clear instruction prompt (more details in the supplementary). We illustrate examples from EgoCVR, as well as how it contrasts with typical samples from WebVid-CoVR in Figure 3.

**Visual Distractors.** We additionally collect distractor video clips for each annotated target video similar to the CIRRR image subsets [37]. We automatically source the distractor clips from the same long-form video provided by the Ego4D FHO task. Our collected distractor clips ensure high visual similarity (*i.e.*

identical camera wearer and scene) and prevent trivial retrieval shortcuts based solely on visual similarity. To obtain the distractor clips, for each target video in EgoCVR, we filter out clips from the Ego4D FHO annotations originating from different long-form videos, clips used as query-target video annotations, and clips depicting the same action as the target. We then rank potential distractor clips by their narration’s CLIP similarity to the target video narration. Finally, we sample up to 6 distractor clips per target video. To represent various semantic similarity levels, we sample one clip from the bottom 10% of similarity scores, four from the middle 80%, and one from the top 10%. We sample a total of 10,522 distractors with an average of 4.2 distractors per target video.

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

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

## 4 Training-Free Re-Ranking Composed Video Retrieval

We adopt several vision-language methods for composed video retrieval. To show that our proposed benchmark focuses on temporal actions, we employ both image processing and video processing models in the evaluation. As image processing models, we employ two widely used image-language models, namely CLIP [43] and BLIP [32]. We also adapt the video-language models EgoVLPv2 [42] and

**LanguageBind** [60] that were specifically designed to learn temporal video representations. EgoVLPv2 was specifically pre-trained on egocentric videos, while LanguageBind aligns various modalities such as video, infrared, depth and audio to a frozen language encoder after pre-training. We also employ the recently introduced composed video retrieval framework **BLIP<sub>CoVR</sub>** [51] and **BLIP<sub>CoVR-ECDE</sub>** [47] which leverages the BLIP model for cross-attention between visual and textual encoders. They were specifically finetuned on WebVid-CoVR for CVR, leading to top performance on the WebVid-CoVR test set. We also evaluate CIREVL [28] on our benchmark since it is a training-free method. Below, we only discuss the self-adapted methods TF-CVR and TFR-CVR.

**Composed Video Retrieval by Language.** We use a methodology very similar to **CIREVL** [28], which has successfully been applied for Composed Image Retrieval. Given a video captioning model such as LaViLa [59], we can obtain the textual caption of the query video. We name this approach training-free CVR (**TF-CVR**). Specifically, given a query video  $q_v$ , and a video captioning model  $\Psi_C$ , we obtain its textual representation as  $c_q = \Psi_C(q_v) \in \mathcal{T}$ . However, this video caption only captures the reference video, not the specified textual modification  $q_t$ . While the two texts could be combined naively using concatenation, we use an LLM to combine the video caption and textual modifier into a coherent target caption, similar to CIREVL [28]. Formally, given access to an LLM  $\Psi_R$ , we generate a target video caption as  $c_q^t = \Psi_R(p \circ c_q \circ q_t)$ , which queries the LLM with a concatenation of the template prompt  $p$ , the generated video caption  $c_q$  and modification instruction  $q_t$ . The template prompt  $p$  consists of a few in-context examples to guide the LLM and a short task description. Concrete examples of this process are shown in the supplementary material. Given this generated target caption  $c_q^t$ , TF-CVR searches the video database  $\mathcal{D}$  alongside  $c_q^t$  using a text-video model (*e.g.* EgoVLPv2 [42], LanguageBind [60]). The retrieved target  $V_q^t$  is:

$$V_q^t = \underset{v \in \mathcal{D}}{\operatorname{argmax}} \frac{\Psi_V(v)^\top \Psi_T(c_q^t)}{\|\Psi_V(v)\| \cdot \|\Psi_T(c_q^t)\|} . \quad (1)$$

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

$$D' = \underset{v \in \mathcal{D}}{\operatorname{top} n_c} \left( \frac{\Psi_V(q_v)^\top \Psi_V(v)}{\|\Psi_V(q_v)\| \cdot \|\Psi_V(v)\|} \right) . \quad (2)$$

After filtering, we apply our proposed approach TF-CVR, except now, the video gallery is restricted to  $D'$ . We refer to this method as training-free re-ranking CVR (**TFR-CVR**). We demonstrate the efficacy of this method in Section 5.2,

especially in settings with a large video gallery. Note that the visual filtering applied in this step can use a different visual encoder than the text-video retrieval step, allowing us to leverage the complementary abilities of different models.

## 5 Experiments

We explain the two proposed evaluation settings and metrics in Section 5.1. Further, we discuss the results obtained on EgoCVR in Section 5.2 along with the ablations, analyses and qualitative examples performed on EgoCVR.

### 5.1 Evaluation Settings and Implementation Details

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

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

**Implementation Details.** We perform our experiments using the publicly available official implementations of various vision-language models [32, 42, 43, 60], using their default configurations to extract both visual and textual features. We employ the ViT-L/14 [16] version of the CLIP model provided by OpenCLIP [26] which was pre-trained on DataComp-1B [18], as well as, the BLIP-Large variant finetuned on COCO [35] and the BLIP model finetuned on WebVid-CoVR [51] by Ventura *et al.* [51]. We use the fully finetuned video encoder for LanguageBind [60] and EgoVLPv2 [42] with full projection. Unless otherwise noted, we use  $n_c = 15$  and employ EgoVLPv2 as the text encoder  $\Psi_T$  in TFR-CVR (Equation 1). CLIP and BLIP visual representations for the videos are obtained by averaging embeddings from 15 uniformly sampled image frames.

### 5.2 Benchmark Evaluation and Model Ablations

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

**Table 1:** Results on both the global and local evaluation settings on EgoCVR. Our proposed **TFR-CVR** achieves state-of-the-art results in both the global and local settings. We also report several baselines that only use the text, the reference video, or a naive average of the visual and textual embeddings (Fusion Strategy Avg). The best and the second best results are in **bold** and underlined, respectively.

Method	Video Model	Textual Input	Visual Input	Fusion Strategy	Global			Local		
					R@1	R@5	R@10	R@1	R@2	R@3
Random	✗	✗	✗	-	0.01	0.05	0.1	25.3	38.2	50.7
CLIP	✗	✓	✗	-	0.7	1.7	2.7	33.5	48.8	61.8
BLIP	✗	✓	✗	-	0.4	1.4	2.7	32.5	46.9	59.7
EgoVLPv2	✓	✓	✗	-	1.7	3.9	7.2	<u>41.0</u>	<u>57.3</u>	<u>69.0</u>
LanguageBind	✓	✓	✗	-	0.9	2.7	4.2	34.2	51.1	64.1
CLIP	✗	✗	✓	-	7.4	33.2	<u>55.3</u>	26.1	43.4	57.7
BLIP	✗	✗	✓	-	6.5	32.6	<u>55.3</u>	26.5	43.7	57.5
EgoVLPv2	✓	✗	✓	-	7.6	32.5	49.6	27.5	44.3	59.1
LanguageBind	✓	✗	✓	-	6.1	33.1	53.4	26.1	42.9	57.7
CLIP	✗	✓	✓	Avg	7.5	33.6	<b>55.6</b>	26.4	43.7	57.9
BLIP	✗	✓	✓	Avg	8.7	32.9	52.8	29.5	45.9	61.0
EgoVLPv2	✓	✓	✓	Avg	<u>9.5</u>	<u>34.9</u>	52.1	30.7	51.3	66.0
LanguageBind	✓	✓	✓	Avg	6.1	33.2	53.5	26.1	43.1	57.8
BLIP <sub>CoVR</sub> [51]	✗	✓	✓	Cross-Attention	5.4	15.2	24.3	33.1	49.5	62.9
BLIP <sub>CoVR-ECDE</sub> [47]	✗	✓	✓	Cross-Attention	6.0	16.3	24.8	33.4	49.3	63.0
CIReVL [28]	✗	✓	✓	Captioning	2.0	6.8	10.6	33.6	49.7	61.4
TFR-CVR (Ours)	✓	✓	✓	Captioning	<b>14.1</b>	<b>39.5</b>	54.4	<b>44.2</b>	<b>61.0</b>	<b>73.2</b>

**Global Setting.** The results for the global evaluation setting are presented in Table 1. We notice the absolute performance of the Recall@k ( $k \in \{1, 5, 10\}$ ) values being quite low due to having thousands of candidate video clips in the gallery for each query. However, we observe that relying solely on the text performs extremely poorly for all methods, as they fail to achieve an R@1 of even 2%. We notice that methods relying on visual features demonstrate competitive performance (up to 7.6% in R@1). In this setting, our proposed **TFR-CVR** achieves the best results (R@1 of 14.1%) due to the combination of LanguageBind-based candidate filtering using visual features, followed by re-ranking using the generated target caption. It is also notable that the BLIP finetuning methods (BLIP<sub>CoVR</sub> and BLIP<sub>CoVR-ECDE</sub>, which attain state-of-the-art results on WebVid-CoVR as well as several CIR benchmarks, does not generalise to our proposed **EgoCVR** benchmark, achieving an R@1 value of only 5.4% and 6.0% respectively. We also observe that CIReVL [28], which was tailored for the task of CIR, does not generalise directly to videos, obtaining an R@1 score of 2%.

**Local Setting.** In the local setting, we notice diminishing returns from methods that rely solely on visual features, performing only marginally better than random selection. This is due to the fact that, in this setting, all the videos in the gallery are by design very similar. Therefore, the visual similarity is not too helpful. Textual search performs much better because in the local setting, the textual information is the main indicator in finding the videos in the gallery.



**Table 2:** Results in terms of R@1, R@5 and R@10 on the global setting that emphasize the importance of temporal information on EgoCVR.

Method	Temporal	R@1	R@5	R@10
LanguageBind	✗	6.4	25.3	38.0
	✓	6.1	<b>33.1</b>	<b>53.4</b>
EgoVLPv2	✗	6.9	24.0	34.8
	✓	<b>9.5</b>	<b>34.9</b>	<b>52.1</b>
BLIP <sub>CoVR</sub>	✗	4.1	12.5	19.6
	✓	<b>5.4</b>	<b>15.2</b>	<b>24.3</b>
TFR-CVR	✗	10.2	27.4	39.2
	✓	<b>14.1</b>	<b>39.5</b>	<b>54.4</b>

**Table 3:** Results in terms of R@1, R@5 and R@10 demonstrating the importance of the two-stage (filtering and re-ranking) process for our proposed TFR-CVR on the global setting of EgoCVR. When applying re-ranking, we show the model that is applied for visual filtering before using TF-CVR.

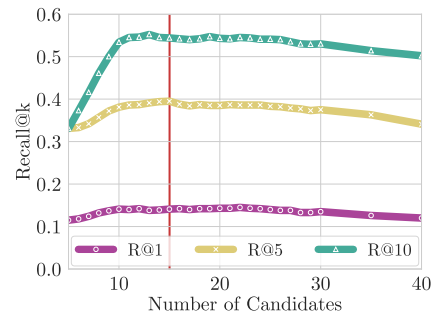
Method	R@1	R@5	R@10
LanguageBind (Stage 1)	6.1	33.1	53.4
EgoVLPv2 (Stage 1)	7.6	32.5	49.6
TF-CVR (Stage 2)	4.4	12.9	18.3
TFR-CVR (EgoVLPv2 → TF-CVR)	12.2	35.1	49.5
TFR-CVR (LanguageBind → TF-CVR)	<b>14.1</b>	<b>39.5</b>	<b>54.4</b>

TFR-CVR performs text-video retrieval with a full caption that also captures the information from the source video, achieves the best result (R@1 of 44.2%).

**Benefits of Temporal Information.** We demonstrate the benefits of using temporal information through processing the whole video compared to using only a single frame sampled from the middle of the video. The results are reported in Table 2. We observe that temporal information improves performance significantly across all methods (up to 12.1 percentage points in terms of R@5), confirming that our benchmark benefits and requires temporal understanding. This is particularly noticeable on the R@5 and R@10 metrics, where we observe TFR-CVR improving from 27.4% to 39.5% and from 39.2% to 54.3%, respectively, emphasising the importance of using temporal information for this task.

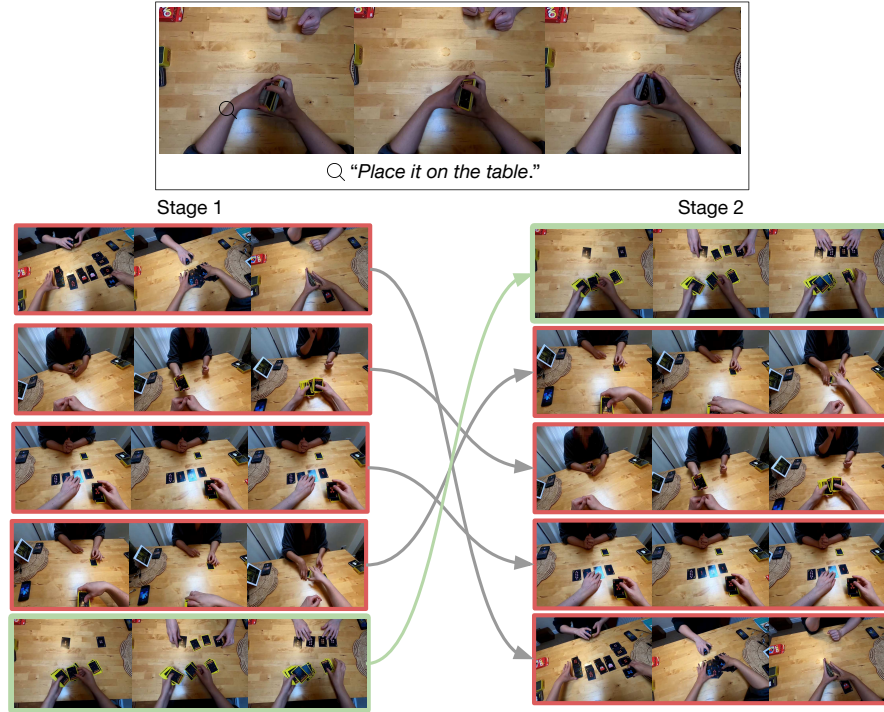
**Benefits of Re-Ranking.** We also demonstrate the efficacy of our two-stage approach, TFR-CVR (*i.e.* first selecting candidates using visual similarity and re-ranking them using text similarity), on the global setting in Table 3. We notice that only using visual similarity or textual search alone is insufficient, while combining the two steps leads to the best-performing results (last two rows). Additionally, we highlight the benefits of TFR-CVR in drawing complementary knowledge from distinct models. For instance, TF-CVR employs the textual encoder from EgoVLPv2. Using LanguageBind for visual filtering in TFR-CVR (last row) instead of EgoVLPv2 improves the retrieval results across all metrics.

In Figure 4, we illustrate the resulting order of the videos obtained after re-ranking. We can notice that in the first stage, while all videos are visually similar,



**Fig. 5:** Effect of the number of candidates  $n_c$  for the visual re-ranking step of TFR-CVR. The vertical line denotes the value of  $n_c$  used in our experiments.





**Fig. 4:** The first and the second stage ranking results of the TFR-CVR method. The arrows indicate how the ranking was changed. The correct video is showcased in green.

the correct video is ranked lower, while after the second stage, the target video is moved to the first position, resulting in a correct retrieval.

**Effect of the Number of Re-Ranking Candidates.** Our approach involves re-ranking the candidates chosen by the first stage of visual filtering. We study the impact of the number of neighbours chosen after the first stage in Figure 5. We observe that the performance stops fluctuating once we select a sufficient number of candidates  $n_c$  for re-ranking ( $n_c > 10$ ). Once the number of candidates becomes too large ( $n_c > 30$ ), the performance starts diminishing and eventually loses the benefits of the visual filtering. In our experiments, we use  $n_c = 15$ . However, the final results are stable within a large range of selected candidates.

**Effect of Text-Caption.** We also investigate the benefits of using an LLM to generate a plausible caption of the video, along with its shortcomings in Table 4. This is achieved by resorting to text-video retrieval with different textual inputs. We experiment with the input textual modification, the predicted caption (as the result of video captioning and LLM reformulation described in Section 4), as well as using the ground-truth target caption provided by Ego4D. The ground-truth target caption serves as a useful upper bound on text-only search. We notice that the video captioning and LLM reformulation consistently improve the results for

**Table 4:** Text-only retrieval results obtained with CLIP [43], LanguageBind [60], and TFR-CVR on EgoCVR. As text query alternatives, we switch among the *instruction*, the caption prediction from video captioning [59] combined with LLM reformulation (*Pred. Caption*), and lastly the ground-truth narration (*GT Caption*) available from Ego4D. The video query is used by TFR-CVR for visual re-ranking on the global evaluation.

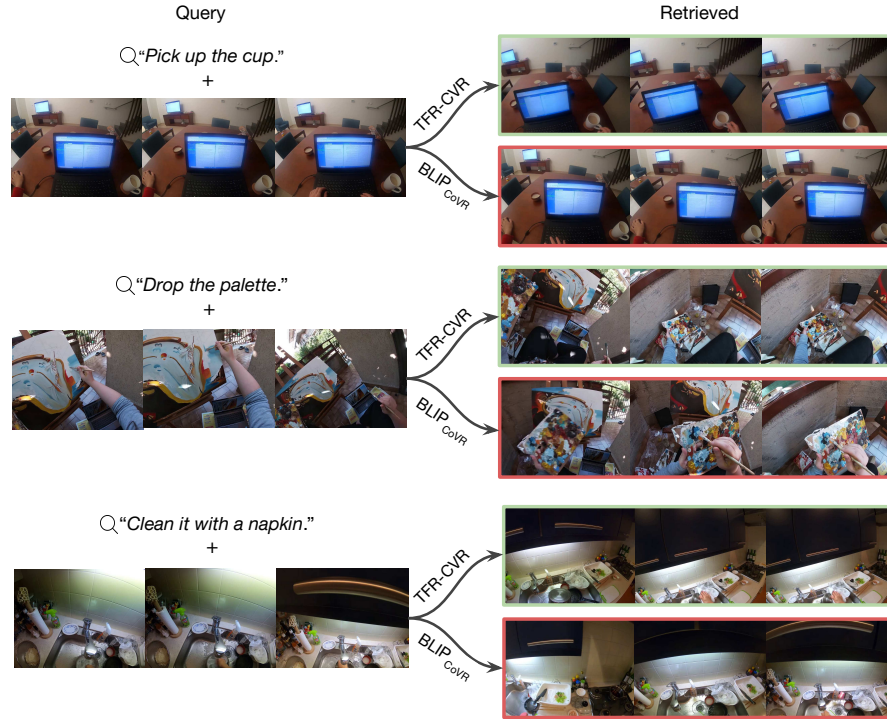
Method	Text Source	Global			Local		
		R@1	R@5	R@10	R@1	R@2	R@3
CLIP	Instruction	0.7	1.7	2.7	33.5	48.8	61.8
	Pred. Caption	1.5	4.2	7.5	34.0	49.8	63.7
	GT Caption	2.1	5.9	9.3	35.1	52.0	69.4
LanguageBind	Instruction	0.9	2.7	4.2	34.2	51.1	64.1
	Pred. Caption	1.7	5.7	8.2	36.6	52.2	64.8
	GT Caption	3.3	8.0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

all the models. In the global setting, the improvement is especially noticeable as the R@1 result increases at least twice (from 0.7% to 2.1% for the CLIP model) due to having a complete caption instead of a brief text. While the ground truth caption naturally improves the results further, it must be noted that improving the caption further does not offer much room for improvement. For instance, in the local setting, the R@1 result improves from 44.2% to 51.7% when our method is employed. Future work on EgoCVR would benefit both from improving the underlying text-video models through better foundation models [33, 58], as well as, from developing methods that can cohesively utilise the reference video and the textual instruction simultaneously.

**Qualitative Examples.** We demonstrate the benefits of our re-ranking approach in Figure 4. We observe that relying on visual features to select the top candidates results in visually similar clips, without capturing subtle actions accurately. After re-ranking with the text-video retrieval using the predicted caption, fine-grained actions are accurately captured in the final ranking.

We additionally illustrate qualitative examples in Figure 6, comparing the retrieved samples of the TFR-CVR and BLIP<sub>CoVR</sub> methods. We observe that our proposed method performs better than BLIP<sub>CoVR</sub>, retrieving all targets. Notably, these examples require fine-grained action understanding. While TFR-CVR returns the correct clip, BLIP<sub>CoVR</sub> retrieves visually similar clips, however, they do not display the correct action. This highlights the inherent limitations of an image-based model despite being finetuned for Composed Video Retrieval.

**Limitations.** We provide a high-quality evaluation benchmark for CVR. However, collecting a training set, even through an automated process, would allow finetuning models for CVR instead of adapting existing vision-language models in a training-free manner. Furthermore, our evaluation benchmark also consists



**Fig. 6:** Qualitative examples of composed video retrieval ranking on EgoCVR. For each example, we show the queries along with the clip retrieved by TFR-CVR and BLIP<sub>CoVR</sub> [51]. The target videos are enclosed in green rectangles.

of egocentric videos, however, it can be employed to assess the generalization of any CVR model. Despite the aforementioned limitations, we believe that our proposed benchmark serves as an intriguing validation ground for adapting existing vision-language models and a valuable evaluation set for high-quality temporal action understanding. Expanding the scope of the benchmark to include different types would also increase the diversity and applicability of the findings.

## 6 Conclusion

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

## Acknowledgements

This work was supported by BMBF FKZ: 01IS18039A, by the ERC (853489 - DEXIM), by EXC number 2064/1 – project number 390727645. Thomas Hummel and Shyamgopal Karthik thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for support. We also thank Yavuz Durmazkeser for his assistance with data labelling, which contributed to increasing our data size.

## References

1. Anwaar, M.U., Labintcev, E., Kleinsteuber, M.: Compositional learning of image-text query for image retrieval. In: WACV (2021) 3
2. Ashutosh, K., Xue, Z., Nagarajan, T., Grauman, K.: Detours for navigating instructional videos. In: CVPR (2024) 4
3. bai, Y., Xu, X., Liu, Y., Khan, S., Khan, F., Zuo, W., Goh, R.S.M., Feng, C.M.: Sentence-level prompts benefit composed image retrieval. In: ICLR (2024) 3
4. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: ICCV (2021) 3
5. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: A clip-hitchhiker’s guide to long video retrieval. arXiv preprint arXiv:2205.08508 (2022) 3
6. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-shot composed image retrieval with textual inversion. In: ICCV (2023) 3
7. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Effective conditioned and composed image retrieval combining clip-based features. In: CVPR Workshops (2022) 3
8. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021) 3
9. Chen, J., Lai, H.: Pretrain like you inference: Masked tuning improves zero-shot composed image retrieval. arXiv preprint arXiv:2311.07622 (2023) 3
10. Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., Liu, J.: Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. NeurIPS (2023) 3
11. Chen, Y., Bazzani, L.: Learning joint visual semantic matching embeddings for language-guided retrieval. In: ECCV (2020) 3
12. Chen, Y., Gong, S., Bazzani, L.: Image search with text feedback by visiolinguistic attention learning. In: CVPR (2020) 3
13. Delmas, G., Rezende, R.S., Csurka, G., Larlus, D.: ARTEMIS: Attention-based retrieval with text-explicit matching and implicit similarity. In: ICLR (2022) 3
14. Dong, J., Li, X., Snoek, C.G.: Predicting visual features from text for image and video caption retrieval. IEEE Transactions on Multimedia (2018) 3
15. Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Li, L., Sui, Z.: A survey on in-context learning (2023) 6
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 9
17. Forbes, M., Kaeser-Chen, C., Sharma, P., Belongie, S.: Neural naturalist: generating fine-grained image comparisons. In: EMNLP (2019) 3

18. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. In: *NeurIPS (2023)* [9](#)
19. Gatti, P., Parikh, K.G., Paul, D.P., Gupta, M., Mishra, A.: Composite sketch+text queries for retrieving objects with elusive names and complex interactions. In: *AAAI (2024)* [3](#)
20. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: *CVPR (2023)* [3](#)
21. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Ham-burger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: *CVPR (2022)* [2](#), [3](#)
22. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In: *CVPR (2024)* [3](#)
23. Gu, G., Chun, S., Kim, W., Jun, H., Kang, Y., Yun, S.: Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916 (2023)* [3](#)
24. Gu, G., Chun, S., Kim, W., Kang, Y., Yun, S.: Language-only efficient training of zero-shot composed image retrieval. In: *CVPR (2024)* [3](#)
25. Han, X., Wu, Z., Huang, P.X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., Davis, L.S.: Automatic spatially-aware fashion concept discovery. In: *ICCV (2017)* [3](#)
26. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: OpenCLIP. <https://doi.org/10.5281/zenodo.5143773> [9](#)
27. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: *ICML (2021)* [3](#)
28. Karthik, S., Roth, K., Mancini, M., Akata, Z.: Vision-by-language for training-free compositional image retrieval. In: *ICLR (2024)* [3](#), [4](#), [8](#), [10](#)
29. Lee, S., Kim, D., Han, B.: Cosmo: Content-style modulation for image retrieval with text feedback. In: *CVPR (2021)* [3](#)
30. Levy, M., Ben-Ari, R., Darshan, N., Lischinski, D.: Data roaming and early fusion for composed image retrieval. In: *AAAI (2024)* [3](#)
31. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: *ICML (2023)* [3](#)
32. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: *ICML (2022)* [2](#), [3](#), [7](#), [9](#)
33. Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122 (2023)* [13](#)
34. Lin, K.Q., Wang, J., Soldan, M., Wray, M., Yan, R., XU, E.Z., Gao, D., Tu, R.C., Zhao, W., Kong, W., et al.: Egocentric video-language pretraining. *NeurIPS (2022)* [2](#), [3](#)
35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *ECCV (2014)* [9](#)
36. Liu, Y., Yao, J., Zhang, Y., Wang, Y., Xie, W.: Zero-shot composed text-image retrieval. In: *BMVC (2023)* [3](#)



37. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: ICCV (2021) [1](#), [3](#), [4](#), [6](#)
38. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neuro-computing* (2022) [3](#)
39. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In: ICCV (2019) [3](#)
40. OpenAI: GPT-4 Technical Report. *arXiv* [abs/2303.08774](#) (2023) [6](#)
41. Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., Yokoya, N.: Learning joint representations of videos and sentences with web image search. In: ECCV Workshops (2016) [3](#)
42. Pramanick, S., Song, Y., Nag, S., Lin, K.Q., Shah, H., Shou, M.Z., Chellappa, R., Zhang, P.: EgoVLPv2: Egocentric video-language pre-training with fusion in the backbone. In: ICCV (2023) [2](#), [3](#), [7](#), [8](#), [9](#)
43. 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: ICML (2021) [2](#), [3](#), [7](#), [9](#), [13](#)
44. Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T.: Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In: CVPR (2023) [3](#)
45. Sun, S., Ye, F., Gong, S.: Training-free zero-shot composed image retrieval with local concept reranking. *arXiv preprint arXiv:2312.08924* (2023) [4](#)
46. Tang, Y., Yu, J., Gai, K., Jiamin, Z., Xiong, G., Hu, Y., Wu, Q.: Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In: AAAI (2024) [3](#)
47. Thawakar, O., Naseer, M., Anwer, R.M., Khan, S., Felsberg, M., Shah, M., Khan, F.S.: Composed video retrieval via enriched context and discriminative embeddings. In: CVPR (2024) [4](#), [8](#), [10](#)
48. Torabi, A., Tandon, N., Sigal, L.: Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124* (2016) [3](#)
49. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023) [2](#)
50. Vaze, S., Carion, N., Misra, I.: Genecis: A benchmark for general conditional image similarity. In: CVPR (2023) [3](#), [4](#)
51. Ventura, L., Yang, A., Schmid, C., Varol, G.: CoVR: Learning composed video retrieval from web video captions. In: AAAI (2024) [1](#), [3](#), [4](#), [5](#), [6](#), [8](#), [9](#), [10](#), [14](#)
52. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J.: Composing text and image for image retrieval-an empirical odyssey. In: CVPR (2019) [1](#), [3](#)
53. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: VATEX: A large-scale, high-quality multilingual dataset for video-and-language research. In: ICCV (2019) [3](#)
54. Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., Feris, R.: The Fashion IQ Dataset: Retrieving images by combining side information and relative natural language feedback. In: CVPR (2021) [1](#), [3](#)
55. Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., et al.: mPLUG-2: A modularized multi-modal foundation model across text, image and video. In: ICML (2023) [3](#)
56. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: A large video description dataset for bridging video and language. In: CVPR (2016) [3](#)

- 57. Xu, R., Xiong, C., Chen, W., Corso, J.: Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In: AAAI (2015) [3](#)
- 58. Zhao, L., Gundavarapu, N.B., Yuan, L., Zhou, H., Yan, S., Sun, J.J., Friedman, L., Qian, R., Weyand, T., Zhao, Y., et al.: VideoPrism: A foundational visual encoder for video understanding. arXiv preprint arXiv:2402.13217 (2024) [13](#)
- 59. Zhao, Y., Misra, I., Krähenbühl, P., Girdhar, R.: Learning video representations from large language models. In: CVPR (2023) [3](#), [8](#), [13](#)
- 60. Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., Wang, H., Pang, Y., Jiang, W., Zhang, J., Li, Z., et al.: LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment. In: ICLR (2024) [2](#), [3](#), [8](#), [9](#), [13](#)