

Beyond Simple Edits: Composed Video Retrieval with Dense Modifications

Omkar Thawakar^{1*} Dmitry Demidov^{1*} Ritesh Thawkar¹ Rao Muhammad Anwer¹
Mubarak Shah² Fahad Shahbaz Khan^{1,3} Salman Khan^{1,4}

¹Mohamed bin Zayed University of AI, ²University of Central Florida,

³Linköping University, ⁴Australian National University

Abstract

Composed video retrieval is a challenging task that strives to retrieve a target video based on a query video and a textual description detailing specific modifications. Standard retrieval frameworks typically struggle to handle the complexity of fine-grained compositional queries and variations in temporal understanding limiting their retrieval ability in the fine-grained setting. To address this issue, we introduce a novel dataset that captures both fine-grained and composed actions across diverse video segments, enabling more detailed compositional changes in retrieved video content. The proposed dataset, named Dense-WebVid-CoVR, consists of 1.6 million samples with dense modification text that is around seven times more than its existing counterpart. We further develop a new model that integrates visual and textual information through Cross-Attention (CA) fusion using grounded text encoder, enabling precise alignment between dense query modifications and target videos. The proposed model achieves state-of-the-art results surpassing existing methods on all metrics. Notably, it achieves 71.3% Recall@1 in visual+text setting and outperforms the state-of-the-art by 3.4%, highlighting its efficacy in terms of leveraging detailed video descriptions and dense modification texts. Our proposed dataset, code, and model are available at : <https://github.com/OmkarThawakar/BSE-CoVR>.

1. Introduction

Composed Image Retrieval (CoIR) aims at retrieving an image from a database based on a reference image and a textual description detailing the desired modifications. Recently, Ventura et al. [30] extend this problem to the video domain, giving rise to Composed Video Retrieval (CoVR), where the aim is to retrieve a target video based on a reference video and a modification text. Fine-grained CoVR task strives to retrieve a target video based on a reference

video and a detailed textual modification, requiring models to capture subtle visual and temporal changes. For example, in video editing and media production, professionals often require systems that can retrieve content with subtle variations in scene composition or actor actions, aiding in locating footage that meets specific creative requirements. As videos grow longer and more complex, users are likely interested in searching for specific moments or actions rather than an entire clip. As a result, fine-grained CoVR task requires a deeper understanding of both the visual and textual information to ensure the modifications described are accurately reflected in the retrieved video.

Fine-grained CoVR task poses unique challenges by requiring the retrieval model to comprehend the visual content as well as the intricate temporal sequences and semantic nuances of textual modifications. To this end, existing CoVR benchmarks such as, WebVid-CoVR [30] are insufficient due to lack of granularity, short or generic modification text, and imprecise temporal understanding, which is crucial for capturing subtle visual and temporal modifications (see Fig. 1). For instance, in Fig. 1 example 2, WebVid-CoVR benchmark's difference "as a child" is insufficient to retrieve video with a "young child playing piano with his instructor".

A robust CoVR benchmark must capture subtle yet meaningful differences between input and target videos using context-rich modification texts. While prior benchmarks, such as EgoCVR [12], focus on specific aspects like temporal changes, they lack dense modification text and dataset diversity, being limited to egocentric videos. These limitations highlight the need for a comprehensive benchmark that effectively encodes visual, semantic, and temporal modifications across a broad range of video content, including lifestyle, nature, sports, and educational domains.

We introduce Dense-WebVid-CoVR, a benchmark designed to enhance retrieval accuracy by leveraging detailed modification texts (see Fig. 1). Our dataset is not merely about lengthening modification text, but about providing richer, more contextually grounded descriptions that help retrieval models distinguish subtle yet important changes

*Equal Contribution

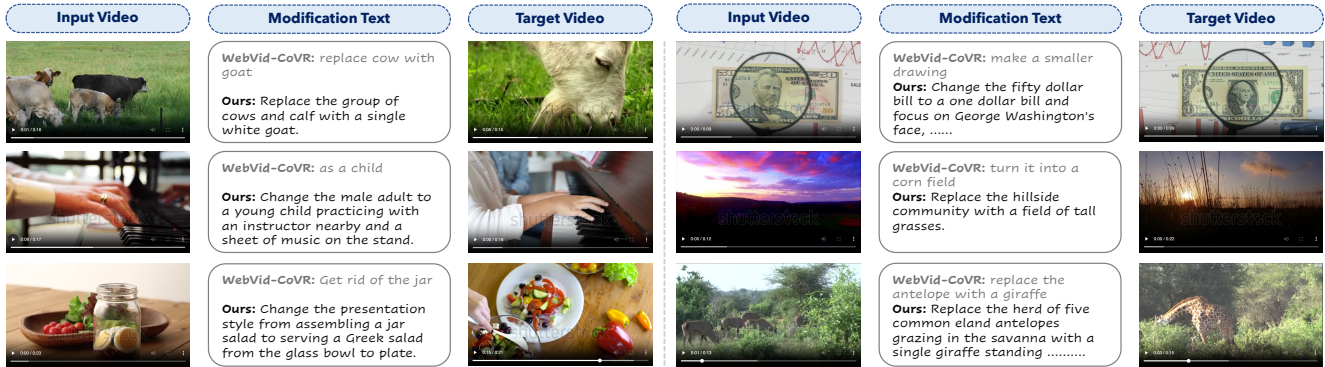


Figure 1. Example composed video retrieval triplets, consisting of the input video, modification text, and the corresponding target video. We compare the basic change text from the existing WebVid-CoVR benchmark [30] with our Dense-WebVid-CoVR dataset that provides a more detailed and context-aware modification text. Additional examples are in the suppl. material.

in videos. Unlike traditional text-to-video retrieval, composed retrieval requires understanding how a target video differs from a reference, making detailed modification texts essential for capturing spatial transformations, object manipulations, and temporal transitions. Dense-WebVid-CoVR is constructed through a two-step process: (1) Detailed video descriptions are generated using Gemini-Pro [28], ensuring high-quality, context-aware textual representation. (2) Source-Modification_Text-Target triplets are then created using GPT-4o [13], which refines video pair relationships with dense, structured modifications. A manual verification step further enhances annotation quality, ensuring that modification texts remain precise, avoid redundancy with the input video, and truly require multimodal understanding. For example, in Fig. 1, our modification text does not simply describe the target video in isolation but explicitly encodes contextual relationships with the input video. In Example 2, rather than a generic phrase like "child playing the piano", our benchmark guides the retrieval system by specifying: "change the male adult to a young child practicing piano with an instructor and a sheet of music on the stand." To further mitigate the risk of text-only retrieval overshadowing multimodal learning, our benchmark introduces query structuring strategies that prevent the modification text from being a direct target description. Building on our Dense-WebVid-CoVR dataset, we develop a new fine-grained CoVR model that effectively encodes the relationships between visual and textual data by fusing input video, description and dense modification text in a single grounding encoder. Our contributions are summarized as:

- We introduce a large-scale fine-grained CoVR dataset, named Dense-WebVid-CoVR, with enriched modification text to encode subtle visual and temporal changes. The dataset comprises 1.6 million samples with an average description length of 81 words and modification text of 31 words, which is around seven times more than the existing CoVR dataset [30]. The test set is fully manually verified to ensure high-annotation quality.

- We further develop a robust CoVR model, leveraging our Dense-WebVid-CoVR dataset, that effectively utilizes the rich text modifications together with input video and textual descriptions in a single grounding encoder. On the Dense-WebVid-CoVR test set, our model achieves a gain of 3.4% over the best existing method [29] when using the same training set, input modalities, and backbone.

2. Related Work

Composed Image Retrieval (CoIR): The task aims to retrieve images based on a reference image and a modification text describing desired changes [32]. Earlier methods rely on manually annotated datasets, e.g., CIRR [22] and FashionIQ [33], which are of high-quality but limited in scale due to the labor-intensive annotation process. More recent approaches aim to scale the task by automatically generating large datasets. Recently, large-scale datasets like LaSCo [20] and SynthTriplets18M [11] have been generated automatically, using visual question answering and text-conditioned image editing frameworks [4]. However, these datasets are not yet publicly available.

Composed Video Retrieval (CoVR): The CoVR task aims to retrieve target videos based on reference videos and textual modification prompts. Recent CoVR methods adapt CoIR techniques to video domain by aggregating multi-frame features [25, 34–36]. Large datasets like WebVid-CoVR [30] and vision-language models (VLMs) [21, 24] have advanced CoVR using contrastive learning and multi-modal embeddings. WebVid-CoVR [30] is constructed by mining video-pairs with similar captions and generating modification texts using large language models (LLMs). Recently, [29] proposes an approach that utilizes language descriptions of source to improve query-specific alignment between the source and target videos. The work of [12] introduces an action retrieval benchmark from egocentric videos. However, the benchmark only has a test set, does not capture dense modification text and is limited to egocentric videos. The benchmark also lacks diversity in terms of video con-

Benchmark	Venue	Type	# Samples	Splits	Data Type	Tools Used	Human Verification	Avg Length of Description	Avg Length of Modification Text	Dense Modification Text	Dense Descriptions	Diversity
InstructPix2Pix	CVPR-2023	Image	454K	train, test	synthetic	GPT-3, Stable Diffusion	None	-	9.4	×	×	✓
CIRR	ICCV-2021	Image	36K	train, test	real-world	AMT	test	-	11.3	×	×	✓
FashionIQ	ICCV-2019W	Image	60K	train, test	synthetic	-	-	-	5.3	×	×	×
CIRCO	ICCV-2023	Image	4.5K	test	synthetic	SEARLE	-	-	8.2	×	×	✓
WebVid-CoVR	AAAI-2024	Video	1.6M	train, test	real-world	MTG-LLM	test	6.68	4.6	×	×	✓
Ego-CVR	ECCV-2024	Video	2.2K	test	real-world	GPT-4	test	-	4	×	×	×
Dense-WebVid-CoV (Ours)	-	Video	1.6M	train, test	real-world	Gemini-Pro, GPT-4	train, test	81.32	31.16	✓	✓	✓

Table 1. **Comparative analysis of various video and image-based benchmarks for composed video retrieval (CoVR).** The benchmarks are categorized by type (video or image), sample size, data type (real-world or synthetic), tools used for generation, and human verification. Compared to existing CoVR datasets, our benchmark provides fine-grained dense descriptions with an average description length of 81.32 words and modification texts of 31.16 words, surpassing them in generating rich, context-aware video retrieval capabilities.

tent (e.g., general purpose videos including, nature, sports, educational visual content and lifestyle), that is available in WebVid-CoVR dataset. In this work, we develop a new benchmark (see Tab. 1) that comprises both training and test set, while containing high-quality detailed modification text about general purpose videos.

3. Dense-WebVid-CoVR Benchmark

To construct a fine-grained CoVR benchmark with detailed modification text, we start with the WebVid-CoVR dataset [30] that contains diverse video content including nature, lifestyle, and professional activities. The dataset contains 1.6 million triplets with videos averaging around 16.8 seconds in length. To effectively create captions for videos, we employ the Gemini-Pro [28] model for video captioning. To ensure caption quality, we apply a hallucination check using the BLIP model [21] to compute the cosine similarity between the video and its caption. Here, captions scoring below a specified threshold that is set empirically are deemed inadequate and recomputed, ensuring alignment with video content and creating a more robust and reliable dataset. Additional details are presented in suppl. material.

3.1. Modification-Text Generation

In the context of fine-grained robust CoVR, modification text is crucial and bridges the gap between two similar videos by explicitly describing the differences between them. These descriptions highlight specific changes, such as alterations in actions, objects, or scenes. Dense modifications are expected to cover nuanced variations in visual content to enable more precise retrieval of videos based on subtle details. For this, we employ GPT-4o [13] to generate modification texts. We provide the model with existing triplets from WebVid, including both the original video captions and their corresponding modification texts, to guide the generation process. Fig. 2 shows the differences between WebVid descriptions and our dense descriptions and detailed modification text. Compared to original WebVid descriptions, subtle changes between videos are captured. For instance Fig. 2 row 1 contains a outdoor nature video. The original WebVid modification text (*put a grassland background*) lacks subtle details. In contrast, our modification text (*Add a serene outdoor scene with*

a lone tree on a grassy hill, and make the camera static to capture the subtle movements of the tree and clouds, evoking tranquility and the beauty of nature better captures) better captures detailed information. The examples in Fig. 2 shows that our modification text provides more details in the form of intricate visual elements, such as environmental settings, colors, lighting, and subtle changes in object focus.

3.2. Modification-Text Verification

To ensure high-quality, we manually verify the generated modification texts. During this process, input and target videos are presented side-by-side along with their generated dense descriptions. The annotators are tasked with assessing the quality of the modification text by comparing it to the visual content and making corrections when necessary. To ensure the quality and accuracy of the generated modification text, the following quality control protocol is used for verification process. (i) A side-by-side comparison to ensure that the changes mentioned accurately reflect the differences between the two videos. (ii) A contextual consistency check to verify that the modification text addresses key changes, such as consistent object movement and transitions in the main scene, related surroundings, and background. (iii) An action and object verification check, where the annotators are asked to check whether the objects and their corresponding actions mentioned in the modification text are present in both the videos. (iv) A temporal alignment check to ensure that the modification text aligns with the actual sequence of actions in the videos. (v) The annotators are asked to check the comprehensive description quality to ensure all relevant changes between the input and target video are covered. (vi) The annotators are asked to ensure that the modification text is clear and concise. (vii) We empirically set a cosine similarity threshold to identify low-quality modification texts for further manual modification. Moreover, it is worth noting that annotators are asked to make manual corrections to the modification text in case of any errors or missing details.

We manually verify *all* 3,000 triplets each from the test set and validation set consisting of WebVid-8M [2] corpus videos. For training set, out of the total of 1.6M triplets consisting WebVid-2M [2] corpus, we carefully select 100k triplets with unique input and target video covering distinct categories representing all training triplets for verification.




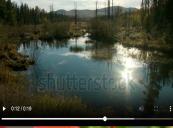


Input (Video-1)	Target (Video-2)	WebVid Description (Video-1)	WebVid Description (Video-2)	WebVid Modification Text	Our Description (Video-1)	Our Description (Video-2)	Our Modification Text
		Clouds and blue sky.	Grassland and blue sky	put a grassland background	The video shows a timelapse of fluffy white cumulus clouds moving across a blue sky. The clouds appear to be moving from left to right, expanding and changing shape as they drift. The sunlight illuminates the clouds, creating a sense	The video displays a serene outdoor scene with a lone tree standing on a grassy hill against a backdrop of a bright blue sky dotted with fluffy white clouds. The tree's leaves rustle gently in the wind, creating a sense of tranquility and peace.	Add a serene outdoor scene with a lone tree on a grassy hill, and make the camera static to capture the subtle movements of tree and clouds.
		Autumn day.	Autumn swamp.	make it a swamp.	The video shows a picturesque landscape with rolling hills covered in a mix of green, yellow, and orange autumnal trees. White buildings peek out from behind the hills on the left side. A body of water, likely a pond or	The video showcases a stunning view of a serene marsh. The sunlight shines brightly, casting a glistening reflection upon the tranquil waters. The sky displays a picturesque scene of blue hues adorned with delicate wisps of white clouds,	Change the landscape to a serene marsh, include bright sunlight, vibrant green plants and trees, a blue sky with white clouds emphasize the tranquil and reflective water.
		Red rose closeup.	Red hibiscus closeup.	make hibiscus.	The video displays a single, vibrant red rose in full bloom, positioned prominently in the foreground. The rose remains stationary throughout the video, its petals slightly curled inward. In the background,	The video showcases a vibrant red hibiscus flower in full bloom, its petals glistening under the sun. The camera focuses on the flower's center, highlighting the intricate details of its reproductive organs.....	make hibiscus, highlight its reproductive organs, and add subtle swaying motion.

Figure 2. Comparison between original WebVid-CoVR [30] descriptions and change-text vs. our generated detailed descriptions and change text. Each row presents an input video (Video-1), a target video (Video-2), and their corresponding descriptions followed by the change-text generated using the descriptions. The original WebVid-CoVR’s [30] change texts lack fine-grained details, whereas our approach offers significantly more comprehensive and context-rich change texts.

To further prevent text-only retrieval from overshadowing multimodal learning, we introduce query structuring strategies that ensure modification texts are not direct target descriptions but require contextual interpretation with the input video. This enforces true multimodal reasoning, ensuring that models cannot retrieve the target video based solely on modification text as stated in CIRCO [1]. These structuring strategies play a crucial role in preserving the core purpose of Composed Video Retrieval (CoVR) by preventing the task from being reduced to text-to-video retrieval. To ensure high-quality annotations, trained annotators manually verified and refined modification texts through multiple rounds. Although around 2-3% of modification texts in the training set may have minor inaccuracies, our experiments show that this has minimal impact on model performance. Instead, the inclusion of detailed modification texts significantly improves retrieval accuracy, leading to a 3.4% gain in Recall@1, proving that our dataset remains highly reliable and effective for fine-grained video retrieval. Additional details are presented in the suppl. material (Section E). Next, we introduce our method that leverages the detailed modification text from the Dense-WebVid-CoVR dataset for fine-grained CoVR.

4. Method

Problem Formulation: In the fine-grained CoVR task, the objective is to retrieve a modified video from a large database using two inputs: a reference video and a detailed textual description that outlines the desired modification. Let \mathcal{V} represent the set of all videos, \mathcal{D} represent the set of corresponding dense descriptions, and \mathcal{T} the space of detailed textual modifications. Given a query video $q \in \mathcal{V}$, its description $d \in \mathcal{D}$ and a corresponding detailed textual modification $t \in \mathcal{T}$, the goal is to identify the target video $v^* \in \mathcal{V}$ that best reflects the described changes. The retrieval system is

desired to leverage both visual and textual embeddings to capture semantic changes, ensuring fine-grained, context-sensitive video retrieval.

4.1. Overall Architecture

For the fine-grained CoVR task, it is desired to accurately encode fine-grained subtle visual and temporal changes by capturing the inter-dependencies between the query video (q), dense detailed description (d), and the modification text (t). Recent methods, such as [29] utilizes a pairwise fusion scheme (see Fig. 4a) that processes each component pair separately (e.g., $f(q, t)$, $f(d, t)$, and $f(q, d)$). We observe such strategy to achieve sub-optimal results likely due to the dilution of the rich modification text details in the final multimodal embedding. To this end, we introduce an approach having a simple yet effective fusion strategy to better align the multimodal query and target videos. Our architecture (see Fig. 3) comprises three components: vision encoder (g), text encoder (e), and grounding text encoder (f). The goal is to retrieve a target video based on a query video (q), a detailed description of the query (d), and a textual modification (t) that describes the desired changes.

Vision Encoder: The Vision Encoder (g) utilizes ViT-L [9] as its backbone and processes the visual input. Instead of processing every frame, the middle frame of the input video is selected to compute the visual embeddings for efficient feature extraction, following [29, 30].

Text Encoder: The text encoder (e), pretrained from the BLIP, processes the detailed textual description (d) that accompanies each video. The comprehensive description is expected to comprise both spatial aspects and temporal elements to effectively summarize all actions within the video. Consequently, the description embeddings capture all video-level features to obtain a holistic representation of the input video. To obtain alignment between visual and textual fea-

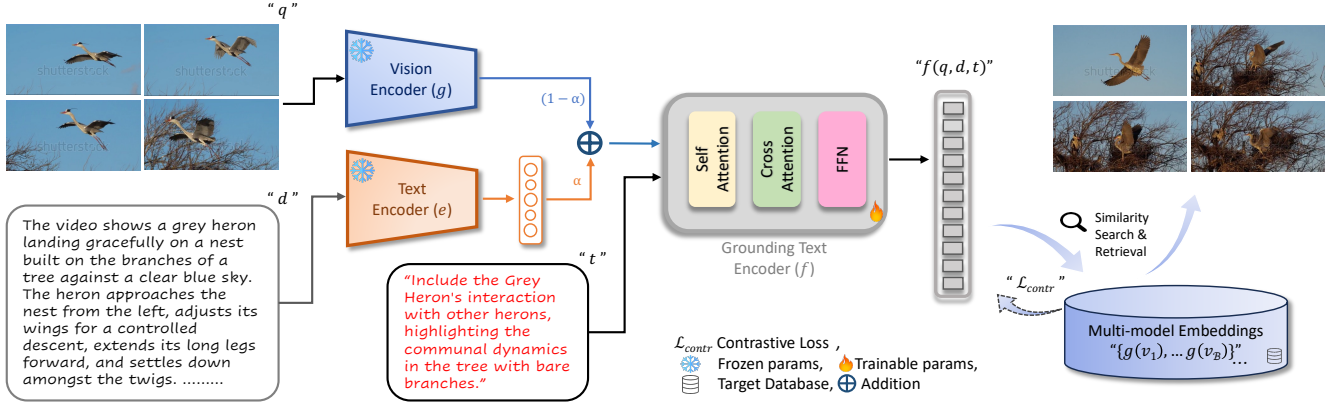


Figure 3. Our proposed fine-grained CoVR architecture comprising three main components: a vision encoder (g), a text encoder (e), and a grounding text encoder (f). The video query (q) and detailed textual description (d) are processed by vision encoder (g) and text encoder (e), with a projection layer aligning their embeddings. Through a unified fusion strategy, the weighted combination of these embeddings is grounded with the modification text (t) in the grounding text encoder (f) using self-attention, cross-attention, and a feed-forward network (FFN). The model is trained with contrastive loss (\mathcal{L}_{contr}), leveraging frozen (g, e) and trainable (f) components to produce multi-modal embeddings for effective video retrieval.

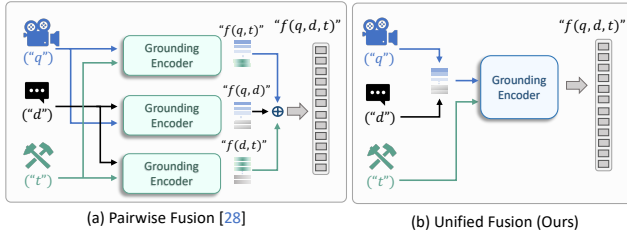


Figure 4. Comparison between the pairwise fusion strategy [29] (left) and our unified fusion scheme (right) for integrating query (q), description (d), and modification text (t) in CoVR.

tures, a projection layer is employed that aligns the text embeddings with the visual embeddings extracted by the vision encoder (g). These aligned embeddings are then integrated using a learnable parameter α , as in:

$$embs = (1 - \alpha)g(q) + \alpha e(d) \quad (1)$$

This enables the model to dynamically balance the impact of visual and textual information (see Fig. 3) based on the optimized α value derived from validation set.

Grounding Text Encoder: The grounding text encoder (f) generates the final composed multimodal embedding. The encoder f fuses the query video and description embeddings with the detailed modification text (t) using a cross-attention mechanism. Cross-attention layers align the visual features from the video with the textual description, grounding the modification in the correct context. The encoder f outputs a fused embedding ($f(q, d, t)$) that encodes the query, description, and modification text in a single representation. This enriched embedding is then compared against embeddings of target videos in a large video database for retrieval.

As discussed above, our approach employs a unified fusion scheme to simultaneously fuse q, d , and t by integrating all three components within a *single* grounding encoder (see

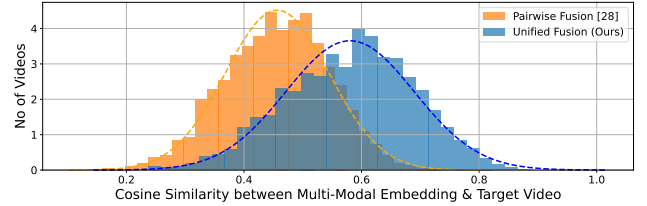


Figure 5. Comparison between the pairwise fusion [29] and our unified fusion in terms of cosine similarity score computed between multimodal query embedding and target video. The results are computed on Dense-WebVid-CoVR validation set. The pairwise embeddings are generated separately, leading to less contextually aligned representations as each pairwise combination is processed separately. In contrast, our method fuses query (q) and description (d) first followed by combination with modification text (t) within a *single* grounding encoder. Our simple yet effective fusion scheme enables richer and granular understanding of the relationships between visual and textual data achieving higher similarity between multi-modal embeddings with target videos.

Fig. 4b). Compared to the pairwise fusion [29], our unified fusion achieves better alignment between multi-modal query and target videos (see Fig. 5). The overall model is trained using a contrastive loss (\mathcal{L}_{contr}), encouraging the alignment of fused embeddings ($f(q, d, t)$) with correct target video embeddings ($g(v)$). The objective is defined as minimizing:

$$\mathcal{L}_{contr} = - \sum_{i \in \mathcal{B}} \log \left(\frac{e^{S_{i,i}/\tau}}{\lambda \cdot e^{S_{i,i}/\tau} + \sum_{j \neq i} e^{S_{i,j}/\tau} w_{i,j}} \right) - \sum_{i \in \mathcal{B}} \log \left(\frac{e^{S_{i,i}/\tau}}{\lambda \cdot e^{S_{i,i}/\tau} + \sum_{j \neq i} e^{S_{j,i}/\tau} w_{j,i}} \right) \quad (2)$$

Here, λ is assigned a value of 1, and the temperature τ is set to 0.07, following [23]. The term $S_{i,j}$ represents the cosine similarity between the joint composed multi-modal embed-

	Model	Training		Modification			Recall@K			
		Dense-WebVid-CoVR	Input Modalities	Text Fusion	Backbone	Frames	R@1	R@5	R@10	R@50
1	Random		-	-	-	-	0.04	0.21	0.32	1.46
2	CoVR-BLIP [30]	✗	Text	-	BLIP	-	24.12	56.02	60.16	82.34
3	CoVR-BLIP [30]	✗	Visual	-	BLIP	15	22.52	53.08	58.34	81.26
4	CoVR-BLIP [30]	✗	Visual + Text	Avg	BLIP	15	38.44	64.96	71.72	87.12
5	Thawakar <i>.et.al.</i> [29]	✗	Visual + Text	Avg	BLIP	15	40.23	66.38	74.84	88.12
6	Our Approach	✗	Visual + Text	Avg	BLIP	15	42.41	68.54	77.07	91.24
7	CoVR-BLIP [30]	✗	Visual + Text	CA	BLIP	15	35.60	60.80	70.31	87.05
8	Thawakar <i>.et.al.</i> [29]	✗	Visual + Text	CA	BLIP	15	39.20	64.40	75.56	88.90
9	Our Approach	✗	Visual + Text	CA	BLIP	15	48.08	73.36	81.06	93.78
10	CoVR-BLIP [30]	✓	Text	-	BLIP	-	38.92	66.36	74.08	90.92
11	Our Approach	✓	Text	-	BLIP	15	50.12	78.36	79.52	92.62
12	CoVR-BLIP [30]	✓	Visual	-	BLIP	15	36.26	64.32	72.18	90.46
13	CoVR-BLIP [30]	✓	Visual + Text	CA	BLIP	15	63.12	85.66	92.56	97.36
14	Thawakar <i>.et.al.</i> [29]	✓	Visual + Text	CA	BLIP	15	67.86	87.72	93.06	98.18
15	Our Approach	✓	Visual + Text	CA	BLIP	15	71.26	89.12	94.56	98.88

Table 2. **Comparison of our approach with existing methods on the Dense-WebVid-CoVR test set.** Our proposed approach consistently outperforms existing methods in *all* settings and Recall@K metrics. Notably in the Visual + Text setting with Cross-Attention (CA), our method improves Recall@1 to 71.26 and Recall@50 to 98.88. Best results are in bold.

ding $f(q, d, t)$ and the corresponding target video $g(v)$. The weight $w_{i,j}$ is configured as in [23], using $\beta = 0.5$, and \mathcal{B} refers to batch size. Finally, a similarity search is performed over video database using the fused multi-modal embeddings. Here, we compare the embedding ($f(q, d, t)$) with embeddings of all target videos ($g(v)$). Videos with highest similarity scores are retrieved as the most likely matches.

5. Experiment

Datasets: We conduct experiments on the proposed Dense-WebVid-CoVR dataset. The training set comprises 131K distinct videos paired with 467K unique change texts, having video descriptions with average words length 81.32 with each video associated with an average of 12.7 triplets, and the change texts averaging 31.2 words in length. The dataset also includes 7K validation triplets and 3.2K manually curated test triplets, ensuring high-quality evaluation sets from the WebVid10M corpus. In addition to Dense-WebVid-CoVR, we conduct experiments on the recent EgoCVR [12] to further evaluate our approach on Ego-Centric videos. EgoCVR consists of 2,295 samples, with 78.9% focusing on temporal events and 21.1% on object-centered modifications, focusing on temporal video understanding.

We also conduct experiments on composed image retrieval (CoIR) task using two standard benchmarks: CIRR [22] and FashionIQ [33]. CIRR contains 36.5K manually annotated open-domain natural image pairs along with their change text. The dataset is split into 28.2K, 16.7K pairs for training, and 41.8K, 22.6K pairs for testing and validation, respectively. FashionIQ focuses on fashion products in three categories such as Shirts, Dresses, and Tops/Tees consisting of 30K images paired with 40.5K change texts. The data distribution includes 18K, 45.5K pairs for training, and 60.2K, 15.4K for testing and validation.

Evaluation Metrics: We follow standard protocols for both

composed video and image retrieval tasks (CoVR and CoIR), as in prior works [12, 22, 30]. Retrieval performance is measured using Recall@K (R@k), where k represents the top-k ranked results. Recall at rank k signifies the percentage of times the correct target is retrieved within the top-k results. Specifically, we report the recall values at ranks 1, 5, 10, and 50 to comprehensively assess the model’s retrieval accuracy across different scenarios.

Implementation Details: We utilize ViT-L [9] as the vision encoder. The text encoder and the grounding text encoder is from BLIP-2 [21], for fusing the modification-text with multimodal features. The model is trained for 5 epochs with a batch size of 1024 (256 per device) and an initial learning rate of $1e - 5$. The value of learnable parameter α derived from validation set is 0.36. For transfer learning on CoIR, we fine-tune on the FashionIQ dataset for 6 epochs using a batch size of 2048/1024 and a learning rate of $1e - 4$. All experiments are run on four NVIDIA A100 40GB GPU’s.

5.1. Results on Composed Video Retrieval (CoVR)

Tab. 2 presents a comparison with existing methods on the proposed Dense-WebVid-CoVR test set with dense modification text under different settings. Our approach achieves consistently improved performance in all settings and metrics. When using the training setting and both input modalities (visual + text), the recent work of Thawakar *.et.al.* [29] obtains Recall@1 and Recall@5 score of 67.9 and 87.7, respectively. Our approach outperforms [29] with Recall@1 and Recall@5 score of 71.3 and 81.1, respectively. In case of no training scenario and cross-attention (CA) based modification text fusion, our approach achieves a larger performance (+8.9% in Recall@1) over [29] likely due to the proposed unified fusion being more effective at leveraging the detailed descriptions. Additionally, Our approach shows impressive performance of Recall@1 of 50.12 with text-based video re-

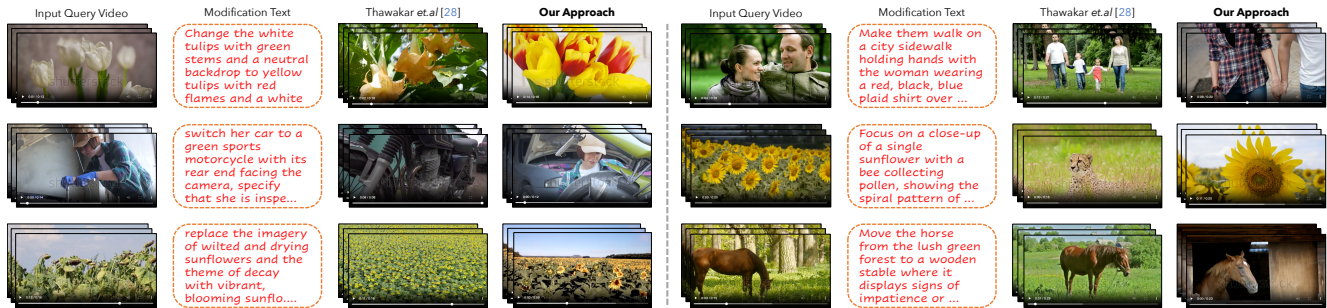


Figure 6. Qualitative comparison between the recent CoVR method [29] and our approach on example videos from the Dense-WebVid-CoVR test set. The approach of [29] based on pairwise fusion misses fine-grained details from the modification text, leading to sub-optimal retrieval performance. For instance in the first example video (row 1 on the left), it retrieves the video with yellow flower but misses the other details of *red flames and a white brick background* in the modification text. Similarly, it misses the fine-grained details in the modification text such as, *specify that she is inspecting ...* in the first example (left) on row 2. Our approach achieves superior retrieval performance by better capturing the fine-grained details and context. Best viewed zoomed in. Additional results are presented in suppl. material.

Method	Global			Local		
	R@1	R@5	R@10	R@1	R@2	R@3
Random	0.01	0.05	0.1	25.3	38.2	50.7
CoVR-BLIP [30]	5.4	15.2	24.3	33.1	49.5	62.9
Thawakar <i>et.al</i> [29]	6.0	14.8	24.3	33.4	49.3	63.0
CIReVL [16]	2.0	6.8	10.2	21.6	35.1	46.0
TFR-CVR [12]	14.1	39.5	54.4	44.2	61.0	73.2
Our Approach	14.6	41.3	54.9	44.8	61.7	74.0

Table 3. Comparison of our method with existing approaches, in a zero-shot setting, on the Ego-CVR test set. Our approach performs favorably in terms of both global and local retrieval metrics (R@K) compared to existing methods. Best results are in bold.

Video Descriptions (d)	Modification Text (t)	R@1	R@5	R@10
WebVid-CoVR	WebVid-CoVR	60.4	84.5	91.4
WebVid-CoVR	Dense-WebVid-CoVR	61.2	84.8	92.6
Dense-WebVid-CoVR	WebVid-CoVR	63.8	87.5	92.4
Dense-WebVid-CoVR	Dense-WebVid-CoVR	71.2	89.1	94.5

Table 4. Ablation study comparing the performance of our proposed model trained with different Video Descriptions (d) and Modification Text (t) combinations. We report Recall@K scores.

retrieval. We further note that our approach is $3\times$ times faster than [29] due to the unified fusion approach that avoids repetition (in contrast to pairwise fusion) and utilizes a single grounding text encoder to construct multimodal embedding.

Fig. 6 presents a qualitative between [29] employing pairwise fusion and our approach based on unified fusion on example videos from Dense-WebVid-CoVR test set. We observe that pairwise fusion-based method [29] that separately process each input component misses fine-grained details within the modification text, leading to sub-optimal retrieval quality. Our approach integrating all input elements within a single grounding encoder better captures the context and fine-grained details within the modification text, leading to superior retrieval performance.

We further conduct experiments on the Ego-CVR dataset in a zero-shot setting. Tab. 3 shows the comparison on Ego-CVR test set. Here, recent methods such as TFR-CVR [12]

uses the pre-trained TFR model that is trained on 10 million corpus of WebVid data for text-to-video retrieval. Our approach performs favorably in terms of global and local retrieval performance, compared to existing works.

Ablation Study: We first perform a study to analyze the impact of using dense modification text for CoVR. Tab. 6 (left) compares the results of using our dense modification texts versus the WebVid-CoVR short modification texts. Our approach consistently outperforms across all recall metrics demonstrating the effectiveness of more detailed modification texts in improving the retrieval accuracy. We further examine the impact of dense descriptions during inference and present the results in Tab. 6 (right). The model utilizing dense descriptions achieves a notable performance gain with a retrieval score of 71.26% Recall@1, compared to 66.08% without dense descriptions.

Table 4 shows that using Dense-WebVid-CoVR modification texts significantly improves retrieval accuracy, with Recall@1 increasing from 63.8 to 71.2 and Recall@5 from 87.5 to 89.1, compared to models trained with WebVid-CoVR modifications. This highlights that shorter modification texts often lead to incorrect target retrieval, causing models to focus on distractor videos similar to the input but lacking the intended changes. In contrast, richer modification texts help models capture subtle transformations more effectively, reinforcing the importance of detailed textual modifications in enhancing multimodal learning and preventing reliance on text-only retrieval.

We conduct a study to understand the impact of different fusion strategies on input video (q) and its description (d) Here, we compared three fusion techniques to fuse visual embedding with description embedding: Addition, Cross-Attention (CA), and the proposed unified (weighted-mean). When using addition strategy in our framework, we achieve Recall@1 of 69.72. The results improve to 70.13 when using CA. The best results of Recall@1 of 71.26 are obtained when using the proposed unified fusion strategy.

Method	Pretrain Data	Recall@K			R _{subset} @K			
		K=1	K=10	K=50	K=1	K=3		
Train CIRR	TIRG [31]	-	14.61	64.08	90.03	22.67	65.14	
	MAAF-RP [8]	-	10.22	48.68	81.84	21.41	61.60	
	ARTEMIS [7]	-	16.96	61.31	87.73	39.99	75.67	
	CIRPLANT [22]	-	19.55	68.39	92.38	39.20	79.49	
	LF-BLIP [3]	-	20.89	61.16	83.71	50.22	86.82	
	CompoDiff [11]	✓	22.35	73.41	91.77	35.84	76.60	
	Combiner [3]	-	33.59	77.35	95.21	62.39	92.02	
	CASE [19]	✓	49.35	88.75	97.47	76.48	95.71	
	CoVR-BLIP [30]	-	48.84	86.10	94.19	75.78	92.80	
	CoVR-BLIP [30]	✓	49.69	86.77	94.31	75.01	93.16	
	Thawakar <i>et al.</i> [29]	✓	51.03	88.93	97.53	76.51	95.76	
	Our Approach	✓	56.30	91.84	98.20	79.16	96.42	
	Zero Shot	Random†	-	0.04	0.44	2.18	16.67	50.00
		CompoDiff [11]	✓	19.37	72.02	90.85	28.96	67.03
Pic2Word [26]		✓	23.90	65.30	87.80	-	-	
CASE [19]		✓	35.40	78.53	94.63	64.29	91.61	
CoVR-BLIP [30]		✓	38.48	77.25	91.47	69.28	91.11	
Thawakar <i>et al.</i> [29]		-	21.34	52.37	74.92	64.66	90.87	
Our Approach		-	32.16	63.34	78.92	68.62	91.06	
Thawakar <i>et al.</i> [29]		✓	40.12	78.86	94.69	70.47	92.12	
Our Approach		✓	44.08	81.72	95.88	74.12	93.18	

Method	Pretrain Data	Dress		Shirt		Toptee		
		R@10	R@50	R@10	R@50	R@10	R@50	
Train FashionIQ	JVSM [5]	-	10.70	25.90	12.00	27.10	13.00	26.90
	CIRPLANT [22]	-	17.45	40.41	17.53	38.81	61.64	45.38
	TRACE [14]	-	22.70	44.91	20.80	40.80	24.22	49.80
	VAL w/GloVe [6]	-	22.53	44.00	22.38	44.15	27.53	51.68
	MAAF [8]	-	23.80	48.60	21.30	44.20	27.90	53.60
	CurlingNet [37]	-	26.15	53.24	21.45	44.56	30.12	55.23
	RTIC-GCN [27]	-	29.15	54.04	23.79	47.25	31.61	57.98
	CoSMo[18]	-	25.64	50.30	24.90	49.18	29.21	57.46
	ARTEMIS[7]	-	27.16	52.40	21.78	43.64	29.20	53.83
	DCNet[17]	-	28.95	56.07	23.95	47.30	30.44	58.29
	SAC [15]	-	26.52	51.01	28.02	51.86	32.70	61.23
	FashionVLP[10]	-	32.42	60.29	31.89	58.44	38.51	68.79
	LF-BLIP [3]	-	25.31	44.05	25.39	43.57	26.54	44.48
	CASE [19]	✓	47.44	69.36	48.48	70.23	50.18	72.24
	CoVR-BLIP [30]	-	43.51	67.94	48.28	66.68	51.53	73.60
	CoVR-BLIP [30]	✓	44.55	69.03	48.43	67.42	52.60	74.31
	Thawakar <i>et al.</i> [29]	✓	46.12	69.52	49.61	68.88	53.79	74.74
	Our Approach	✓	48.12	71.48	51.38	70.38	55.08	75.96
Zero Shot	Random	-	0.26	1.31	0.16	0.79	0.19	0.95
	Pic2Word [26]	✓	20.00	40.20	26.20	43.60	27.90	47.40
	CoVR-BLIP [30]	✓	21.95	39.05	30.37	46.12	30.78	48.73
	Thawakar <i>et al.</i> [29]	-	15.24	34.12	18.36	32.54	19.56	37.54
	Our Approach	✓	21.08	38.26	22.18	36.72	25.06	44.28
	Thawakar <i>et al.</i> [29]	✓	24.57	40.93	33.12	48.42	33.16	50.24
	Our Approach	✓	26.12	42.88	35.32	49.92	35.44	51.66

Table 5. **Left: Comparison of our approach with existing methods on the CIRR [22] test set.** We present the results in both training and zero-shot settings in terms of Recall@K and R@K. Our approach consistently outperforms existing methods, achieving Recall@K=1 score of 44.08 and Recall@K=50 score of 95.88. **Right: Comparison with existing methods on retrieval tasks for specific attributes such as dress, shirt, and toptee on FashionIQ [33] validation set.** Best results are in bold.

Modification-Text	R@1	R@5	R@10	R@50
WebVid-CoVR	68.88	88.62	94.20	98.62
Dense-WebVid-CoVR	71.26	89.12	94.56	98.88

Our Model	R@1	R@5	R@10	R@50
Without using dense descriptions in inference	66.08	88.32	93.82	98.36
With using dense descriptions in inference	71.26	89.12	94.56	98.88

Table 6. **Left: Impact of using the modification text from the original WebVid-CoVR or from our Dense-WebVid-CoVR on the model performance at inference.** Our method achieves best performance when incorporating dense and detailed modification texts from Dense-WebVid-CoVR dataset. Notably, our approach achieves an R@1 of 71.26, surpassing its performance of 68.88 with WebVid-CoVR. **Right: Impact of using dense video descriptions during inference.** When utilizing dense descriptions, our method achieves superior performance across all metrics, achieving an R@1 of 71.26, compared to 66.08% without using dense descriptions. Best results are in bold.

5.2. Results on Composed Image Retrieval (CoIR)

In addition to the CoVR task, we evaluate our approach on composed image retrieval (CoIR) task using both training and zero-shot settings. We conduct experiments on CIRR [22] test set and FashionIQ [33] validation set. The results are presented in Tab. 5. On the left and when using the train CIRR setting, our method achieves Recall@1 score of 56.30, outperforming recent approaches like CoVR-BLIP [30] and Thawakar *et al.* [29] by a significant margin. In the zero-shot setting when the methods are not trained on CIRR data, our approach performs favorably against existing methods with a Recall@1 score of 44.08. These results suggest the generalizability of our method in terms of handling unseen data. We also analyze the performance on specific attribute-based retrieval tasks, as presented in Tab. 5 (right). Our method achieves consistent improvements across various categories. Notably in the dress category, our method obtains R@50 score of 49.92. On the toptee category, our approach achieves R@50 score of 51.66, outperforming existing methods in the zero-Shot setting. These results suggest that our method has the ability to handle attribute-specific composed image retrieval with better precision and accuracy.

6. Conclusion

We investigate the problem of composed video retrieval (CoVR) and propose a new dataset with detailed video descriptions and dense modification texts, capturing subtle visual and temporal changes. Our dataset, named Dense-WebVid-CoVR, comprises 1.6 million samples with dense modification text with an average length of 31.2 words. In addition, we propose an approach that encodes subtle changes by simultaneously processing input video, description, and dense modification text in a single grounding encoder. For CoVR, we conduct experiments on two datasets: Dense-WebVid-CoVR and Ego-CVR. Our approach achieves favorable results on both datasets. We further evaluate our approach for CoIR task on two datasets: CIRR and FashionIQ. Our approach achieves state-of-the-art performance on both datasets. A potential future direction is to explore CoVR task in a multilingual setting, especially for low-resource languages, to expand its real-world applicability to diverse populations. Another potential research direction is to explore efficient techniques for processing very long videos and their corresponding descriptions for the CoVR task.

7. 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.

References

- [1] Lorenzo Agnolucci, Alberto Baldrati, Marco Bertini, and Alberto Del Bimbo. isearle: Improving textual inversion for zero-shot composed image retrieval. *arXiv preprint arXiv:2405.02951*, 2024. 4
- [2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. A clip-hitchhiker’s guide to long video retrieval. *arXiv preprint arXiv:2205.08508*, 2022. 3
- [3] 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
- [4] 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. 2
- [5] 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
- [6] 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
- [7] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csukka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. *arXiv preprint arXiv:2203.08101*, 2022. 8
- [8] 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
- [9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 6
- [10] 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
- [11] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoon Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023. 2, 8
- [12] Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. Egocvr: An egocentric benchmark for fine-grained composed video retrieval. *European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 6, 7
- [13] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 3
- [14] 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
- [15] 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
- [16] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*, 2023. 7
- [17] 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
- [18] 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
- [19] 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
- [20] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and early fusion for composed image retrieval. *arXiv preprint arXiv:2303.09429*, 2(3):7, 2023. 2
- [21] 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. 2, 3, 6
- [22] 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. 2, 6, 8
- [23] 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. 5, 6
- [24] 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 supervi-

- sion. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2
- [25] 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
- [26] 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
- [27] 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
- [28] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 3
- [29] Omkar Thawakar, Muzammal Naseer, Rao Muhammad Anwer, Salman Khan, Michael Felsberg, Mubarak Shah, and Fahad Shahbaz Khan. Composed video retrieval via enriched context and discriminative embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26896–26906, 2024. 2, 4, 5, 6, 7, 8
- [30] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5270–5279, 2024. 1, 2, 3, 4, 6, 7, 8
- [31] 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
- [32] 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
- [33] 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. 2, 6, 8
- [34] 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
- [35] 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.
- [36] 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. 2
- [37] 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