CoVR: Learning Composed Video Retrieval from Web Video Captions

Lucas Ventura[1,2], Antoine Yang[2], Cordelia Schmid[2], Gül Varol[1]

[1]LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France
[2] Inria, ENS, CNRS, PSL Research University, France
lucas.ventura@enpc.fr

Abstract

Composed Image Retrieval (CoIR) has recently gained popularity as a task that considers *both* text and image queries together, to search for relevant images in a database. Most CoIR approaches require manually annotated datasets, comprising image-text-image triplets, where the text describes a modification from the query image to the target image. However, manual curation of CoIR *triplets* is expensive and prevents scalability. In this work, we instead propose a scalable automatic dataset creation methodology that generates triplets given video-caption *pairs*, while also expanding the scope of the task to include composed *video* retrieval (CoVR). To this end, we mine paired videos with a similar caption from a large database, and leverage a large language model to generate the corresponding modification text. Applying this methodology to the extensive WebVid2M collection, we automatically construct our WebVid-CoVR dataset, resulting in 1.6 million triplets. Moreover, we introduce a new benchmark for CoVR with a manually annotated evaluation set, along with baseline results. Our experiments further demonstrate that training a CoVR model on our dataset effectively transfers to CoIR, leading to improved state-of-the-art performance in the zero-shot setup on both the CIRR and FashionIQ benchmarks. Our code, datasets, and models are publicly available at https://imagine.enpc.fr/~ventural/covr.

Figure 1: Task: Composed Video Retrieval (CoVR) seeks to retrieve *videos* from a database by searching with both a query image and a query text. The text typically specifies the desired modification to the query image. In this example, a traveller might wonder how the photographed place looks like during a fountain show, by describing several modifications, such as "during show at night, with fireworks".

1 Introduction

Consider the scenario where a traveller takes a picture of a landmark or scenic spot and wants to discover videos that capture the essence of that location, by specifying certain conditions via text. For example, the query image in Figure 1 (of a fountain in Barcelona), along with the text "during show" should bring the video showcasing the fountain show. Further refining the text query such as "during show at night", would allow the traveller to decide whether to wait for the show until the night time. In this work, our goal is composed video retrieval (CoVR), where the user performs such multimodal search, by querying an image of a particular visual concept and a modification text, to find videos that exhibit the similar visual characteristics with the desired modification, in a dynamic context.

CoVR has many use cases, including but not limited to searching online videos for finding reviews of a specific product, how-to videos of a tool for specific usages, live events in specific locations, sports matches of specific players. Similar to composed image retrieval (CoIR), CoVR is also particularly useful when conveying a concept with a visual is easier and/or more accurate than only using words (e.g., unknown location/object, a specific camera view, a specific color).

Given the increased momentum in vision and language research in the recent years (Li et al. 2022b; Radford et al. 2021), CoIR has emerged as a new task (Vo et al. 2019), and since then witnessed improvements of both models and benchmarks (Baldrati et al. 2023, 2022; Gu et al. 2023; Levy et al. 2024; Liu et al. 2021; Wu et al. 2021). However, to the best of our knowledge, CoVR was not studied before. A key challenge in building CoVR models is the difficulty of gathering suitable training data of video-text-video triplets. We overcome this limitation by developing an automatic approach to generate triplets from existing video-caption collections. Specifically, we mine video pairs whose corresponding captions slightly differ in text space. We automatically describe this difference with a language model, which we train for a *modification-text generation* task. In particular, we use manually annotated triplets, each containing: (a) source caption, (b) target caption, (c) the modification text. We then finetune a large language model (LLM) (Touvron et al. 2023) by inputting (a-b), and outputting (c). We assume the resulting modification to describe the difference between the corresponding videos, thus obtaining video-text-video triplets (see Figure 2 for an overview). When training our CoVR/CoIR

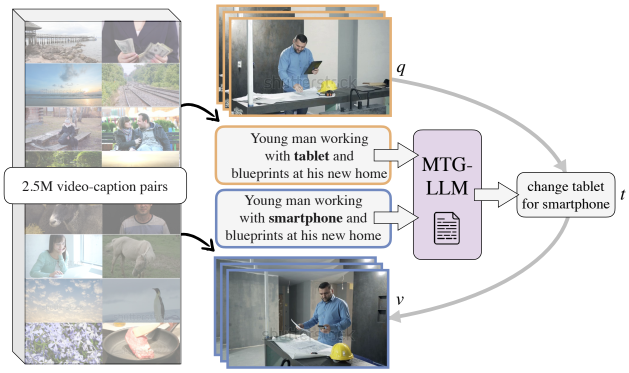

Figure 2: Method overview: We automatically mine similar caption pairs from a large video-caption database from the Web, and use our modification text generation language model (MTG-LLM) to describe the difference between the two captions. MTG-LLM is trained on a dataset of 715 triplet text annotations (Brooks, Holynski, and Efros 2023). The resulting triplet with the two corresponding videos (query q and target video v) and the modification text (t) is therefore obtained fully automatically, allowing a scalable CoVR training data generation.

models, we can flexibly select one or more frames from the videos, enabling multiple settings (i.e., retrieving images or videos).

We apply our triplet generation approach to the WebVid2M dataset (Bain et al. 2021) which contains 2.5M Web-scraped video-caption pairs. This results in the WebVid-CoVR training dataset with 1.6M CoVR triplets. By virtue of its automatic generation procedure, WebVid-CoVR is inherently noisy. To efficiently train on such large-scale and noisy data, we use a contrastive loss (van den Oord, Li, and Vinyals 2018), adopting the HN-NCE variant from (Radenovic et al. 2023) to upsample the significance of hard negatives. We design a CoVR model based on the cross-modal BLIP (Li et al. 2022b) and use query scoring (Bain et al. 2022) to exploit information from multiple video frames. Training this model on WebVid-CoVR shows strong transferability to the CoIR task, in both zero-shot and finetuning settings, achieving state-of-the-art results on the standard CIRR (Liu et al. 2021) and FashionIQ (Wu et al. 2021) benchmarks in the zero-shot setup. Finally, to foster research in CoVR, we repeat our generation procedure on a distinct subset of the WebVid10M dataset (Bain et al. 2021) and manually select correctly generated samples to constitute WebVid-CoVR-Test, a test set of 2,435 CoVR triplets. We find that our model achieves promising results on WebVid-CoVR-Test compared to standard baselines.

To summarize, our contributions are: (i) We propose a scalable approach to automatically generate composed visual retrieval training data. We apply this pipeline to the WebVid2M dataset and generate the WebVid-CoVR training dataset with 1.6M CoVR triplets. (ii) We show that training a CoVR model on WebVid-CoVR transfers well to the CoIR task, and achieves state-of-the-art results on the CIRR and FashionIQ benchmarks in the zero-shot setup. (iii) We evaluate our model on WebVid-CoVR-Test, a new CoVR benchmark that we manually annotate. Our code, datasets, and models are available at https://imagine.enpc.fr/~ventural/covr. Appendix is included in the arXiv version (Ventura et al. 2023).

2 Related Work

Composed image retrieval (CoIR). CoIR (Vo et al. 2019) has been an active area of research in recent years (Guo et al. 2018; Vo et al. 2019; Wu et al. 2021; Baldrati et al. 2022; Liu et al. 2021; Delmas et al. 2022; Kim et al. 2021; Baldrati et al. 2023; Saito et al. 2023; Gu et al. 2023). Most methods designed for this problem use manually annotated image-text-image triplets for training (Wu et al. 2021; Liu et al. 2021; Baldrati et al. 2022; Delmas et al. 2022). Very recent works, such as Pic2Word (Saito et al. 2023) and SEARLE (Baldrati et al. 2023), explore zero-shot CoIR setups where no manually annotated CoIR triplet is used. These approaches build on CLIP (Radford et al. 2021) and train a mapping network using image-only data for text inversion so that they can be flexibly composed with text descriptions. Our approach is similar in that it avoids collecting manual triplets; however, we instead perform supervised training on automatically generated image-text-video triplets given only video-text pairs. We also differ from above works by focusing on the composed video retrieval (CoVR) task, as opposed to CoIR.

Datasets for composed image retrieval. CIRR (Liu et al. 2021) and Fashion-IQ (Wu et al. 2021) are the two most widely used CoIR benchmarks. Both are manually annotated, hence small scale (about 30k triplets, see Table 1) due to the high cost implied in collecting CoIR triplets. To scale up, two concurrent works proposed larger, automatically generated CoIR datasets: LaSCo (Levy et al. 2024) and SynthTriplets18M (Gu et al. 2023). However, these two datasets are currently not publicly available. The LaSCo dataset (Levy et al. 2024) is generated using the visual question answering annotations and the pairing between images and counterfactual images in the VQAv2 dataset (Antol et al. 2015). In detail, this dataset provides for each (image, question, answer) triplet a counterfactual triplet with the same question and different image and answer. In contrast, we do not rely on such expensive annotation schemes. SynthTriplets18M (Gu et al. 2023) uses the text-conditioned image editing framework InstructPix2Pix (Brooks, Holynski, and Efros 2023) to automatically generate CoIR data. Their edit text generation process is similar to ours, but our generation process differs in that we automatically mine similar videos from a dataset of video-text pairs to construct CoVR triplets instead of generating visual data. In experiments, we show the superiority of our triplet construction procedure as we achieve much higher CoIR results (e.g., 38% vs 19% zero-shot R@1 on CIRR while generating fewer data). Lastly, our WebVid-CoVR dataset is not limited to still images and considers videos, while standing out as the largest composed retrieval dataset in the natural domain, as depicted in Table 1.

Vision-language pretraining. Many strong multi-modal models have been pretrained on large datasets of image-caption pairs (Alayrac et al. 2022; Chen et al. 2020; Li et al. 2021, 2023, 2020b; Radford et al. 2021; Jia et al. 2021; Su et al. 2019; Schuhmann et al. 2022) or video-caption

Dataset	Type	#Triplets	#Visuals	#Unique words	Avg. text length	Domain
CIRR (Liu et al. 2021)	📷	36,554	21,185	7,129	59.51	Natural
FashionIQ (Wu et al. 2021)	📷	30,132	7,988	4,425	27.13	Fashion
CIRCO-Test (Baldrati 2023)	📷	800	-	870	50.94	Natural
LaSCo (Levy et al. 2024)	📷	389,305	121,479	13,488	30.70	Natural
SynthTriplets18M (Gu et al. 2023)	📷	18,000,000	-	-	-	Synthetic
WebVid-CoVR	📹	1,648,789	130,775	19,163	23.36	Natural
WebVid-CoVR-Test	📹	2,556	2,444	1,935	21.97	Natural

Table 1: Existing datasets: We compare our proposed WebVid-CoVR training dataset and its manually annotated test set WebVid-CoVR-Test with existing composed visual retrieval datasets. 📷 denotes image, 📹 denotes video datasets. We contribute the largest training dataset for the natural domain. Note that, while SynthTriplets18M is larger, the transfer performance to real images is ineffective potentially due to a domain gap (see Table 3).

pairs (Akbari et al. 2021; Li et al. 2022a, 2020a; Miech et al. 2019, 2020; Xu et al. 2021; Sun et al. 2022; Xue et al. 2022a; Zhao et al. 2023). In contrast, we generate CoVR training data from video-caption pairs instead of directly training on them. Our data generation approach is also related to other generation approaches used for other tasks, e.g., action recognition (Nagrani et al. 2020), visual question answering (Yang et al. 2021), and visual dialog (Liu et al. 2023). However, unlike all these tasks, the CoVR task requires retrieving visual data.

Video retrieval. Text-to-video retrieval has received great attention over the last few years (Fang et al. 2021; Gao et al. 2021; Ge et al. 2022; Liu et al. 2022; Luo et al. 2021; Ma et al. 2022; Rasheed et al. 2023; Xue et al. 2022b; Yang, Bisk, and Gao 2021; Yao et al. 2022; Xu et al. 2021). We also make use of multiple video frames with query scoring similar to (Bain et al. 2022). However, different from these methods, we focus on *composed* video retrieval, where the query consists of both text and visual data.

3 Automatic Triplet Generation and Training

The goal of our composed video retrieval (CoVR) task is, given an input image q and a modification text t, to retrieve a modified video v in a large database of videos[1]. Our goal is to avoid the manual annotation of (q, t, v) triplets for training. Hence we automatically generate such triplets from Web-scraped video-caption pairs, as explained in Section 3.1 and illustrated in Figure 2. The resulting WebVid-CoVR dataset, together with its manually curated evaluation set, is presented in Section 3.2. Finally, we present how we train a CoVR model using WebVid-CoVR in Section 3.3.

3.1 Generating Composed Video Retrieval Triplets

Given a large (Web-scraped) dataset of video-caption pairs, we wish to automatically generate video-text-video CoVR triplets (q, t, v) where the text t describes a modification to the visual query q. However, the dataset of video-caption pairs neither contains annotations of paired videos, nor modification text that describes their difference. Hence we propose a methodology to automatically mine paired videos and describe their difference, as described below. Note that for illustration, we take as an example the WebVid2M dataset (Bain et al. 2021) with 2.5M video-caption pairs, but this methodology could potentially be applied to other large datasets of video-text (or image-text) pairs.

Mining paired videos by pairing captions. In order to obtain video pairs that exhibit visual similarity while differing in certain aspects, we leverage their associated captions. The core idea is that videos with similar captions are likely to have similar visual content. Specifically, we consider captions that differ by a single word, excluding punctuation marks. For instance, the caption *"Young woman smiling"* is paired with *"Old woman smiling"* and *"Young couple smiling"*. In the 2M distinct captions from WebVid2M, this process allows us to identify a vast pool of 1.2M distinct caption pairs with 177k distinct captions, resulting in 3.1M paired videos. In the following, we describe further steps to filter the data into a smaller set.

Filtering caption pairs. We wish to automatically generate the modification text between paired videos using their (paired) captions. However, caption pairs with the same meaning are likely to result in meaningless differences. On the contrary, caption pairs that differ too much are likely to result in large visual differences that cannot be easily described. To address these issues, we filter out caption pairs that are too similar and too dissimilar. Specifically, we exclude caption pairs with CLIP text embedding similarity ≥ 0.96 (e.g., *"Fit and happy young couple playing in the park"* and *"Fit and happy young couple play in the park"*) and caption pairs with CLIP text embedding similarity ≤ 0.6 (e.g., *"Zebra on a white background"* and *"Coins on a white background"*). We also exclude pairs where the captions differ by a digit (which mostly consist of a date in practice), a word not part of the English dictionary, or by a rare word. Rare words are detected based on the zipf`zipf` frequency (Speer 2022). Finally, we remove templated captions such as *"abstract of"*, *"concept of"*, and *"flag of"* which are over-represented in WebVid2M. At the end of this filtering stage, we have 370k distinct caption pairs with 92k distinct captions, resulting in

[1]Note that q could also be a video query, but in our main experiments we focus on an image query, and provide more results in the supplementary material (Section C.2) with video queries.

1.2M paired videos that we will use to generate the modification text. Note that we can use these paired videos in both directions to generate triplets, as the source and target videos can be swapped.

Generating a modification text from paired captions. In order to generate a modification text between paired videos, we develop and apply a "modification text generation large language model" (MTG-LLM) to their corresponding paired captions. We describe the MTG-LLM inference process below and then explain its training details. The MTG-LLM takes as input two paired captions and generates a modification text that describes the difference between the two captions (see Figure 2). In detail, the generation is auto-regressive, i.e., we recursively sample from the token likelihood distribution conditioned on the previously generated tokens until an end-of-sentence token is reached. Examples of the input-output, and details about the prompt format, which involves concatenating the two captions with a delimiter, can be found in Section B.4 of the Appendix in (Ventura et al. 2023). We use top-k sampling (Fan, Lewis, and Dauphin 2018) for generating the tokens instead of maximum-likelihood methods such as beam search. Note that we only generate a single modification text per caption pair for computational efficiency, but the MTG-LLM could be used to generate multiple modification texts per caption pair which could serve as a data augmentation in future work.

We now describe the training details of the MTG-LLM. We start from a LLM pretrained with a next token prediction objective on a Web-scale text dataset, namely LLaMA (Touvron et al. 2023). We then finetune this LLM for the MTG task on a manually annotated text dataset. In particular, we repurpose the editing dataset from InstructPix2Pix (Brooks, Holynski, and Efros 2023), which provides a modification text and a target caption for 700 input captions. We augment this dataset with 15 annotations that cover additional cases. More details about the additional examples can be found in Section B.4 of the Appendix in (Ventura et al. 2023).

Filtering video pairs. We wish to avoid some modification texts being over-represented in the dataset as it could harm training. Hence, if there are more than 10 video pairs associated with the same pair of captions (therefore leading to the same modification text), we only select top 10 video pairs. As the CoVR task typically involves similar query-target video pairs, we choose pairs of videos with the highest visual similarity, as measured by the CLIP visual embedding similarity computed at the middle frame of the videos.

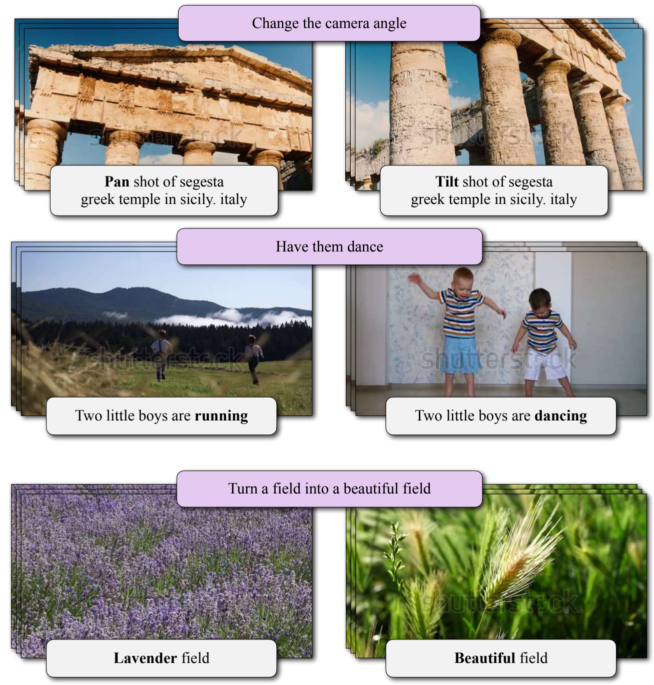

Figure 3: Examples of generated CoVR triplets in WebVid-CoVR: The middle frame of each video is shown with its corresponding caption, with the distinct word highlighted in bold. Additionally, the generated modification text is displayed on top of each pair of videos. The bottom example illustrates a noisy generated modification text, as 'beautiful' is subjective and both target and query videos can be considered as beautiful fields.

3.2 Our Resulting WebVid-CoVR Dataset

In the following, we describe the training and test partitions of our CoVR data. While our training set is automatically generated, our test set is manually verified.

WebVid-CoVR: a large-scale CoVR training dataset. By applying the previously described pipeline to the WebVid2M dataset (Bain et al. 2021), we generate WebVid-CoVR, a dataset containing 1.6M CoVR triplets, which is significantly larger than prior datasets (see Table 1). On average, a video lasts 16.8 seconds, a modification text contains 4.8 words, and one target video is associated with 12.7 triplets. We study the effect of the modification text length in Section C.6 of the Appendix in (Ventura et al. 2023). WebVid-CoVR is highly diverse with 131k distinct videos and 467k distinct modification texts. Examples of CoVR triplets from the WebVid-CoVR dataset are illustrated in Figure 3. These examples (along with additional ones included in Section D.3 of the Appendix) demonstrate the diversity present in WebVid-CoVR, highlighting a wide range of content and variations in the modification texts. However, it is important to acknowledge that some noise naturally exists in the dataset, as shown in the bottom example of Figure 3, where the text does not describe the difference between the two videos due to both videos describing beautiful fields. We provide further analysis such as removal of inappropriate content, and dataset statistics of WebVid-CoVR in Section A of the Appendix.

WebVid-CoVR-Test: a new CoVR evaluation benchmark. Due to the noise in WebVid-CoVR, we manually annotate a small test set, dubbed WebVid-CoVR-Test, for evaluation. For this, we first repeat the data generation procedure described in Section 3.1, but on a different corpus of video-caption pairs. Specifically, we consider video-caption pairs from the WebVid10M corpus (Bain et al. 2021) that are not included in the WebVid2M dataset, resulting in a pool of 8 million video-caption pairs. This ensures that other models using WebVid2M for pretraining have not been exposed to any of the test examples. In the video pairs filtering stage, for each pair of captions, we here only keep one pair of videos (the one with the highest visual similarity). This results in 163k candidate triplets that could be used for testing purposes. We randomly sample 7k triplets that we use for validation and randomly sample 3.2k other triplets that we manually annotate as described below.

We augment the 3.2k triplets by generating two additional modification texts with the MTG-LLM. The annotator reads the three generated modification texts, looks at three frames from the query and target videos, and either keeps the best modification text if at least one is valid or discards the sample. Through this meticulous annotation process, we ensure that the test set comprises high-quality and meaningful CoVR triplets. This results in a test set of 2.5k triplets, i.e., about 22% of the examples are considered as noisy and are discarded.

3.3 Training on WebVid-CoVR

Here, we describe our CoVR model architecture and how we train it on our WebVid-CoVR dataset.

CoVR-BLIP model architecture. Our model architecture builds on a pretrained image-text model, BLIP (Li et al. 2022b). BLIP is pretrained on a large dataset of image-caption pairs with three vision-language objectives: image-text contrastive learning, image-text matching, and image-conditioned language modeling. However, BLIP is not trained for composed visual retrieval with both visual and text inputs. Therefore we adapt BLIP to the CoIR/CoVR task as follows.

We use the BLIP image encoder to encode the image query q (which corresponds to the middle frame of the video in case of WebVid-CoVR). The resulting visual features and the modification text (t) are then forwarded to the BLIP image-grounded text encoder together, which outputs a multi-modal embedding $f(q,t) \in \mathbb{R}^d$ where d is the embedding dimension. To retrieve a target video v_k from a database of videos V, we compute embedding vectors for all gallery videos as follows. We uniformly sample N frames from the video and compute a weighted mean of the BLIP image embeddings to obtain the video embedding vector $h(v_k) \in \mathbb{R}^d$. The weights are obtained by computing the similarity between the corresponding frame and the modification text using the pretrained BLIP image and text encoders, respectively (similar to (Bain et al. 2022) in the context of text-to-video retrieval). Using pretrained and frozen BLIP embeddings allows us to precompute and store all the weights. Finally, given a multi-modal embedding $f(q,t)$, the retrieved video is the one that maximizes the embedding similarity, i.e., $\arg\max_{v_k \in V}(h(v_k) \cdot f(q,t)^T)$.

Training. In order to train on WebVid-CoVR, we use a contrastive learning approach (van den Oord, Li, and Vinyals 2018; Radenovic et al. 2023), as it has been shown to be effective to learn strong multi-modal representations from large-scale noisy data (Radford et al. 2021). We make the following design choices. First, we create a training batch by sampling distinct target videos; and for each target video, we randomly sample an associated image-text query pair. Iterating over videos ensures that the same target video appears only once in a batch and maximizes the number of different target videos that can be used as negatives in contrastive learning. We show the benefit of this approach in Section 4.4 (Table 6). Second, we employ HN-NCE (Radenovic et al. 2023) which increases the weight of most similar samples and uses as negatives all target videos $v_{j \in \mathcal{B}}$ in the batch $\mathcal{B}$. Formally, given a training batch $\mathcal{B}$ of triplets (q_i, t_i, v_i), we minimize the following loss:

$$\mathcal{L}(\mathcal{B}) = -\sum_{i \in \mathcal{B}} \log\left(\frac{e^{S_{i,i}/\tau}}{\alpha \cdot e^{S_{i,i}/\tau} + \sum_{j \neq i} e^{S_{i,j}/\tau} w_{i,j}}\right) - \sum_{i \in \mathcal{B}} \log\left(\frac{e^{S_{i,i}/\tau}}{\alpha \cdot e^{S_{i,i}/\tau} + \sum_{j \neq i} e^{S_{j,i}/\tau} w_{j,i}}\right)$$

where α is set to 1, temperature τ is set to 0.07, $S_{i,j}$ is the cosine similarity between the multi-modal embedding f_i and the target video embedding $\hat{v}_j$, and $w_{i,j}$ is set as in (Radenovic et al. 2023) with $\beta = 0.5$.

4 Experiments

We first describe the experimental protocol including the datasets, evaluation metrics, and implementation details (Section 4.1). We then present the results of CoVR on our new video benchmark (Section 4.2), as well as transfer results of CoIR on standard image benchmarks (Section 4.3). Finally, we provide ablations on our key components (Section 4.4).

4.1 Experimental Setup

Datasets. WebVid-CoVR is our proposed training CoVR dataset, and WebVid-CoVR-Test is our new CoVR benchmark, both presented in Section 3.2.

CIRR (Liu et al. 2021) is a manually annotated CoIR dataset that contains open-domain natural images from NLVR2 (Suhr et al. 2019), comprising 36.5k queries annotated on 19k images. CIRR includes two evaluation protocols: a standard one with the entire validation set as the search gallery, and a fine-grained *subset*, where the search space is a subgroup of six images similar to the query image (based on pretrained ResNet15 feature distance). The dataset is divided into training, validation, and testing splits with 28225/16742, 4181/2265 and 4148/2178 queries/images, respectively.

FashionIQ (Wu et al. 2021) is another CoIR dataset that contains images of fashion products, divided into three categories of Shirts, Dresses, and Tops/Tees. The query and target images were automatically paired based on title similarities (crawled from the web), and modification texts were then manually annotated. This dataset consists of 30k queries annotated on 40.5k different images. It is divided into training and validation splits with 18000/45429 and 6016/15415 queries/images, respectively.

Evaluation metrics. Following standard evaluation protocols (Liu et al. 2021), we report the video retrieval recall at rank 1, 5, 10, and 50. Recall at rank k (R@k) quantifies the number of times the correct video is among the top k results. MeanR denotes the average of R@1, R@5, R@10, and R@50. Higher recall means better performance.

Implementation details. For our MTG-LLM, we use LLaMA 7B model (Touvron et al. 2023) that we finetune for one epoch with an initial learning rate of 3e−5. For our

	Input modalities	Fusion	Backbone	#frames	WebVid-CoVR-Test R@1	R@5	R@10	R@50
Random	-	-	-	-	0.08	0.23	0.35	1.76
Not finetuned on WebVid-CoVR	Text	-	BLIP	-	19.68	37.09	45.85	65.14
	Visual	-	BLIP	15	34.90	59.23	68.04	85.95
	Visual + Text	Avg	CLIP	15	44.37	69.13	77.62	93.00
	Visual + Text	Avg	BLIP	15	45.46	70.46	79.54	93.27
Finetuned on WebVid-CoVR	Text	-	BLIP	-	23.67	45.89	55.13	77.03
	Visual	-	BLIP	15	38.89	64.98	74.02	92.06
	Visual + Text	MLP	CLIP	1	50.55	77.11	85.05	96.60
	Visual + Text	MLP	BLIP	1	50.63	74.80	83.37	95.54
	Visual + Text	CA	BLIP	1	51.80	78.29	85.84	97.07
	Visual + Text	**CA**	**BLIP**	**15**	**53.13**	**79.93**	**86.85**	**97.69**

Table 2: Benchmarking on the WebVid-CoVR-Test set: We observe that using both the visual and text input modalities performs better than individual modalities alone, both with/without finetuning on WebVid-CoVR (shown at the top/bottom of the table, respectively). When using pretraining models without finetuning, we apply average fusion (Avg) for the embeddings. BLIP performs slightly better than CLIP on this benchmark. Finetuning on WebVid-CoVR brings significant benefits. In this case, fusing with the pretrained cross-attention (CA) from BLIP is more effective than training a randomly-initialized MLP fusion as done in (Baldrati et al. 2022). Moreover, using multiple frames to embed the target video brings further improvements over using the middle frame. The first row represents the random performance.

CoVR model, we use the BLIP with ViT-L (Dosovitskiy et al. 2021) at 384 pixels finetuned for text-image retrieval on COCO and freeze the ViT for computational efficiency. We train our CoVR model on WebVid-CoVR for 4 epochs with a batch size of 2048 and an initial learning rate of $1e{-}5$. To finetune on CIRR/FashionIQ, we train for 6 epochs with a batch size of 2048/1024 and an initial learning rate of $1e{-}4$. We set hyperparameters based on the validation curve of WebVid-CoVR. Experiments are conducted on 4 NVIDIA A100-SXM4-80GB GPUs. More details are included in Section B of the Appendix.

4.2 Composed Video Retrieval Results

We provide a number of baselines for our new benchmark on WebVid-CoVR-Test. Table 2 summarizes these CoVR results. We first report the random chance performance in the first row. The rest of the table is split into two. The top block uses existing pretrained text and image encoders from BLIP (Li et al. 2022b) or CLIP (Radford et al. 2021) backbones without any finetuning. Models in the bottom block are finetuned on WebVid-CoVR. We report results with the composed query, as well as with the individual modalities. For combining modalities, we experiment with the simple average fusion baseline (Avg) when using frozen embeddings, and fusion with a randomly-initialized MLP or BLIP-pretrained cross-attention (CA) layers when finetuning. Note that the MLP fusion baseline is similar to Combiner (Baldrati et al. 2022) that concatenates the image and text embeddings from CLIP (or BLIP in CASE (Levy et al. 2024)), and is referred to as late fusion by CASE. For finetuning individual modalities, we train and test either with text-only query using the modification text, or with the visual-only image query. Finally, we experiment with using only the middle frame embedding or the weighted average of target video frame embeddings as explained in Section 3.3 (with the exception that visual-only experiments use equal weights due to not having access to the modification text for computing the scores).

We make several conclusions. (i) Combining both visual and text modalities yields better performance than the models with individual modalities. This result highlights that our new CoVR benchmark requires paying attention to both modalities. (ii) Visual-only outperforms text-only suggesting that the video pairs automatically mined through their caption similarity indeed exhibits visual similarity, and that the image captures the target video better than the modification text. (iii) Finetuning on WebVid-CoVR obtains substantial improvements over using pretrained and frozen embeddings. (iv) When finetuning, fusion with BLIP cross-attention (CA) performs better than the MLP fusion. (v) Results with the BLIP backbone are marginally higher than those with CLIP. (vi) Using $N = 15$ target video frames further boosts the performance over using only the middle frame. We further analyse the effect of the number of frames in Section C.7 of the Appendix in (Ventura et al. 2023).

4.3 Transfer Learning to CoIR

While our focus is video retrieval, we also experiment with transferring our CoVR models to image retrieval tasks on standard CoIR benchmarks. We define zero-shot CoIR as not using any manually annotated CoIR triplet for training. We perform zero-shot CoIR by directly applying our model trained on our automatically generated WebVid-CoVR dataset to CoIR tasks and also explore finetuning our model on the training set of the downstream benchmark.

Table 3 report results on CIRR and Fashion-IQ datasets. These results show that our model highly benefits from training on WebVid-CoVR, especially in the zero-shot setting, on both datasets. In addition, our model achieves state-of-the-art zero-shot performance on both CIRR and FashionIQ, and performs competitively in the finetuning setting.

Mode	Method	Pretraining Data	CIRR R@1	CIRR R@5	CIRR R@10	CIRR R@50	FashionIQ R@10	FashionIQ R@50
Train	TIRG+LastConv (Vo '19) †	-	11.04	35.68	51.27	83.29	-	-
	TIRG (Vo '19)†	-	14.61	48.37	64.08	90.03	-	-
	JVSM (Chen '20)	-	-	-	-	-	11.90	26.60
	TRACE w/BER (Jandial '20)	-	-	-	-	-	22.57	46.19
	VAL w/GloVe (Chen '20)	-	-	-	-	-	24.15	46.61
	CurlingNet (Yu '20)‡	-	-	-	-	-	25.90	51.01
	RTIC-GCN (Shin '21)‡	-	-	-	-	-	28.18	53.09
	CoSMo (Lee '21)	-	-	-	-	-	26.58	52.31
	DCNet (Kim '21)	-	-	-	-	-	27.78	53.89
	MAAF (Doods '20)†	-	10.31	33.03	48.30	80.06	24.30	48.80
	MAAF-BERT (Doods '20)†	-	10.12	33.10	48.01	80.57	-	-
	MAAF-IT (Doods '20)†	-	9.90	32.86	48.83	80.27	-	-
	MAAF-RP (Doods '20)†	-	10.22	33.32	48.68	81.84	-	-
	CIRPLANT (Liu '21)†	-	19.55	52.55	68.39	92.38	18.87	41.53
	SAC w/BERT (Jandial '22)	-	-	-	-	-	29.08	54.70
	FashionVLP (Goenka '22)	-	-	-	-	-	34.27	62.51
	ARTEMIS (Delmas '22)	-	16.96	46.10	61.31	87.73	26.05	50.29
	LF-CLIP (Baldrati '22)	-	-	-	-	-	35.39	59.03
	LF-BLIP (Levy 2023)	-	20.89	48.07	61.16	83.71	25.75	43.98
	Combiner (Baldrati '22)	-	33.59	65.35	77.35	95.21	-	-
	CompoDiff (Gu '23)	ST18M	22.35	54.36	73.41	91.77	40.98	62.23
	CASE (Levy '23)	-	48.00	79.11	87.25	**97.57**	-	-
	CASE (Levy '23)	LaSCo	48.68	79.98	88.51	97.49	48.79	70.68
	CASE (Levy '23)	LaSCo+COCO	49.35	**80.02**	**88.75**	97.47	-	-
	CoVR-BLIP	-	48.84	78.05	86.10	94.19	49.02	70.24
	CoVR-BLIP **(Ours)**	WebVid-CoVR	**49.69**	78.60	86.77	94.31	**49.40**	**70.98**
Zero Shot	Random†	-	0.04	0.22	0.44	2.18	0.06	0.32
	CompoDiff (Gu '23)	ST18M	19.37	53.81	72.02	90.85		
	Pic2Word (Saito '23)	CC3M	23.90	51.70	65.30	87.80	24.70	43.70
	CASE (Levy '23)	LaSCo	30.89	60.75	73.88	92.84	-	-
	CASE (Levy '23)	LaSCo+COCO	35.40	65.78	**78.53**	**94.63**	-	-
	CoVR-BLIP	-	19.76	41.23	50.89	71.64	16.00	32.77
	CoVR-BLIP **(Ours)**	WebVid-CoVR	**38.48**	**66.70**	77.25	91.47	**27.70**	**44.63**

Table 3: State-of-the-art comparison on the CIRR test set and FashionIQ validation set: Our model benefits from training on WebVid-CoVR in the zero-shot setting, and in the finetuning setting where it performs competitively. † denotes results reported by (Liu et al. 2021). For methods that pretrain specifically for composed retrieval, we indicate their pretraining data. CC3M denotes Conceptual Captions 3M (Sharma et al. 2018), ST18M denotes (Gu et al. 2023), and COCO denotes (Lin et al. 2014).

4.4 Ablation Studies

We now ablate the importance of several key aspects of our method by training only on WebVid-CoVR.

Importance of data scale. In Table 4, we evaluate the effect of the number of video-caption pairs used as a seed for our triplet generation pipeline. We construct subsets of videos such that larger ones include smaller ones, and only keep triplets that contain the sampled videos for training. We find that results steadily increase when using more videos, demonstrating that our method largely benefits from scaling the size of the seed dataset of video-captions. We also observe the importance of the filtering techniques described in Section 3.1, as the model trained on unfiltered data underperforms.

Modification text generation. We use a large language model finetuned for modification text generation (MTG-LLM) as explained in Section 3.1. We here compare this solution to a simple rule-based baseline that uses several templates to generate the modification text given the two captions

Initial #videos	Generated #triplets	Fil.	W-C-T MeanR	CIRR MeanR	FashionIQ MeanR
0	-	-	36.80	45.88	24.39
200k	11k	✓	62.20	65.25	**37.04**
500k	66k	✓	76.12	68.09	36.98
1M	269k	✓	77.89	68.20	36.47
2.5M	1.6M	✓	**79.40**	**68.48**	36.17
2.5M	3.6M	✗	78.92	67.41	32.31

Table 4: Data size: We measure the importance of the number of videos used for data generation and of filtering (Fil. column) the generated data, by evaluating on WebVid-CoVR-Test (W-C-T), CIRR, and FashionIQ. All models are trained for the same number of iterations on the generated data.

Model	WebVid-CoVR-T. R@1	WebVid-CoVR-T. MeanR	CIRR R@1	CIRR MeanR
Rule-based	41.39	70.36	15.66	46.16
Rule-based paraph.	52.39	78.75	33.54	64.33
MTG-LLM	**53.13**	**79.40**	**38.48**	**68.48**

Table 5: Modification text generation: We compare our MTG-LLM (LLaMA 7B parameters) against both a rule-based MTG baseline and a paraphrased rule-based MTG baseline (using GPT-3.5-turbo from OpenAI). We observe important gains in the downstream performance of the model trained on the generated data.

Iteration	HN-NCE	WebVid-CoVR-T. R@1	WebVid-CoVR-T. MeanR	CIRR R@1	CIRR MeanR
Triplets	✓	49.80	78.15	35.35	65.59
Videos	✗	49.02	76.62	35.57	66.07
Videos	✓	**53.13**	**79.40**	**38.48**	**68.48**

Table 6: Training strategies: Iterating on batches of distinct target videos (instead of triplets) and up-sampling hard negatives (HN-NCE (Radenovic et al. 2023)) both benefit the CoVR/CoIR performance.

that differ by one word. Specifically, the modification text is based on the two different words from the captions. We generate templates that use these words and choose one at random during training. These templates include variations such as "Remove `txt_diff`$_1$" and *"Change* `txt_diff`$_1$ *for* `txt_diff`$_2$*"*. A full list of all the templates can be seen in Section B.3 of the Appendix in (Ventura et al. 2023). Additionally, we investigate the possibility of paraphrasing the rule-based modification texts using `GPT-3.5-turbo` from OpenAI (Brown et al. 2020) as a source of augmentation, by prompting *"Paraphrase the following sentence:* `{Rule-base modification text}`*"*. In preliminary analysis, we qualitatively observed that LLaMA (Touvron et al. 2023) and LLaMA 2 (Touvron et al. 2023) alternatives were overly verbose when used for paraphrasing; however, GPT-3.5 outputs were satisfactory.

In Table 5, we show that our MTG-LLM generates better modification texts than the rule-based baseline, by evaluating the results of the model trained on the generated data. Paraphrasing the rule-based examples significantly boosts the performance (from 41 to 52 R@1), while still being worse than our MTG-LLM, especially on the CIRR benchmark. Note that the paraphrasing comes with the cost of running an expensive LLM ($43 cost for this experiment for 1 paraphrasing per modification text on the entire dataset). On the other hand, our MTG-LLM finetuning only requires 715 text examples. Qualitative examples comparing MTG-LLM and rule-based are provided in Table A.5 of the Appendix.

Training strategies. In Table 6, we first show the benefit on WebVid-CoVR of training by iterating on target videos instead of CoVR triplets. This is to avoid having the same target video appearing multiple times in a training batch, hence increasing the number of correct negatives that are used in the contrastive loss. Furthermore, up-sampling hard negatives adopting the HN-NCE loss formulation from (Radenovic et al. 2023) also slightly benefits the performance.

5 Conclusions and Limitations

In this work, we studied the new task of CoVR by proposing a simple yet effective methodology to create automatic training data. Our results on several benchmarks (including our manually curated video benchmark, as well as existing image benchmarks) suggest that, while noisy, such an automated and scalable approach can provide effective CoVR model training. One potential limitation of our method is that the generated modification text may not depict some visible changes due to not considering the image pair, but only their captions. Moreover, our modification text is suboptimal due to only inputting one-word difference caption pairs (i.e., focusing only on one change, and not considering multi-word differences). For example, the following modification with multiple changes from the CIRR dataset would not exist in our data "close up of a similar dog, but it is swimming on its own with a tennis ball in its mouth". Future work can incorporate visually-grounded modification generation and multiple modifications between query and target video pairs.

Ethics statement

Our model constitutes a generic multi-modal search tool, but is not intended for a specific application. While there are helpful use cases such as online shopping, traveling, and personal development (i.e., how-to), there may be potential privacy and harmful risks when training the model on datasets with inappropriate content. The risks include surveillance applications such as searching for a specific person, and looking up violent and graphic videos. For our WebVid-CoVR data release, we refer to Section A for further analysis about removal of inappropriate content. We note that our dataset users must also adhere to the terms of use stipulated by WebVid (Bain et al. 2021).

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011014223 made by GENCI. The authors would like to acknowledge the research gift from Google, the ANR project CorVis ANR-21-CE23-0003-01, Antoine Yang's Google PhD fellowship, and thank Mathis Petrovich, Nicolas Dufour, Charles Raude, and Andrea Blazquez for their helpful feedback.

References

Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.-H.; Chang, S.-F.; Cui, Y.; and Gong, B. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *NeurIPS*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV*.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *ICCV*.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2022. A CLIP-Hitchhiker's Guide to Long Video Retrieval. *arXiv:2205.08508*.

Baldrati, A.; Agnolucci, L.; Bertini, M.; and Del Bimbo, A. 2023. Zero-Shot Composed Image Retrieval with Textual Inversion. *ICCV*.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Bimbo, A. D. 2022. Effective conditioned and composed image retrieval combining CLIP-based features. In *CVPR*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*.

Brown et al., T. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

Chen, Y.-C.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *ECCV*.

Delmas, G.; de Rezende, R. S.; Csurka, G.; and Larlus, D. 2022. ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity. In *ICLR*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical Neural Story Generation. In *ACL*.

Fang, H.; Xiong, P.; Xu, L.; and Chen, Y. 2021. CLIP2Video: Mastering Video-Text Retrieval via Image CLIP. *arXiv:2106.11097*.

Gao, Z.; Liu, J.; Chen, S.; Chang, D.; Zhang, H.; and Yuan, J. 2021. CLIP2TV: An Empirical Study on Transformer-based Methods for Video-Text Retrieval. *arXiv:2111.05610*.

Ge, Y.; Ge, Y.; Liu, X.; Li, D.; Shan, Y.; Qie, X.; and Luo, P. 2022. BridgeFormer: Bridging Video-text Retrieval with Multiple Choice Questions. In *CVPR*.

Gu, G.; Chun, S.; Kim, W.; Jun, H.; Kang, Y.; and Yun, S. 2023. CompoDiff: Versatile Composed Image Retrieval With Latent Diffusion. *arXiv:2303.11916*.

Guo, X.; Wu, H.; Cheng, Y.; Rennie, S.; Tesauro, G.; and Feris, R. S. 2018. Dialog-based Interactive Image Retrieval. In *NeurIPS*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

Kim, J.; Yu, Y.; Kim, H.; and Kim, G. 2021. Dual Compositional Learning in Interactive Image Retrieval. *AAAI*.

Levy, M.; Ben-Ari, R.; Darshan, N.; and Lischinski, D. 2024. Data Roaming and Early Fusion for Composed Image Retrieval. *AAAI*.

Li, D.; Li, J.; Li, H.; Niebles, J. C.; and Hoi, S. C. 2022a. Align and Prompt: Video-and-Language Pre-training with Entity Prompts. In *CVPR*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. H. 2022b. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

Li, L.; Chen, Y.-C.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020a. HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training. In *EMNLP*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020b. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *ECCV*.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. In *NeurIPS*.

Liu, Y.; Xiong, P.; Xu, L.; Cao, S.; and Jin, Q. 2022. TS2-Net: Token Shift and Selection Transformer for Text-Video Retrieval. In *ECCV*.

Liu, Z.; Rodriguez-Opazo, C.; Teney, D.; and Gould, S. 2021. Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models. In *ICCV*.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2021. CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval. *arXiv:2104.08860*.

Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; and Ji, R. 2022. X-CLIP: End-to-End Multi-Grained Contrastive Learning for Video-Text Retrieval. In *ACMMM*.

Miech, A.; Alayrac, J.-B.; Smaira, L.; Laptev, I.; Sivic, J.; and Zisserman, A. 2020. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*.

Nagrani, A.; Sun, C.; Ross, D.; Sukthankar, R.; Schmid, C.; and Zisserman, A. 2020. Speech2Action: Cross-modal Supervision for Action Recognition. In *CVPR*.

Radenovic, F.; Dubey, A.; Kadian, A.; Mihaylov, T.; Vandenhende, S.; Patel, Y.; Wen, Y.; Ramanathan, V.; and Mahajan, D. 2023. Filtering, Distillation, and Hard Negatives for Vision-Language Pre-Training. In *CVPR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Rasheed, H.; Khattak, M. U.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Fine-tuned CLIP Models are Efficient Video Learners. In *CVPR*.

Saito, K.; Sohn, K.; Zhang, X.; Li, C.-L.; Lee, C.-Y.; Saenko, K.; and Pfister, T. 2023. Pic2Word: Mapping Pictures to Words for Zero-shot Composed Image Retrieval. *CVPR*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C. W.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS Datasets and Benchmarks Track*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL*.

Speer, R. 2022. rspeer/wordfreq: v3.0.2. https://doi.org/10.5281/zenodo.7199437. Zenodo.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *ICLR*.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *ACL*.

Sun, Y.; Xue, H.; Song, R.; Liu, B.; Yang, H.; and Fu, J. 2022. Long-Form Video-Language Pre-Training with Multimodal Temporal Contrastive Learning. In *NeurIPS*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971*.

Touvron et al., H. 2023. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. *arXiv:2307.09288*.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748*.

Ventura, L.; Yang, A.; Schmid, C.; and Varol, G. 2023. CoVR: Learning Composed Video Retrieval from Web Video Captions. arXiv:2308.14746.

Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.-J.; Fei-Fei, L.; and Hays, J. 2019. Composing Text and Image for Image Retrieval - an Empirical Odyssey. In *CVPR*.

Wu, H.; Gao, Y.; Guo, X.; Al-Halah, Z.; Rennie, S.; Grauman, K.; and Feris, R. 2021. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *CVPR*.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*.

Xue, H.; Hang, T.; Zeng, Y.; Sun, Y.; Liu, B.; Yang, H.; Fu, J.; and Guo, B. 2022a. Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions. In *CVPR*.

Xue, H.; Sun, Y.; Liu, B.; Fu, J.; Song, R.; Li, H.; and Luo, J. 2022b. CLIP-ViP: Adapting Pre-trained Image-Text Model to Video-Language Representation Alignment. *arXiv*.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2021. Just Ask: Learning To Answer Questions From Millions of Narrated Videos. In *ICCV*.

Yang, J.; Bisk, Y.; and Gao, J. 2021. TACo: Token-aware Cascade Contrastive Learning for Video-Text Alignment. In *ICCV*.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2022. FILIP: Fine-grained Interactive Language-Image Pre-Training. In *ICLR*.

Zhao, Y.; Misra, I.; Krähenbühl, P.; and Girdhar, R. 2023. Learning Video Representations from Large Language Models. In *CVPR*.