This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Composed Video Retrieval via Enriched Context and Discriminative Embeddings

Omkar Thawakar[1] **Muzammal Naseer**[1] **Rao Muhammad Anwer**[1,2] **Salman Khan**[1,3]
Michael Felsberg[4] **Mubarak Shah**[5] **Fahad Shahbaz Khan**[1,4]
[1]Mohamed bin Zayed University of AI [2]Aalto University [3]Australian National University
[4]Linköping University [5]University of Central Florida

Abstract

Composed video retrieval (CoVR) is a challenging problem in computer vision which has recently highlighted the integration of modification text with visual queries for more sophisticated video search in large databases. Existing works predominantly rely on visual queries combined with modification text to distinguish relevant videos. However, such a strategy struggles to fully preserve the rich query-specific context in retrieved target videos and only represents the target video using visual embedding. We introduce a novel CoVR framework that leverages detailed language descriptions to explicitly encode query-specific contextual information and learns discriminative embeddings of vision only, text only and vision-text for better alignment to accurately retrieve matched target videos. Our proposed framework can be flexibly employed for both composed video (CoVR) and image (CoIR) retrieval tasks. Experiments on three datasets show that our approach obtains state-of-the-art performance for both CovR and zero-shot CoIR tasks, achieving gains as high as around 7% in terms of recall@K=1 score. Our code, detailed language descriptions for WebViD-CoVR dataset are available at https://github.com/OmkarThawakar/composed-video-retrieval.

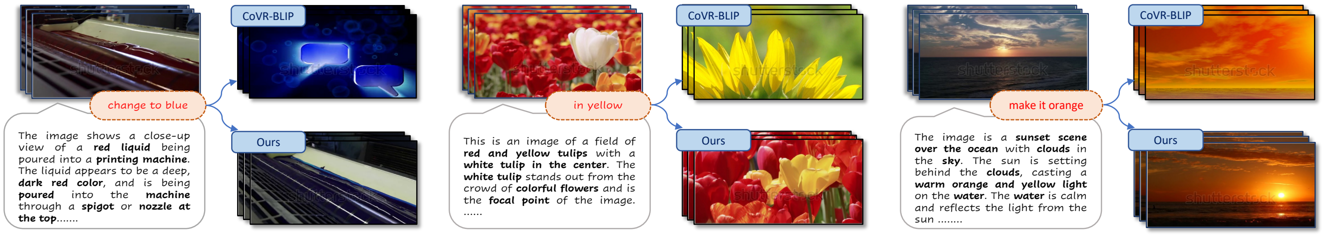
1. Introduction

Composed image retrieval (CoIR) is the task of retrieving matching images, given a query composed of an image along with natural language description (text). Compared to the classical problem of content-based image retrieval that utilizes a single (visual) modality, composed image retrieval (CoIR) uses multi-modal information (query comprising image and text) that aids in alleviating miss-interpretations by incorporating user's intent specified in the form of language descriptions (e.g., text-based modification to the query image). Following CoIR, composed video retrieval (CoVR) has been recently explored in the literature [43] where the multimodal search is performed to retrieve *videos* that display almost identical visual characteristics with the desired user intent, given a query image of a specific visual theme along with the modifier (change) text. CoVR is a challenging problem with various real-world applications, e.g., e-commerce and fashion, internet video search, finding live events in specific locations, and retrieving sports videos of particular players. In this work, we investigate the problem of composed video retrieval (CoVR).

The problem of CoVR poses two unique challenges: a) bridging the domain gap between the input query and the modification text, and b) simultaneously aligning the multimodal feature embedding with the feature embedding of the target videos that are inherently dynamic. Further, their context can also vary across different video frames. To address the problem of CoVR, the recent work [43] introduces an annotation pipeline to generate video-text-video triplets from existing video-caption datasets. The curated triplets contain the source and target video along with the change text describing the differences between the two videos. These triplets are then used to train a CoVR model, where a multimodal encoder encodes the image query to obtain visual features which are passed along with the change text to an image-grounded text encoder, thereby generating the feature embedding. In this way, a correspondence is established between the latent embedding of input visual query's and the desired change text to retrieve a target video.

We note that the aforementioned framework [43] struggles (see Fig. 1) since the latent embedding of a query visual input (image/video) is likely to be insufficient to provide necessary semantic details about the query image/video due to the following reasons: a) visual inputs are high-dimensional and offer details, most of which are not related to the given context, b) the visual depiction often shows a part of the broader context and there exist non-visual contextual cues that play a crucial role in understanding the given inputs. This motivates us to look into an alternative approach that explicitly encodes contextual information beyond what is apparent through only the visual input.

In this work, we argue that the detailed language descriptions of the visual content is likely to provide complementary



The image shows a close-up view of a **red liquid** being poured into a **printing machine**. The liquid appears to be a deep, **dark red color**, and is being **poured** into the **machine** through a **spigot** or **nozzle at the top**.......

This is an image of a field of **red and yellow tulips** with a **white tulip in the center**. The **white tulip** stands out from the crowd of **colorful flowers** and is the **focal point** of the image. ......

The image is a **sunset scene over the ocean** with **clouds** in the **sky**. The sun is setting behind the **clouds**, casting a **warm orange and yellow light** on the **water**. The **water** is calm and reflects the light from the sun ....

Figure 1. Comparison between the baseline CoVR-BLIP [43] (top row) and our approach (bottom row) on example video samples from the WebVid-CoVR testset. Here, the change text is highlighted in red. We observe that the baseline typically focuses only on the change while ignoring the semantic alignment of the target video with the query input video (e.g., the composed target video in the second example from the left should have change "yellow" reflected on the salient white tulip surrounded by red flowers, as in the query input). However, the retrieved target video loses the context (red flowers surrounding the yellow tulip). This suggests that it is particularly challenging for the model to understand the correspondence between the change text and the relevant target video using *only* the visual input. In contrast, our retrieved target videos are visually similar to the input query composed with the change text. Our approach leveraging detailed descriptions (highlighted in white boxes) for joint multi-modal embedding alignment encodes the necessary context to alter the composition of the video (e.g., changing the color of the "white flower" in 2nd video to yellow and changing the color of the "sky and clouds" to orange in 3rd video).

contextual information that is otherwise difficult to encode through visual input only. For instance, consider example query videos of red liquid, flowers, and sunset in Fig. 1. We can observe that the context becomes clear with language descriptions of these query videos; the red liquid is for the printing machine not immediately visible in the input, the white tulip stands out from the crowd of colorful red and yellow flowers, and the sun is setting over the ocean and behind the clouds. Here, richer semantics and a better context reduce the ambiguities while emphasizing important relationships e.g., the saliency of the white tulip means the change text relates to its color change, existing colors of the sky at sunset mean the change to orange should pertain to it.

Contributions: We propose a framework that explicitly leverages detailed language descriptions to preserve the query-specific contextual information, thereby reducing the domain gap with respect to the change text for CoVR. To this end, we utilize recent multi-modal conversational model to generate detailed textual descriptions which are then used during the training to complement the query videos. Furthermore, we learn discriminative embeddings of vision, text and vision-text during contrastive training to align the composed input query and change text with target semantics for enhanced CoVR. Our framework can be flexibly employed for both CoVR and CoIR tasks.

Extensive experiments on three datasets reveal the merits of our proposed contributions leading to state-of-the-art performance on both CoVR and zero-shot CoIR tasks. On the WebVid-CoVR dataset, our approach achieves a significant gain of ≈7% in terms of recall@K=1 score compared to recent CoVR-BLIP [43]. On the CIRR test set for the zero-shot setup, our approach achieves recall@K=1 score of 40.12. Fig. 1 shows a comparison on example WebVid-CoVR test set example video samples between our approach and the recent CoVR-BLIP method [43].

2. Related Work

Composed Image Retrieval (CoIR): A significant progress has been made in the field of content-based image retrieval thanks to recent advances in deep learning techniques [7, 16, 35, 46]. The problem holds extensive practical significance finding applications in diverse domains such as, product search, face recognition, and image geo-localization [18, 30, 34, 41]. Following the advances in cross-modal image retrieval, the scope has been extended to multiple query modalities such as, text-to-image retrieval, sketch-to-image retrieval, cross-view image retrieval, event detection and also to the problem of composed image retrieval (CoIR) [21, 28, 40, 45, 47]. CoIR is challenging since it requires image retrieval based on its reference image and corresponding relative change text. Most existing CoIR approaches are built on top of CLIP [37] and learn the multi-model embeddings comprising reference image and relative change text caption for target image retrieval [12–14, 29, 32]. These methods carefully harness the capabilities of large-scale pretrained image and text encoders, effectively amalgamating compositional image and text features to achieve improved performance.

Composed Video Retrieval (CoVR): The field of text-to-video retrieval has witnessed significant breakthroughs as a pivotal sub-domain within the broader context of multimedia information retrieval [12–14, 29, 32, 33]. Early efforts in this domain predominantly explored content-based retrieval approaches, leveraging key-frame analysis, color histograms, and local feature matching. The advent of deep learning techniques has further revolutionized text-to-video retrieval, with the emergence of multi-modal embeddings and attention mechanisms [38, 49–52]. Recently, [43] explored the problem of composed video retrieval (CoVR) where the objective is to retrieve the target video, given the reference video and its corresponding compositional change text. Due to the unavailability of a benchmark and following existing CoIR works [3, 31], [43] propose a new benchmark for CoVR, named WebVid-CoVR, which comprises a synthetic training set and a manually curated test set. Further, the authors also propose a framework, named CoVR-BLIP, that is built on top of BLIP [26] where an image grounded text encoder is utilized to generate multi-model features and aligns it with target video embeddings using a contrastive loss [36].

Our Approach: Different from COVR-BLIP [43], our approach leverages detailed language descriptions of the reference video that are automatically generated through a multi-modal conversation model and provide with following advantages. First, it helps in preserving the query-specific contextual information and aids in reducing the domain gap with the change text. Second, rather than relying on only using the visual embedding to represent target video as in [43], learning discriminative embeddings through vision, text, and vision-text enables improved alignment due to the extracting complementary target video representations. It is worth mentioning that these automatically generated detailed language descriptions can be effectively utilized within our framework either only during training or at both training and inference. In both cases, our approach leads to superior performance compared to original [43] as well as using default (short) text captions within [43]. Furthermore, our approach exhibits notable competitive capabilities in both transfer learning and zero-shot learning contexts for the CoIR task.

3. Method

Problem Statement: Composed Video Retrieval (CoVR) strives to retrieve a target video from a database. This target video is desired to be aligned with the visual cues from a query video but with the characteristics of the desired change represented by the text. Formally, for a given embedding of input query $q \in Q$ and the desired modification text $t \in T$, we optimize for a multi-modal encoder f and a visual encoder g, such that $f(q,t) \approx g(v)$, where $v \in V$ is the target video from a database. As discussed earlier, the problem of CoVR is challenging since it requires bridging the domain gap between input query q and the modification text t. Furthermore, it requires simultaneously aligning the multi-modal feature embedding $f(q,t)$ with the feature embedding of target videos that are inherently dynamic, and their context also varies across different video frames.

Baseline Framework: To address the above problem, we base our method on the recently introduced framework [43], named CoVR-BLIP, that trains the multi-modal encoder f which takes the representations from the visual encoder g. The visual encoder g remains frozen and is used to get the latent embeddings for visual input query which are then provided to multi-modal encoder f along with the tokenized change text t to produce multi-modal embedding $f(q,t)$. Then, the input visual query and the change text t are aligned with the desired target videos using a contrastive loss between $f(q,t)$ and $g(v)$ (Fig. 2). This results in a direct correspondence between the visual latent embedding of the input query and the desired change text for retrieving a target video. For more details, we refer to [43].

We note that the baseline CoVR-BLIP framework struggles to effectively preserve the contextual information of the query sample, since the multi-modal information is likely biased towards the change text. This is evident in Fig. 1, where the dominant feature in the baseline is the change text mixed with the holistic representation of the visual query e.g., yellow follower or orange sky. Instead, here the objective was to convert only the white tulip to yellow with the surrounding red flowers or orange sky over the beach. Next, we propose our approach that aims to alleviate these limitations for improved CoVR performance.

3.1. Architecture Design

Motivation: To motivate our proposed approach, we distinguish two desirable characteristics that are to be considered when designing an approach for the CoVR task.

Query-specific Contextual Information Preservation: As discussed earlier, compositional video retrieval (CoVR) relies on reducing the domain gap between the visual input and the change text. This is typically achieved by leveraging an image-grounded text encoder, where the cross-attention layers are trained using the changing text and embedding from a frozen visual encoder [43]. As the training occurs within the image-grounded text encoder only, the multi-modal representation is likely to get predominately biased towards the change text. As a result, it looses the context of the query sample which is essential for the task.

In this work, we argue that such a query-specific contextual information can be incorporated in the image-grounded text encoder through detailed language descriptions of these query videos; the red liquid is for printing machine not immediately visible in the input, the white tulip stands out from the crowd of colorful red and yellow flowers, and the sun is setting over the ocean and behind the clouds (see Fig. 1). Therefore, the correspondence between embedding of the visual input (q) and its corresponding detailed description (d) results in an enhanced vision-text representation $f(q,d)$ of q, thereby ensuring a contextualized understanding of the query video. Further, the complementary nature of detailed descriptions aids in reducing the domain gap between the input query and the modification text by establishing correspondence between the detailed description of the input query and the modification text, as $f(d,t)$. Thus, we seek to improve CoVR by minimizing the following objective:

$$v^* = \underset{v \in V}{\arg\max} \;\; \mathcal{L}\left(\tilde{f}(q,d,t), \;\; g(v)\right), \tag{1}$$

$$\tilde{f}(q,d,t) = f(q,t) + f(q,d) + f(e(d),t). \tag{2}$$

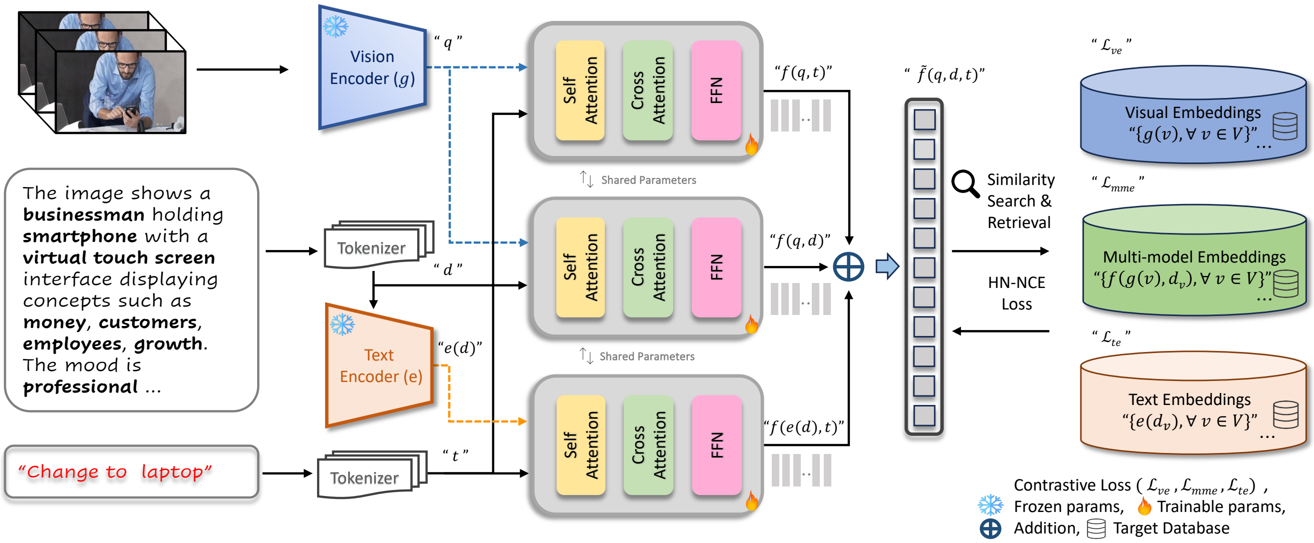

Vision Encoder (g)
" q "
The image shows a **businessman** holding **smartphone** with a **virtual touch screen** interface displaying concepts such as **money**, **customers**, **employees**, **growth**. The mood is **professional** ...
Tokenizer
" d "
Text Encoder (e)
"$e(d)$"
"Change to laptop"
Tokenizer
" t "
Self Attention
Cross Attention
FFN
Shared Parameters
Self Attention
Cross Attention
FFN
Shared Parameters
Self Attention
Cross Attention
FFN
"$f(q,t)$"
"$f(q,d)$"
"$f(e(d),t)$"
" $\tilde{f}(q,d,t)$"
Similarity Search & Retrieval
HN-NCE Loss
" $\mathcal{L}_{ve}$ "
Visual Embeddings "$\{g(v), \forall v \in V\}$" ...
" $\mathcal{L}_{mme}$ "
Multi-model Embeddings "$\{f(g(v), d_v), \forall v \in V\}$" ...
" $\mathcal{L}_{te}$ "
Text Embeddings "$\{e(d_v), \forall v \in V\}$" ...
Contrastive Loss ($\mathcal{L}_{ve}$, $\mathcal{L}_{mme}$, $\mathcal{L}_{te}$) ,
Frozen params, Trainable params,
Addition, Target Database

Figure 2. Our framework comprises three inputs: the reference video, a detailed visual description of an input video, and a change text corresponding to the target video. The input video is encoded by the vision encoder g, and the description is encoded by the frozen text encoder e. The default tokenizer tokenizes the change text. The encoded input triplet (q,d,t) is then processed by the multi-model encoder (f) grounding two inputs a time. The dotted lines shown are going to the cross-attention for grounding. During training, we add outputs of the multi-model encoder to obtain the joint multi-model embedding $\tilde{f}(q,d,t)$ that is aligned across three target databases using hard negative contrastive losses (HN-NCE): $\mathcal{L}_{ve}$, $\mathcal{L}_{mme}$, and $\mathcal{L}_{te}$. During inference, our approach can utilize input query or a combination of input query along with its description to retrieve a composed target video.

where, q and d represent input query (image/video) and its corresponding language description, t is the desired modification text, $\tilde{f}(q,d,t)$ is the pairwise summation of individual correspondence embeddings and $\mathcal{L}$ is a similarity-based loss.

Learning Discriminative Embeddings for Alignment: In the CoVR task, the model is desired to learn to align its output with the target video after mixing the change text with the query video. Instead of only representing the target video in the latent space through a visual embedding [43], a multiple discriminative embedding of vision, text, and vision-text is expected to provide better alignment due to complementary target video representation.

Overall Architecture: Figure 2 presents our proposed architecture comprising three inputs: the reference video, the text corresponding to the change, and the detailed video description. Compared to the baseline framework, the focus of our design is to effectively align the joint multi-modal embedding, comprised of the three inputs ($\tilde{f}(q,d,t)$), with the target video database to achieve enhanced contextual understanding during training for composed video retrieval. Within our proposed framework, we first process the reference video and its description using pre-trained [26] image encoder g and text encoder e to produce their latent embedding of the same dimension as, $q \in \mathbb{R}^m$ and $d \in \mathbb{R}^m$. We use the same multi-modal encoder f, as in the baseline [43]. This multi-model encoder takes the visual embeddings from a pre-trained visual encoder g along with tokenized textual inputs and produces a multi-modal embedding. Given the tokenized change text t, and embeddings of the reference video and its descriptions, q and d, we fuse any two inputs at a time using the multi-modal encoder f comprising of cross-attention layers, to produce joint multi-modal embeddings ($\tilde{f}(q,d,t)$), as shown in Fig. 2. The input query video and its corresponding description are processed by visual encoder g and text encoder, respectively. It is worth mentioning that the only difference between the text encoder e and the multi-modal encoder f are cross attention layers. In other words, if we remove the cross attention layers from multi-modal encoder f, it converts to text-only encoder e. We use the text encoder e to process the language descriptions of an input video.

Within the proposed framework, we only train the multi-modal encoder f whereas the image and text encoders remain frozen. During training, we provide the change text t and the visual query embeddings q to the encoder f for obtaining the multi-model embeddings $f(q,t)$ corresponding to the change text t. As shown in Fig. 2, here grounding occurs via cross-attention between q and t. In a similar manner, we obtain an enhanced contextualized multi-model representation of embeddings of input video q and its tokenized description d from f as $f(q,d)$. As a final step, we provide the change text t to the text encoder f, and ground it with the embedding of description $e(d)$ to obtain $f(e(d),t)$. Consequently, we combine these grounded embeddings in a pairwise summation manner as shown in Eq. (2) to obtain the joint multi-model embeddings $\tilde{f}(q,d,t)$. These joint multi-model embeddings are then utilized to retrieve the target video from the database. In order to train the multi-modal encoder f, we employ hard-negative contrastive loss [36, 43] between $\tilde{f}(q, d, t)$ and the target database, as shown in Fig. 2. The loss is as follows,

$$\mathcal{L} = -\sum_{i\in\mathcal{B}} \log\left(\frac{e^{S_{i,i}/\tau}}{\alpha\cdot e^{S_{i,i}/\tau}+\sum_{j\neq i} e^{S_{i,j}/\tau} w_{i,j}}\right) - \sum_{i\in\mathcal{B}} \log\left(\frac{e^{S_{i,i}/\tau}}{\alpha\cdot e^{S_{i,i}/\tau}+\sum_{j\neq i} e^{S_{j,i}/\tau} w_{j,i}}\right) \tag{3}$$

where α is set to 1 and temperature τ is set to 0.07 as in [36], $S_{i,j}$ is the cosine similarity between the joint multi-modal embedding $\tilde{f}(q_i, d_i, t_i)$ and the corresponding target video $g(v_i)$, $w_{i,j}$ is set as in [36] with $\beta = 0.5$, and $\mathcal{B}$ is the batch size. Next, we describe how to effectively utilize the recent multi-modal conversation models [54] to obtain query-specific detailed language descriptions for composed video retrieval.

3.2. Query-specific Language Descriptions

In order to obtain the video descriptions, we employ a recent open-source multi-modal conversation model [54]. Generally, multi-modal conversation models learn alignment between a pretrained large language model such as, Vicunna [6] and a pretrained vision encoder of vision language model such as, CLIP [37] or BLIP [25]. This alignment enables these multi-modal conversation models to reason and contextualize a given visual input. Since these are image models and for our case of video inputs, we sample the middle frame of the video and generate its detailed description using a multi-modal conversation model by prompting the model with "Describe the input image in detail". We further remove the noise within these descriptions by removing the manually curated unnecessary symbols, tokens, or special characters. Further, these models can hallucinate about a given visual sample. To identify hallucinated descriptions, we first measure the lower bound of cosine similarity between default WebVid captions [1] and visual inputs within BLIP latent space to identify a hallucination threshold. We then discard those descriptions, where the cosine similarity between our generated description and the visual input is lower than the hallucination threshold. Consequently, the resulting enriched descriptions are better aligned with the videos (Fig. 3). As discussed earlier, the base framework [43] only aligns the input video with the target video database. To further enhance the alignment of our joint multi-modal $\tilde{f}(q, d, t)$, we introduce multiple target datasets as explained next.

3.3. Enhancing Diversity in Target Database

The proposed method takes three inputs (video, modification text, and video description) and three target databases to

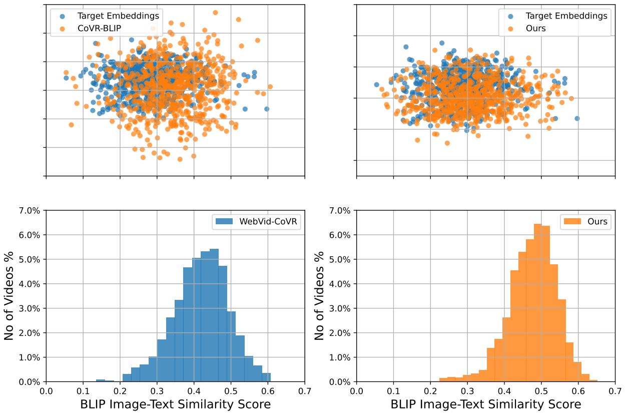


Figure 3. **First row:** Comparison between the baseline and our approach in terms of proximity of the output embedding with the target videos on WebVid dataset. Here, each data sample represents the projection of the embedding from $\mathbb{R}^m$ to $\mathbb{R}^2$. Our joint multi-modal embeddings leveraging the language information are closer to the target embeddings, compared to the baseline embedding utilizing only the visual input. **Second row:** the cosine similarity between video embeddings and the WebVid dataset captions (on the left), compared to the similarity between video embeddings and our generated textual descriptions (on the right). Here, the Y-axis corresponds to the number of videos whereas the X-axis denotes the cosine similarity. Our approach utilizing the generated descriptions achieves better alignment with the video embeddings.

train the model. The first target database is based on the visual embedding of input videos generated by a pretrained vision encoder of BLIP-2 [26]. Our second target database is based on multi-model embeddings derived from the pretrained multi-modal encoder of BLIP-2 [26]. The final target database is based on a text-only embedding of the video description generated by the pretrained BLIP-2 [26] text encoder. We use these additional databases only during training time to compute the hard negative contrastive loss between our joint multi-model embeddings and target datasets.

Overall Loss Formulation: For a given batch $\mathcal{B}$, we formulate hard negative contrastive loss for each of our three target databases as follows,

$$\mathcal{L}_{contr} = \lambda * \mathcal{L}_{ve} + \mu * \mathcal{L}_{mme} + \delta * \mathcal{L}_{te}, \tag{4}$$

where, $\mathcal{L}_{ve}$, $\mathcal{L}_{mme}$, and $\mathcal{L}_{te}$ are the contrastive loss represented by Eq. (3). We compute the similarity of $\tilde{f}(q_i, d_i, t_i)$ with the corresponding target video embedding $g(v_i)$, target multi-modal embedding $f(g(v_i), d_{v_i})$, and the target text embedding d_{v_i} for $\mathcal{L}_{ve}$, $\mathcal{L}_{mme}$, and $\mathcal{L}_{te}$, respectively. λ, μ, and δ are learnable parameters that scale the weightage of each loss during training.

Inference: During inference, for the 3 given inputs: reference video, description and change text, we first process the reference video and its description using pre-trained frozen image encoder g and text encoder e to produce their latent embedding. The change text is simply tokenized as shown

		Training					Recall@K			
	Model	WebVid-CoVR	Input Modalities	Fusion	Backbone	Frames	R@1	R@5	R@10	R@50
1	Random		-	-	-	-	0.08	0.23	0.35	1.76
2	CoVR-BLIP [43]	✘	Text	-	BLIP	-	19.68	37.09	45.85	65.14
3	CoVR-BLIP [43]	✘	Visual	-	BLIP	15	34.90	59.23	68.04	85.95
4	CoVR-BLIP [43]	✘	Visual + Text	Avg	CLIP	15	44.37	69.13	77.62	93.00
5	CoVR-BLIP [43]	✘	Visual + Text	Avg	BLIP	15	45.46	70.46	79.54	93.27
6	**Our Approach**	✘	Visual + Text	Avg	BLIP	15	**47.52**	**72.18**	**82.37**	**95.06**
7	CoVR-BLIP [43]	✘	Visual + Text	CA	BLIP	15	15.85	32.79	40.3	58.33
8	**Our Approach**	✘	Visual + Text	CA	BLIP	15	**20.85**	**41.2**	**50.2**	**72.1**
9	CoVR-BLIP [43]	✔	Text	-	BLIP	-	23.67	45.89	55.13	77.03
10	CoVR-BLIP [43]	✔	Visual	-	BLIP	15	38.89	64.98	74.02	92.06
11	CoVR-BLIP [43]	✔	Visual + Text	MLP	CLIP	1	50.55	77.11	85.05	96.06
12	CoVR-BLIP [43]	✔	Visual + Text	MLP	BLIP	1	50.63	74.8	83.37	95.54
13	CoVR-BLIP [43]	✔	Visual + Text	CA	BLIP	1	51.80	78.29	85.84	97.07
14	CoVR-BLIP [43]	✔	Visual + Text	CA	BLIP	15	53.13	79.93	86.85	97.69
15	**Our Approach**	✔	Visual + Text	CA	BLIP	15	**60.12**	**84.32**	**91.27**	**98.72**

Table 1. **Baseline comparison on the WebVid-CoVR test set**. Without training on the WebVid-CoVR and using averaging as fusion, our approach (row 6) achieves a gain of xx over the baseline (row 5). A consistent improvement in performance is also obtained over the baseline (row 7 vs. row 8) when using cross-attention (CA) as a fusion scheme. The performance is improved when performing training on the WebVid-CoVR training set. Using the same input modalities, fusion scheme, backbone, and frames, our approach (row 15) achieves a significant gain of 6.9% in terms of Recall@K=1 over the baseline [43] (row 14). Best results are in bold.

in Figure 2. We use the multi-modal encoder f and gather the multi-model embeddings from 2 inputs at a time such as $f(q,t)$, $f(q,d)$ and $f(e(d),t)$. Consequently, we simply do the pairwise addition of three (3) multi-model embeddings to produce joint multi-modal embeddings $\tilde{f}(q,d,t)$ for target video retrieval. Note that, this pairwise addition allows us to use any combination of inputs as illustrated in ablative analysis (refer Tab. 2). Finally, similar to CoVR [43] the target videos are retrieved by mapping the similarity between the joint multi-model embeddings $\tilde{f}(q,d,t)$ and the visual embedding database $g(V)$.

4. Experiments

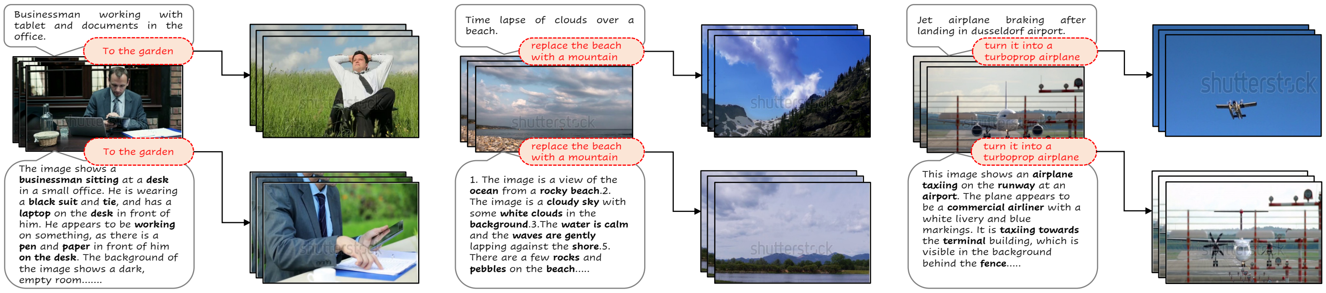
4.1. Experimental Setup and Protocols

Dataset for Composed Video Retrieval (CoVR): We evaluate our approach on the recently introduced WebVid-CoVR dataset [43]. The training set of WebVid-CoVR consists of triplets (input video, change text, and target video) and is generated synthetically, whereas the test set is manually curated using the model in the loop. The change text within a triplet is generated by comparing captions of the input and target videos using an LLM. It represents the differences between the input and target videos. The WebVid-CoVR training set consists of 131K distinct videos and 467K distinct change texts. One video is associated with each of the 12.7 triplets and the average change text length is 4.8 words. WebVid-CoVR also includes validation and test sets gathered from the WebVid10M corpus. In the validation set there are 7K triplets, whereas in the test set there are 3.2K triplets that have been manually curated to ensure high quality.

Datasets for Composed Image Retrieval (CoIR): We use CIRR [31] and FashionIQ [48] benchmarks for composed image retrieval. CIRR [31] consists of manually annotated open-domain natural image and change text pairs with (36.5K, 19K) distinct pairs. The data distribution of this image and change text pairs is around (28.2K,16.7K), (41.8K, 22.6K), (41.5K, 21.8K) for training, testing, and validation set, respectively. The FashionIQ [48] dataset consists of images of fashion products in three categories: Shirts, Dresses, and Tops/Tees. The reference query and target image are paired based on their category. The corresponding change text is manually annotated. This dataset consists of (30K, 40.5K) images and change text pairs queries annotated on 40.5K distinct images. The data distribution of this image and change text pairs is around (18K, 45.5K), (60.2K, 15.4K) for training, testing, and validation, respectively.

Evaluation Metrics: We follow standard evaluation protocol for the composed image as well video retrieval from [31, 43]. We report the retrieval results using recall values at rank 1, 5, 10, 50. Recall at rank k (R@k) denotes the number of times the correct retrieval occurred among the top-k results.

Implementation Details: We use a multi-modal conversational model [54] to generate the visual descriptions. As discussed earlier, we built our approach on the recent CoVR-BLIP [43] and use the same components without adding any *additional parameters*. We use ViT-L [11] as the frozen vision encoder g, which is pretrained for text-image retrieval on COCO [27]. The frozen text encoder e is from BLIP-2 [26] without cross-attention with pretrained weights of BERT$_{base}$ [9]. We train our model for 20 epochs with a batch size of 1024 (256 batch size per device) with an initial learning rate of $1e-5$. For a fair comparison, we report the


Businessman working with tablet and documents in the office.

The image shows a **businessman sitting** at a **desk** in a small office. He is wearing a **black suit** and **tie**, and has a **laptop** on the **desk** in front of him. He appears to be **working** on something, as there is a **pen** and **paper** in front of him **on the desk**. The background of the image shows a dark, empty room.......

Time lapse of clouds over a beach.

replace the beach with a mountain
replace the beach with a mountain

1. The image is a view of the **ocean** from a **rocky beach**.2. The image is a **cloudy sky** with some **white clouds** in the **background**.3.The **water is calm** and the **waves are gently** lapping against the **shore**.5. There are a few **rocks** and **pebbles** on the **beach**.....

Jet airplane braking after landing in dusseldorf airport.

turn it into a turboprop airplane
turn it into a turboprop airplane

This image shows an **airplane taxiing** on the **runway** at an **airport**. The plane appears to be a **commercial airliner** with a white livery and blue markings. It is **taxiing towards** the **terminal** building, which is visible in the background behind the **fence**.....

Figure 4. Qualitative Comparison between default WebVid-CoVR short captions (top row) with our generated detailed descriptions (bottom row) within our framework. The change text is highlighted in red and the text (default short captions in top row, detailed description in bottom row) are highlighted in black. Here in all three examples from CovR-Vid testset, we observe that the default WebVid-CoVR short captions struggle to fully preserve the contextual information in the retrieved target video (top row). In comparison, our approach leveraging detailed descriptions is able to correctly retrieve the target video with most relevant contextual match with reference video (bottom row). For instance, keeping person working while putting him beside garden in *video-1*, keeping the sea while replacing the beach with mountains in *video-2* and keeping the turboprop airplane on airport behind fence in *video-3*. Best viewed zoomed in. Additional examples are in the suppl.

results of our baseline CoVR-BLIP [43] in the same settings. For transfer learning on CoIR, we fine-tuned the model on the FashionIQ dataset for 6 epochs. We use a batch size of 2048/1024 and an initial learning rate of $1e-4$. After training, our learnable parameters λ, μ and δ for scaling the weightage of each loss was optimized based on validation set with values 0.83, 0.08 and 0.07. We use four NVIDIA A100 GPUS for all the experiments.

4.2. Composed Video Retrieval

Baseline Comparison: We present a baseline comparison in Tab. 1. Compared to the baseline CoVR-BLIP [43], our approach achieves consistent improvement in performance across different recall rates. Without training on WebVid-CoVR benchmark, our approach achieves a significant gain of 5.0% in terms of Recall@K=1 and 13.8% in terms of Recall@K=50. When conducting training on WebVid-CoVR dataset, our approach achieves a significant improvement of 7% in terms of Recall@K=1.

Ablation Study: We first analyze the *impact of inputs* on CoVR performance. Here, we train our model as described in Sec. 3. We freeze our model and study the effect of inputs: reference video, descriptions, and change text with their different combinations for CoVR during inference (Tab. 2) on WebVid-CoVR test set. The performance increases as we replace the input from change text to descriptions to the reference video. As soon as the video and descriptions are provided, the performance improves. This shows that the detailed descriptions provide additional information that complements the input video. Since modification text is not part of the input, the model did not have instructions regarding how to change the composition of the input video and behaves as a plain retrieval task. Further, we obtain superior results after providing modification text along with the detailed descriptions and input video.

Table 2. **The impact of Inputs on our model performance on WebVid-CoVR testset.** The best performance is obtained when using all inputs (video, detailed description and change text), indicating the complementary nature of videos and their detailed language descriptions. Best results in bold.

Input			Recall@K			
Video (q)	description (d)	change text (t)	R@1	R@5	R@10	R@50
✗	✗	✓	26.95	51.25	62.19	83.24
✗	✓	✗	39.92	65.62	76.21	92.23
✓	✗	✗	40.51	66.56	76.95	92.66
✓	✓	✗	42.19	69.06	78.95	95.0
✗	✓	✓	45.51	73.12	82.7	95.78
✓	✗	✓	56.26	81.46	88.97	98.0
✓	✓	✓	**60.12**	**84.32**	**91.27**	**98.72**

Table 3. **The impact of target datasets on our model performance on WebVid-CoVR test set.** We study the impact of different configurations of losses ($\mathcal{L}_{ve}$, $\mathcal{L}_{mme}$, $\mathcal{L}_{te}$) during training. The best results are obtained with all loss terms, indicating the importance of diversity in target datasets. Best results are in bold.

Training Loss			Recall@K			
$\mathcal{L}_{ve}$	$\mathcal{L}_{mme}$	$\mathcal{L}_{te}$	R@1	R@5	R@10	R@50
✗	✗	✓	33.79	65.47	77.93	94.65
✗	✓	✗	42.97	69.49	78.71	93.2
✓	✗	✗	58.37	83.72	89.79	98.16
✗	✓	✓	58.94	83.86	89.76	98.64
✓	✓	✗	59.12	84.18	90.36	98.54
✓	✓	✓	**60.12**	**84.32**	**91.27**	**98.72**

Next, we study in Tab. 3 the *effect of different target datasets*: visual embeddings, multi-modal embeddings, and text-only embedding as explained in Sec. 3.3. This implies that we use any or different combination of the three contrastive losses introduced in Eq. (4) ($\mathcal{L}_{ve}$, $\mathcal{L}_{mme}$, $\mathcal{L}_{te}$). We use all three inputs at inference for CoVR. As a result of training only with $\mathcal{L}_{te}$, we observe an improvement indicating that $\mathcal{L}_{te}$ plays a role in refining our joint multi-modal embedding. Introducing $\mathcal{L}_{mme}$ as a training loss further enhances recall rates, compared to $\mathcal{L}_{te}$, emphasizing its positive impact on the model's ability to retrieve relevant instances. When $\mathcal{L}_{ve}$ is employed as a training loss, the performance

Table 4. **State-of-the-art comparison on CIRR test set**. Our approach achieves consistent improvement in performance on both transfer learning (row 1-12) and zero-shot settings (row 13-20). On the challenging zero-shot setting and using the same pretraining WebVid-CoVR dataset, our approach achieves an absolute gain of 1.6% in terms of Recall@K=1 over [43]. Best results are in bold.

	Method	Pretrain Data	Recall@K K=1	Recall@K K=10	Recall@K K=50	R_{subset}@K K=1	R_{subset}@K K=3
Train CIRR	TIRG [44]	-	14.61	64.08	90.03	22.67	65.14
	MAAF-RP [10]	-	10.22	48.68	81.84	21.41	61.60
	ARTEMIS [8]	-	16.96	61.31	87.73	39.99	75.67
	CIRPLANT [31]	-	19.55	68.39	92.38	39.20	79.49
	LF-BLIP [2, 24]	-	20.89	61.16	83.71	50.22	86.82
	CompoDiff [17]	✓	22.35	73.41	91.77	35.84	76.60
	Combiner [2]	-	33.59	77.35	95.21	62.39	92.02
	CASE [24]	✓	49.35	88.75	97.47	76.48	95.71
	CoVR-BLIP [43]	-	48.84	86.10	94.19	75.78	92.80
	Ours	-	49.18	87.06	94.72	75.66	93.16
	CoVR-BLIP [43]	✓	49.69	86.77	94.31	75.01	93.16
	Ours	✓	**51.03**	**88.93**	**97.53**	**76.51**	**95.76**
Zero Shot	Random†	-	0.04	0.44	2.18	16.67	50.00
	CompoDiff [17]	✓	19.37	72.02	90.85	28.96	67.03
	Pic2Word [39]	✓	23.90	65.30	87.80	-	-
	CASE [24]	✓	35.40	78.53	94.63	64.29	91.61
	CoVR-BLIP [43]	-	19.76	50.89	71.64	63.04	89.37
	Ours	-	21.34	52.37	74.92	64.66	90.87
	CoVR-BLIP [43]	✓	38.48	77.25	91.47	69.28	91.11
	Ours	✓	**40.12**	**78.86**	**94.69**	**70.47**	**92.12**

Table 5. **State-of-the-art comparison on FashionIQ val. set**. Our method obtains favorable results on both transfer learning (row 1-19) and zero-shot (20-25) settings. On the challenging zero-shot setting and using same pretraining, our method obtains an absolute gain of 2.6% (average over three classes: *Dress*, *Shirt* and *Toptee*) in terms of Recall@R=10 over [43]. Best results are in bold.

	Method	Pretrain Data	Dress R@10	Dress R@50	Shirt R@10	Shirt R@50	Toptee R@10	Toptee R@50
Train FashionIQ	JVSM [4]	-	10.70	25.90	12.00	27.10	13.00	26.90
	CIRPLANT [31]	-	17.45	40.41	17.53	38.81	61.64	45.38
	TRACE [19]	-	22.70	44.91	20.80	40.80	24.22	49.80
	VAL w/GloVe [5]	-	22.53	44.00	22.38	44.15	27.53	51.68
	MAAF [10]	-	23.80	48.60	21.30	44.20	27.90	53.60
	CurlingNet [53]	-	26.15	53.24	21.45	44.56	30.12	55.23
	RTIC-GCN [42]	-	29.15	54.04	23.79	47.25	31.61	57.98
	CoSMo[23]	-	25.64	50.30	24.90	49.18	29.21	57.46
	ARTEMIS[8]	-	27.16	52.40	21.78	43.64	29.20	53.83
	DCNet[22]	-	28.95	56.07	23.95	47.30	30.44	58.29
	SAC [20]	-	26.52	51.01	28.02	51.86	32.70	61.23
	FashionVLP[15]	-	32.42	60.29	31.89	58.44	38.51	68.79
	LF-BLIP [2, 24]	-	25.31	44.05	25.39	43.57	26.54	44.48
	CASE [24]	✓	**47.44**	69.36	48.48	**70.23**	50.18	72.24
	CoVR-BLIP [43]	-	43.51	67.94	48.28	66.68	51.53	73.60
	Ours	-	44.39	68.86	49.17	67.54	52.47	74.28
	CoVR-BLIP [43]	✓	44.55	69.03	48.43	67.42	52.60	74.31
	Ours	✓	46.12	**69.52**	**49.61**	68.88	**53.79**	**74.74**
Zero Shot	Random	-	0.26	1.31	0.16	0.79	0.19	0.95
	Pic2Word [39]	✓	20.00	40.20	26.20	43.60	27.90	47.40
	CoVR-BLIP [43]	-	13.48	31.96	16.68	30.67	17.84	35.68
	Ours	-	15.24	34.12	18.36	32.54	19.56	37.54
	CoVR-BLIP [43]	✓	21.95	39.05	30.37	46.12	30.78	48.73
	Ours	✓	**24.57**	**40.93**	**33.12**	**48.42**	**33.16**	**50.24**

improves across all metrics. Similarly, the combination of these losses gives further improvement in performance, indicating their complementing nature.

Lastly, we analyze impact of *detailed language descriptions* on CoVR. We train our model as described in Sec. 3 and freeze our model to study impact of video description quality for CoVR. We compare our detailed description with short WebVid captions (Tab. 6) and observe that the performance is inferior when using default WebVid captions, likely because they do not provide richer context compared to detailed descriptions. Fig. 4 further shows the comparison on example video samples from WebVid-CoVR test set.

Table 6. **The impact of detailed descriptions on our model performance on WebVid-CoVR test set.** Our approach leveraging detailed descriptions achieves consistent improvement in performance, compared to the default captions. Best results are in bold.

Our Model	R@1	R@5	R@10	R@50
using webvid captions	58.23	83.31	90.08	98.05
using our detailed descriptions	**60.12**	**84.32**	**91.27**	**98.72**

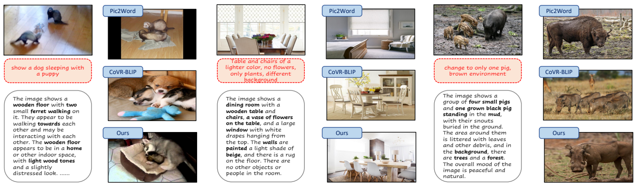

Pic2Word
CoVR-BLIP
Ours
show a dog sleeping with a puppy
The image shows a wooden floor with two small ferret walking on it. They appear to be walking towards each other and may be interacting with each other. The wooden floor appears to be in a home or other indoor space, with light wood tones and a slightly distressed look. ......
Table and chairs of a lighter color, no flowers, only plants, different background
The image shows a dining room with a wooden table and chairs, a vase of flowers on the table, and a large window with white drapes hanging from the top. The walls are painted a light shade of beige, and there is a rug on the floor. There are no other objects or people in the room.
change to only one pig, brown environment
The image shows a group of four small pigs and one grown black pig standing in the mud, with their snouts buried in the ground. The area around them is littered with leaves and other debris, and in the background, there are trees and a forest. The overall mood of the image is peaceful and natural.

Figure 5. Qualitative Comparison between Pic2Word [39](top row), CoVR-BLIP [43] (mid-row) and our proposed method (bottom-row) in zero-shot CoIR task. Here in all three examples from CIRR test set, we observe that using only reference image and change text (in red) Pic2Word [39] and CoVR-BLIP [43] struggle to correctly retrieved target video (top and mid row). In comparison, our approach leveraging detailed descriptions is accurately retrieving the target video with most relevant contextual match with reference video (bottom row). Best viewed zoomed in.

4.3. Composed Image Retrieval

We present state-of-the-art comparison in Tab. 4 and Tab. 5 on CIRR [31] and FashionIQ [48], respectively. Here, the target embeddings are w.r.t a single image. We report results in two settings: zero-shot and transfer learning on respective datasets. In zero-shot setting, we use our model trained on WebVid-CoVR and directly apply it to these respective benchmarks. In both cases, our approach achieves superior performance. Fig. 5 presents a qualitative comparison with existing works on example samples from CIRR [31] test set.

5. Conclusion

We propose an approach that effectively contributes to CoVR and CoIR tasks by integrating detailed visual descriptions. The descriptions are generated from advanced vision-language conversational models with relative change text and visual features aiding to bridge a critical gap in the retrieval process. The enhanced contextual understanding and richer content interpretation offered by our approach leads to superior performance on multiple datasets.

Acknowledgement The computations were enabled by resources provided by NAISS at Alvis partially funded by Swedish Research Council through grant agreement no. 2022-06725, LUMI hosted by CSC (Finland) and LUMI consortium, and by Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the NSC. The work is supported by Swedish Research Council grant 2022-04266.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 5

[2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21466–21474, 2022. 8

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[4] Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 136–152. Springer, 2020. 8

[5] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3001–3011, 2020. 8

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 5

[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 539–546. IEEE, 2005. 2

[8] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. *arXiv preprint arXiv:2203.08101*, 2022. 8

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 6

[10] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020. 8

[11] A Dosovitskiy, L Beyer, A Kolesnikov, D Weissenborn, X Zhai, and T Unterthiner. Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[12] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 2

[13] Zijian Gao, Jingyu Liu, Sheng Chen, Dedan Chang, Hao Zhang, and Jinwei Yuan. Clip2tv: An empirical study on transformer-based methods for video-text retrieval. *arXiv preprint arXiv:2111.05610*, 1(2):6, 2021.

[14] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176, 2022. 2

[15] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14105–14115, 2022. 8

[16] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 241–257. Springer, 2016. 2

[17] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023. 8

[18] James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 2

[19] Surgan Jandial, Ayush Chopra, Pinkesh Badjatiya, Pranit Chawla, Mausoom Sarkar, and Balaji Krishnamurthy. Trace: Transform aggregate and compose visiolinguistic representations for image search with text feedback. *arXiv preprint arXiv:2009.01485*, 7:7, 2020. 8

[20] Surgan Jandial, Pinkesh Badjatiya, Pranit Chawla, Ayush Chopra, Mausoom Sarkar, and Balaji Krishnamurthy. Sac: Semantic attention composition for text-conditioned image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4021–4030, 2022. 8

[21] Lu Jiang, Shoou-I Yu, Deyu Meng, Teruko Mitamura, and Alexander G Hauptmann. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 27–34, 2015. 2

[22] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1771–1779, 2021. 8

[23] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 802–812, 2021. 8

[24] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and early fusion for composed image retrieval. *arXiv preprint arXiv:2303.09429*, 2023. 8

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 5

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen

image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3, 4, 5, 6

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[28] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5007–5015, 2015. 2

[29] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European Conference on Computer Vision*, pages 319–335. Springer, 2022. 2

[30] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 2

[31] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021. 3, 6, 8

[32] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2

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

[34] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015. 2

[35] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 3–20. Springer, 2016. 2

[36] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6967–6977, 2023. 3, 5

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 5

[38] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554, 2023. 2

[39] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314, 2023. 8

[40] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 2

[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[42] Minchul Shin, Yoonjae Cho, Byungsoo Ko, and Geonmo Gu. Rtic: Residual learning for text and image composition using graph convolutional network. *arXiv preprint arXiv:2104.03015*, 2021. 8

[43] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. CoVR: Learning composed video retrieval from web video captions. *AAAI*, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[44] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448, 2019. 8

[45] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448, 2019. 2

[46] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1386–1393, 2014. 2

[47] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016. 2

[48] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021. 6, 8

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

[50] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-

trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022.

[51] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11562–11572, 2021.

[52] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2

[53] Youngjae Yu, Seunghwan Lee, Yuncheol Choi, and Gunhee Kim. Curlingnet: Compositional learning between images and text for fashion iq data. *arXiv preprint arXiv:2003.12299*, 2020. 8

[54] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 5, 6