SLERP$^+$: SPHERICAL LINEAR INTERPOLATION FOR UNIFIED COMPOSITIONAL RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot composed image/video retrieval is a challenging task that involves using a combination of a reference visual input and a relative caption as a query to search for target visual data. Earlier studies have treated composed image retrieval and composed video retrieval methods separately, potentially neglecting the benefits of integrating image-video-text representation learning. In this paper, we consolidate these tasks into a single Composed *Visual* Retrieval (CVR) task, which requires the composition of image and video samples with textual modifications using a unified retrieval model. Our principal insight is that the video modality can be effectively added to existing vision-language pretrained models. When integrated with the Spherical Linear Interpolation (Slerp) method previously proposed for Composed Image Retrieval (CoIR), we found that it results in an effective approach for solving the CVR task, which we called Slerp$^+$. Extensive experiments demonstrate Slerp$^+$'s superiority across various composed image and video retrieval benchmarks, including our newly proposed video benchmark. Notably, Slerp$^+$ mutually enhances image and video retrieval performance over single-modality models, underscoring its potential to transform the field of compositional visual retrieval.

1 INTRODUCTION

Composed Image Retrieval (CoIR) and composed Video Retrieval (CoVR) tasks take visual data and user-provided textual instructions as queries to retrieve relevant data samples from the gallery. Given that certain characteristics are more accurately described through language while others are more effectively conveyed visually, the multi-modality of the query enhances the quality of search results. For example, a customer may combine text descriptions with images to find specific products, such as searching for "white shirts similar to this red one". The potential of composed retrieval in a wide range of practical applications has led to a significant rise in the level of interest within the retrieval community.

In the supervised setting, the task of composed retrieval requires expensive annotations of triplets for model training, each of which consists of a reference visual input, a relative caption, and target data Liu et al. (2021); Baldrati et al. (2022); Ventura et al. (2023); Xu et al. (2024). As a result, zero-shot learning-based methods Saito et al. (2023); Baldrati et al. (2023); Gu et al. (2023); Du et al. (2024) as well as pseudo-triplet based methods, as proposed in Jang et al. (2024b); Ventura et al. (2023), have gained popularity in recent years. However, both approaches require a complex additional layer in their designs, complicating the process and restricting their applicability to specific contexts. For instance, zero-shot methods require an additional projection module to transform images into pseudo-word tokens. In the case of pseudo-triplet methods, the generation of pseudo-triplets necessitates costly tuning of a Large Language Model (LLM), and potentially suffers from the inherent hallucinations that LLMs are especially prone to. Lastly, it is important to note that the CoIR and CoVR tasks have so far been isolated from each other, potentially overlooking the overall effectiveness that a unified retrieval system could offer.

In this paper, we propose an extension of the Spherical Linear Interpolation (Slerp) Jang et al. (2024a)-based CoIR method for the unified image-video-text Composed Visual Retrieval (CVR) task, which we call Slerp$^+$. Our method provides a simple yet effective solution that integrates image and video representations with text, improving both image and video composed retrieval performances. As

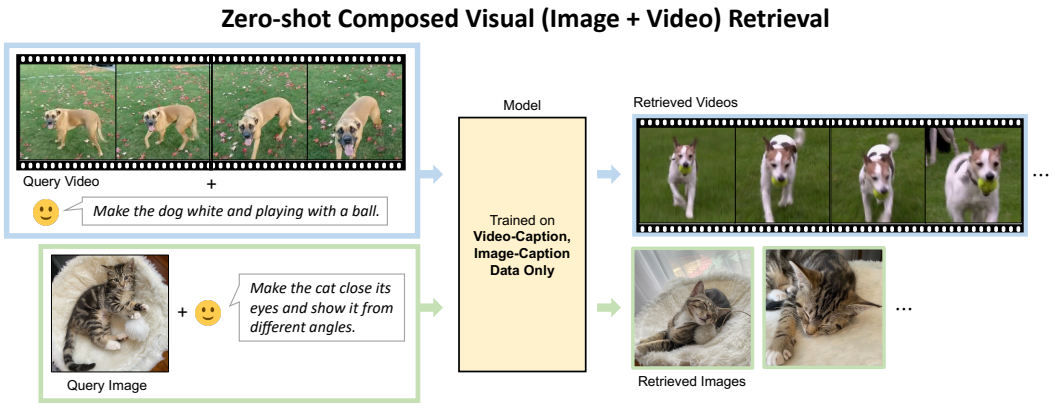

Figure 1: Expanding on the current composed retrieval task, which typically considers only dual modalities of either video-text or image-text, we propose a new task where a query, composed of either an image or video accompanied by user modification text, can retrieve corresponding images and/or videos. To address this task, we introduce a novel method: Slerp$^+$. Our model is trained in a unified manner, utilizing both video-caption and image-caption pairs. Like other zero-shot composed retrieval methods, Slerp$^+$ is designed to be trained without compositional supervision of modification text, relying solely on the captions corresponding to the image or video samples.

shown in Figure 1, our objective is to construct a single retrieval system that can compose both image and video samples with modification text in a zero-shot manner. We find that existing vision-language pretraining objectives (e.g., BLIP's Li & et al. (2022) image-text contrastive learning and image-text matching), when applied to image, video, and text samples in a unified manner, are sufficient for enabling CVR with Slerp, eliminating the need for complicated design choices.

Specifically, Slerp$^+$ employs a ViT vision encoder Dosovitskiy et al. (2020), and a BERT-style text encoder that includes cross-attention layers for vision-language understanding Devlin et al. (2018); Li & et al. (2022); Li et al. (2023), as our baseline. These attention layers process textual queries to attend to visual keys and values, facilitating the learning of a seamless understanding between image-text or video-text. During the training phase, a dataset configured with image-caption and video-caption pairs is utilized to train the image-video-text-unified model of Slerp$^+$. Vision-Text Contrastive learning (VTC) and Vision-Text Matching (VTM), where 'Vision' refers to both images and videos, losses are incorporated to foster alignment between the representations of the images, videos, and their corresponding captions.

Without additional fine-tuning for the CVR task, in the retrieval phase, we apply Slerp Shoemake (1985); Jang et al. (2024a) to the visual and textual embeddings produced by the model to create a composed embedding efficiently. Comprehensive experiments with these embeddings on existing CoIR and CoVR datasets Liu et al. (2021); Wu et al. (2021); Ventura et al. (2023) provide strong evidence on the superiority of our proposed unified method.

Finally, despite the growing significance of zero-shot composed retrieval, the field of CoVR remains largely under-explored. To encourage more robust research discoveries, we introduce a new video *evaluation* benchmark based on Activitynet-captions Krishna et al. (2017) dataset that incorporates more complex textual modifications to increase the task's complexity. The robustness and adaptability of our Slerp$^+$, as demonstrated on this dataset, underscores its potential to significantly advance the field of compositional vision-language retrieval.

Our contributions can be summarized as:

- To the best of our knowledge, we are the first to attempt to build a unified composed retrieval system that integrates image, video, and text modalities, marking a new and significant area yet to be explored.
- We propose Slerp$^+$, which despite its simplicity, is an extremely effective method. Trained solely with image-caption and video-caption pairs, Slerp$^+$ comprehends images, videos, and text to produce aligned embeddings for composed retrieval.

- By leveraging Slerp, Slerp$^+$ achieves strong results on existing composed image and video retrieval tasks, as well as our newly introduced video benchmark, demonstrating the effectiveness of the proposed method.

2 Related Work

Composed Retrieval with Supervised Triplets Supervised learning-based methods for Composed Image Retrieval (CoIR) Liu et al. (2021); Baldrati et al. (2022); Kim et al. (2021); Goenka et al. (2022); Delmas et al. (2022); Xu et al. (2024) have garnered significant interest due to their decent performance and domain-specific applicability. These methods utilize human-annotated triplets, comprising reference images, textual descriptions of desired modifications, and target images, to train CoIR models for natural images Liu et al. (2021) and fashion images Wu et al. (2021); Goenka et al. (2022). An intriguing approach to compose triplets follows a semi-supervised paradigm. Here, Visual Delta Generation (VDG) Jang et al. (2024b) uses both supervised triplets and unlabeled data samples to enhance the performance of the CoIR model. Ventura et al. (2023) further expanded the composed retrieval task to the video modality, proposing the CoVR task. They construct triplets using videos instead of images, after which they train a CoVR model that shows promising performance. However, approaches utilizing supervised triplets face significant challenges when there is a need to scale up the training set due to the high labeling cost. Additionally, domain-specific supervised triplets often result in low generalization capability. In this work, we propose a unified method for both images and videos that doesn't rely on any human-annotated triplets, thereby reducing domain-specific bias and offering flexibility for various composed retrieval use cases.

Zero-shot Composed Retrieval Zero-shot CoIR models, which use large-scale image-caption pairs instead of annotated triplets, have also been proposed Saito et al. (2023); Cohen et al. (2022); Baldrati et al. (2023); Du et al. (2024); Gu et al. (2023); Jang et al. (2024a). These zero-shot methods leverage vision-language pretraining models Radford et al. (2021); Li & et al. (2022); Li et al. (2023) to provide image-text joint representation. The ease of large scale data collection through web-crawling has accelerated the development of these vision-language pretrained models, and in turn more generalized zero-shot CoIR models. However, zero-shot learning has not yet been extended to the video modality. This is primarily due to the greater difficulty in gathering video-caption pairs compared to image-caption pairs and the increased computational cost. Additionally, the complexity of video data may require more data samples to achieve satisfactory zero-shot composed retrieval performance. To address these challenges, we propose a simple and effective unified CVR method, Slerp$^+$, that utilizes both image-caption and video-caption pairs for training. This approach not only allows both modalities to complement each other, it is also more practical due to its capability to perform image or video composed retrieval with an unique integrated model.

3 Method

In this section, we detail our unique, unified Slerp$^+$ framework for the Composed Visual Retrieval (CVR) task. Similar to zero-shot CoIR methods Saito et al. (2023); Cohen et al. (2022); Baldrati et al. (2023); Levy et al. (2023); Du et al. (2024); Jang et al. (2024a), Slerp$^+$ utilizes image-caption pairs for training while also incorporating video-caption pairs to achieve a seamless understanding of image, video, and text modalities. We first discuss the embedding-based learning scheme that trains a transformer-based vision-language encoders. Subsequently, we describe the retrieval inference process, which employs linear interpolation between visual and textual embeddings to generate a composed embedding.

3.1 Preliminaries

To generate cross-modal aligned embeddings from image-text pairs, vision-language pretraining, such as CLIP Radford et al. (2021), are trained on a large dataset $\mathcal{D}_{(x,t)} = \{x_n, t_n\}_{n=1}^{N}$, where each pair consists of an image (x) and its corresponding caption (t). These models are equipped with a trainable vision encoder E_v and a text encoder E_t, which produce image embedding $\mathbf{v} = E_v(x)$ and text embedding $\mathbf{w} = E_t(t)$, respectively. Both $\mathbf{v}$ and $\mathbf{w}$ are $l2$-normalized d-dimensional vectors, i.e., $\mathbf{v}, \mathbf{w} \in \mathbb{R}^{d \times 1}$.

Then, contrastive loss $\mathcal{L}_{cont.}$ is applied on these embeddings as:

$$\min_{\{\theta_{E_v},\theta_{E_t}\}} \mathcal{L}_{cont.} = \mathcal{L}_{V2T} + \mathcal{L}_{T2V}, \tag{1}$$

where the parameters of the vision encoder θ_{E_v} and the text encoder θ_{E_t} are trained using two terms of normalized temperature-scaled cross entropy loss Oord et al. (2018) which are defined as:

$$\mathcal{L}_{V2T} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log \frac{\exp\left(\mathbf{v}_i^T \cdot \mathbf{w}_i/\tau\right)}{\sum_{j\in\mathcal{B}} \exp\left(\mathbf{v}_i^T \cdot \mathbf{w}_j/\tau\right)}, \tag{2}$$

$$\mathcal{L}_{T2V} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log \frac{\exp\left(\mathbf{w}_i^T \cdot \mathbf{v}_i/\tau\right)}{\sum_{j\in\mathcal{B}} \exp\left(\mathbf{w}_i^T \cdot \mathbf{v}_j/\tau\right)}. \tag{3}$$

Here, $\mathcal{B}$ denotes a training batch sub-sampled from $\mathcal{D}$, and τ is the temperature for scaling similarity.

In another vision-language pretraining, BLIP Li & et al. (2022), the learning process is improved with the introduction of cross-attention layers to the text encoder, where the vision encoder output tokens from image patches are attended with a cross-attention layer to learn a joint representation between image and text. Moreover, an additional binary classification loss known as image-text matching loss is employed. This loss function is designed to capture the fine-grained alignment between vision and language, which is formulated as:

$$\min_{\{\theta_{E_v},\theta_{E_t},\theta_m\}} \mathcal{L}_{mat.} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} [y_i \log(p_i) + (1-y_i)\log(1-p_i)], \tag{4}$$

where y_i is the true label denoting if the i-th image-text pair in the batch is matched (1) or not (0), and p_i is the predicted probability of the i-th pair being a match. This probability is computed by passing the combined image and text embeddings through a matching head, represented by a linear layer, parameterized as θ_m.

The contrastive learning scheme aligns an image with its corresponding text caption, effectively distinguishing it from unpaired ones. This process is further refined through image-text matching, which promotes a detailed comprehension of the relationship between visual and textual data. Ultimately, both the image and text encoders are optimized to generate embeddings that capture the semantic alignment between images and their corresponding text captions.

3.2 Unified Training of Image-Video-Text

In this work, we extend the concept of zero-shot composed retrieval to encompass images, videos, and text within a unified Slerp$^+$ framework. To accomplish this, we utilize a vision encoder E_v and a text encoder E_t from BLIP Li & et al. (2022), instead of from CLIP Radford et al. (2021). Unlike CLIP, the text encoder in BLIP incorporates cross-attention layers, enabling it to attend to the entire tokens of vision and text simultaneously. This design is especially advantageous for understanding the complex relationship between visual data and text, such as compositional reasoning. As a result, we establish our baseline based on the BLIP model, and fine-tune only the text encoder while freezing the vision encoder. This ensures that the vision encoder produces consistent outputs from both images and videos, and allows the text encoder to learn how to effectively integrate visual inputs.

For a given datasets of image-text paired $\mathcal{D}_{(x,t)} = \{x_n, t_n\}_{n=1}^{N_x}$, and a video-text paired $\mathcal{D}_{(z,t)} = \{z_n, t_n\}_{n=1}^{N_z}$ where z represents video of M-frames (images) as $z = [x_1, ..., x_M]$, we train Slerp$^+$ model in a image-video-text holistic manner as shown in Figure 2. Note that, the vision [CLS] token and text [CLS] token are input into the vision and text encoder respectively for each sample, and the output embedding from these [CLS] tokens is used to represent each data sample.

Specifically for contrastive learning, we first forward image samples x from a training batch $\mathcal{B}$, where $\mathcal{B} \sim \mathcal{D}(x,t) \cup \mathcal{D}(z,t)$, to the vision encoder to obtain the image embedding $\mathbf{v}_x$, where $\mathbf{v}_x = E_v(x)$. We also forward frames of video in $\mathcal{B}$ to obtain the video embedding $\mathbf{v}_z$, where $\mathbf{v}_z = \mathbb{E}_{x\sim z}[E_v(x)]$ and $\mathbb{E}$ denotes expectation (averaging). Next, we obtain the text embedding using the captions from

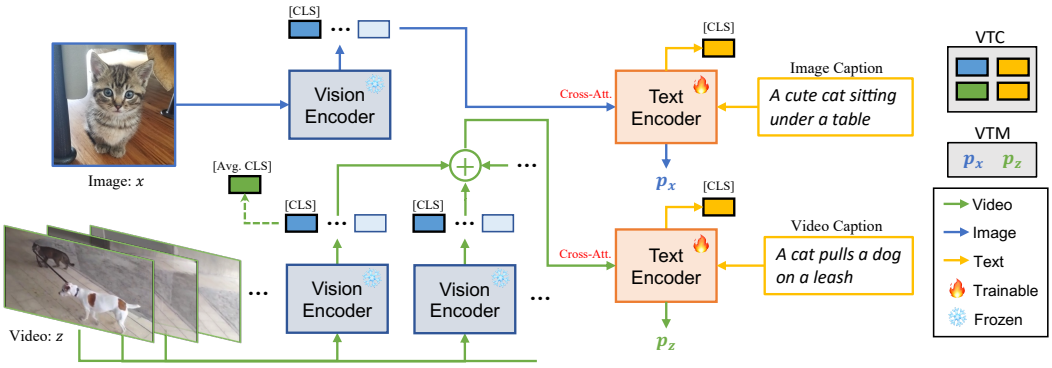

Figure 2: Overview of Slerp$^+$'s learning process. Avg. is an abbreviation for average, and Cross-Att. stands for cross-attention. We forward image-caption and video-caption pairs to generate embeddings using [CLS] token. All images and video frames are processed using the same frozen vision encoder, and their output patch-wise tokens are forwarded to the text encoder to compute cross-attention with the corresponding text. The text encoder is then updated using VTC and VTM losses.

images and videos, denoted as $\mathbf{w}$, where $\mathbf{w} = E_t(t)$. Finally, to achieve comprehensive representation learning that integrates image, video, and text, we concatenate the image and video embeddings. Separately, we also concatenate the embeddings of image captions and video captions within a batch. Using these, we compute the Vision-Text Contrastive (VTC) loss using Equation 1.

To enable the text encoder to better understand visual inputs and identify fine-grained alignment between visual and text elements, we introduce the Vision-Text Matching (VTM) loss using Equation 4. From the image perspective, we process the image with the vision encoder to obtain the last hidden state tokens (patch-wise tokens). These tokens are then forwarded to the text encoder to compute cross-attention with text tokens from the image caption. We calculate the binary classification prediction p_x using a trainable matching head from the text [CLS] embedding.

For video, we process frames with the vision encoder to obtain the last hidden state tokens separately, then concatenate them in a temporally sequential manner. The text encoder calculates cross-attention between these tokens with the video caption text tokens, and the matching head predicts the score p_z from text [CLS] embedding. Additionally, we employ the hard negative mining strategy Li et al. (2021) to identify more informative negatives, and we involve all of image, video, text embeddings in this process. Consequently, by learning to discriminate whether the matched pairs are positive or negative regardless of the data modality, the model is able to comprehend a holistic representation of image-video and text.

3.3 Inference

After training the model, it can be utilized to process images, videos, and text, generating embeddings for retrieval. However, executing composed retrieval requires a method to combine visual and text embeddings. One such approach is the early-fusion strategy. This strategy simultaneously feeds the text encoder of the model with visual patch tokens and text modifications, using the output [CLS] token as the composed embedding. However, this approach may not be entirely accurate. The model, trained to handle multi-modal inputs with VTM - a system designed to classify whether visual data and text match, is not intended to compose visual and text elements to project a composed output.

In light of this, an alternative late-fusion approach that utilizes interpolation between visual and text embeddings, also known as Slerp Shoemake (1985); Jang et al. (2024a), can be considered. This strategy employs individual modality encoders to generate separate visual and text embeddings, which are then linearly interpolated to obtain a composed embedding $\mathbf{c}$ as:

$$\mathbf{c} : \text{Slerp}(\mathbf{v}, \mathbf{w}; t) = (\sin((1-t)\alpha) \cdot \mathbf{v}, + \sin(t\alpha) \cdot \mathbf{w}) \,/\, \sin(\alpha) \tag{5}$$

where t is balancing scalar value and α is the angle between embeddings which is computed as $\alpha = \cos^{-1}(\mathbf{v} \cdot \mathbf{w})$. This process can be seamlessly applied to the model trained in Section 3.2. The

reason lies in the contribution of both VTC and VTM losses to the discovery of a fine-grained semantic alignment among the embeddings of images, videos, and text simultaneously. Such alignment leads to densely distributed embeddings, covering the intermediate compositional representation successfully.

Finally, we can construct a image-video-text unified retrieval system using the Slerp process, which we call Slerp$^+$. The retrieval gallery is established using visual embeddings $\mathbf{v}$ obtained from image and video samples. Users can input a query image or video, along with a text modification, to create a composed embedding $\mathbf{c}$ using the Slerp method. The system then searches relevant images or videos from the gallery by calculating the cosine similarity between the composed query embedding and each individual image or video embedding.

4 EXPERIMENTS

In this section, we first outline the experimental settings including the datasets, evaluation metrics, and implementation details (Section 4.1). Following this, we present the results of composed video retrieval on existing benchmark as well as our newly introduced one (Section 4.2) and composed image retrieval (Section 4.3). Finally, we provide further analysis with ablation study (Section 4.4).

4.1 SETUP

Datasets. To train the Slerp$^+$ model, we utilize two types of datasets: (1) image-caption pairs and (2) video-caption pairs. Specifically for image-caption pairs, we use CC3M dataset Sharma et al. (2018), which was collected through web crawling. This is a common dataset for training zero-shot composed IR methods Saito et al. (2023); Gu et al. (2023); Du et al. (2024), and we use a subset of 2.3M pairs that was accessible to us to train the model.

For video-caption pairs, we use a subset of the WebVid dataset Bain et al. (2021). Following the setup proposed in CoVR Ventura et al. (2023), we select its training split, which consists of a total of 130K video-caption pairs, ensuring no overlap between the training and test sets. Furthermore, we clean up the dataset by removing noisy samples with too few frames or exceptionally long videos, resulting in 94K video-caption pairs for the training set of the Slerp$^+$ model.

To evaluate composed retrieval performance, we utilize two image-based benchmarks and two video-based benchmarks. For images, first, we utilize *CIRR* Liu et al. (2021), which deals with natural images, where the test split consists of 503 subgroups of 2,178 images. Second, we utilize *FashionIQ* Wu et al. (2021), which focuses on fashion domain images of three categories: Dress, Shirt, and Toptee. Following the literature, we employ the validation split for evaluation, which consists of 15,415 images.

For videos, first, we utilize the CoVR test splits named *WebVid-CoVR-Test* set. A high-quality set of 2,556 triplets is carefully selected and curated to form the separate corpus of the WebVid dataset. Second, to further evaluate the performance of models on composed VR tasks, we establish a new experimental protocol of more complex text modifications, *Activitynet-CoVR*, based on the validation split of the Activitynet-captions dataset Krishna et al. (2017). This dataset labels each video with corresponding sentences (captions) and timestamps. We create reference and target video pairs using two methods: (1) Intra-pair, selecting two video clips from the same video, with the earliest caption and the corresponding video clip as the reference and the last clip as the target, and (2) Inter-pair, selecting two video clips from different videos by calculating the similarity between all possible video clips and choosing the closest two with cosine similarity scores over 0.8, measured with the VideoMAE-Large model Tong et al. (2022). This results in a total of 1,260 pairs. Similar to the annotation steps for the CoVR test set, we use the instruction-tuned LLM, LLaMA3 LLa (2024). We prompt it directly with [*Please give an imperative that makes video A to video B*], including the caption of each video clip. Finally, annotators carefully filter out noisy triplets, resulting in 800 composed VR triplets for evaluation.

Evaluation Metrics. In line with the protocols adopted in benchmarks Saito et al. (2023); Ventura et al. (2023), we evaluate the model's performance using recall scores at the top K retrieval results (R@K) for ranks 1, 5, 10, and 50. The metric Recall at rank k (R@k) measures the frequency at which the correct image or video appears among the top k results. A higher recall score indicates superior performance.

Table 1: Retrieval results on *WebVid-CoVR-Test* set.

Type	Method	Query Modality	Fusion	R@1	R@5	R@10	R@50
Zero-shot	BLIP† Li & et al. (2022)	Text	-	21.17	41.78	50.47	70.42
	BLIP† Li & et al. (2022)	Video	Avg	39.44	62.99	72.30	89.71
	BLIP† Li & et al. (2022)	Video + Text	Avg	45.62	70.55	79.67	93.40
	X-CLIP† Ni et al. (2022)	Video + Text	Slerp	42.09	66.19	74.72	89.39
	Slerp$^+$	Text	-	21.95	42.92	51.88	71.67
	Slerp$^+$	Video	Avg	41.82	63.26	72.34	89.87
	Slerp$^+$	Video + Text	Slerp	**57.82**	**80.16**	<u>86.38</u>	<u>96.60</u>
Fine-tuned on WebVid-CoVR	CoVR‡ Ventura et al. (2023)	Text	-	23.67	45.89	55.13	77.03
	CoVR‡ Ventura et al. (2023)	Video	Avg	38.89	64.98	74.02	92.06
	CoVR† Ventura et al. (2023)	Video + Text	Slerp	44.44	66.20	75.31	91.98
	CoVR‡ Ventura et al. (2023)	Video + Text	CA	<u>53.13</u>	<u>79.93</u>	**86.85**	**97.69**

Implementation Details. As a baseline for the Slerp$^+$ model, we utilize the BLIP model Li & et al. (2022), configured with a ViT-L/16 vision encoder Dosovitskiy et al. (2020) and a Bert-based text encoder Devlin et al. (2018). We use a fine-tuned version of the BLIP model on the MS COCO dataset Lin et al. (2014). The pretrained weights provided by HuggingFace[1] Wolf et al. (2020) are applied to this baseline model under the identifier: `Salesforce/blip-itm-large-coco`.

In the training of the Slerp$^+$ model (Section 3.2), we aim to retain the baseline's knowledge acquired from large-scale pretraining. To achieve this, we employ the parameter-efficient fine-tuning technique, LoRA Hu et al. (2021), to the text encoder. This technique allows us to keep the original parameters intact while introducing small adaptation weights for tuning. The additional LoRA parameters are configured as: LoRA$_\alpha = 16$, rank $= 16$, and dropout $= 0.1$. Throughout the training process, we maintain the entire set of parameters for E_v and E_t as fixed. Only the parameters for LoRA, the text projection linear layer, and the matching head θ_m are updated. The models are trained using $8 \times$ A100-80GB GPUs. For video training, we use a batch size of 4, sampling 8 equally-spaced frames per video, resulting in a total of 32 frames. For image training, we use a batch size of 32. Therefore, each GPU processes a total of 64 samples (frames + images). The initial temperature τ for scaling is set to 1/0.07 and is continuously updated during training. We employ the AdamW optimizer Loshchilov & Hutter (2018) with a fixed learning rate of 3e-5 and a weight decay of 0.01. The training is conducted for single epoch, with the trainable parameters constituting less than 0.32% of the total parameters, ensuring efficiency.

During the evaluation, we adopt the same frame sampling strategy as in the training stage, sampling 8 equally-spaced frames per video for both query and gallery videos. To compose the visual and text embeddings with late-fusion (Slerp), we set t to 0.6 for videos and 0.7 for images by default. For comparison, we also utilize the X-CLIP Ni et al. (2022) model, a CLIP-style video-text pretraining, with the same Slerp inference. Additionally, in line with the experiments conducted in CoVR Ventura et al. (2023) that evaluate the performance of image-text models in a video-text scenario, we input the selected frames separately into the vision encoder of BLIP. We then either average (Avg) the image embeddings or apply an early-fusion strategy with the pretrained Cross-Attention (CA) as used in CoVR. We use the pretrained checkpoint from Huggingface: `microsoft/xclip-large-patch14`, and the pretrained checkpoint provided by the official implementation of the CoVR model.

4.2 Composed Video Retrieval Results

In the subsequent sections, we compare our Slerp$^+$ method with existing zero-shot CoIR, supervised CoVR, and video-text pretraining methods Saito et al. (2023); Baldrati et al. (2023); Gu et al. (2023); Du et al. (2024); Jang et al. (2024a); Ventura et al. (2023); Ni et al. (2022). The symbol † indicates that we conducted experiments using provided checkpoints on our setup, while ‡ denotes that the results were directly obtained from the best scores reported by each method. The best scores are marked in bold, while the second best are underlined.

[1]https://huggingface.co/models

Table 2: Retrieval results on *Activitynet-CoVR* set.

Type	Method	Query Modality	Fusion	R@1	R@5	R@10	R@50
Zero-shot	BLIP† Li & et al. (2022)	Text	-	23.87	52.25	62.75	83.38
	BLIP† Li & et al. (2022)	Video	Avg	34.50	46.00	48.87	57.25
	BLIP† Li & et al. (2022)	Video + Text	Avg	35.60	47.32	49.12	59.10
	X-CLIP† Ni et al. (2022)	Video + Text	Slerp	33.00	50.25	54.50	67.50
	Slerp$^+$	Text	-	24.62	52.13	65.62	<u>84.00</u>
	Slerp$^+$	Video	Avg	<u>38.12</u>	46.50	49.12	57.63
	Slerp$^+$	Video + Text	Slerp	**43.00**	**60.62**	**68.37**	**84.50**
Fine-tuned on WebVid-CoVR	CoVR† Ventura et al. (2023)	Text	-	23.67	45.89	55.13	77.03
	CoVR† Ventura et al. (2023)	Video	Avg	36.62	46.12	49.37	57.37
	CoVR† Ventura et al. (2023)	Video + Text	Slerp	37.38	47.87	51.75	61.50
	CoVR† Ventura et al. (2023)	Video + Text	CA	35.12	<u>60.25</u>	<u>67.87</u>	83.25

Table 3: Zero-shot CoIR results on *CIRR test set*.

Method	Recall@K				Recall$_{\text{subset}}$@K		
	K=1	K=5	K=10	K=50	K=1	K=2	K=3
Pic2Word‡ Saito et al. (2023)	23.90	51.70	65.30	87.80	-	-	-
SEARLE‡ Baldrati et al. (2023)	24.22	52.41	66.29	88.63	53.71	74.63	87.61
LinCIR‡ Gu et al. (2023)	25.04	53.25	66.68	-	57.11	77.37	88.89
Image2Sentence‡ Du et al. (2024)	30.84	61.06	73.57	**92.43**	-	-	-
Slerp + TAT‡ Jang et al. (2024a)	33.98	61.74	72.70	88.94	68.55	<u>85.11</u>	<u>93.21</u>
CoVR‡ Ventura et al. (2023)	<u>38.48</u>	<u>66.70</u>	<u>77.25</u>	91.47	<u>69.28</u>	83.76	91.11
Slerp$^+$	**39.74**	**67.74**	**77.40**	<u>91.55</u>	**70.65**	**86.72**	**94.36**

WebVid-CoVR-Test. The retrieval results on the CoVR-test set are shown in Table 1. We divide the methods into two categories: zero-shot, which does not use supervised composed VR triplets used in CoVR Ventura et al. (2023), and the supervised CoVR approach, which is fine-tuned on the WebVid-CoVR-Training set. Despite not using any supervised triplets and being trained in a zero-shot manner, Slerp$^+$ achieves the highest scores for R@1 and R@5. The R@1 score shows a significant gap, while the R@10 and R@50 scores are not far behind the top scores. When comparing Slerp$^+$ in terms of query modality, it is evident that Slerp$^+$ successfully combines video and text queries to retrieve relevant samples from the gallery, achieving significantly better retrieval scores than text-only or video-only. The success of Slerp$^+$ is not solely due to the use of Slerp, as evidenced by the results of CoVR with Slerp and X-CLIP with Slerp, which are far inferior to Slerp$^+$ using the same fusion approach. In this experiment, Slerp$^+$ demonstrates the benefits of holistic representation learning over image-caption only pretrained BLIP, video-caption only pretrained X-CLIP, and video composed triplet supervised CoVR.

Activitynet-CoVR. The experimental results on Activitynet-CoVR with longer text modifications are presented in Table 2. In this setup, Slerp$^+$ outperforms all other zero-shot methods across all Recall ranks, even surpassing the fine-tuned CoVR. The supervised training of the CoVR model with WebVid results in less generalization with Activitynet, leading to lower scores in R@1. Notably, Slerp$^+$ with only video achieves the second highest score in R@1, indicating its precision in retrieving the most relevant result. Conversely, Slerp$^+$ with only text secures the second highest score in R@50, demonstrating its effectiveness in retrieving a broader set of relevant results. These scores confirm the proficiency of the Slerp$^+$ model in balancing precision and recall, effectively leveraging information from both video and text modalities.

4.3 Composed Image Retrieval Results

CIRR. The retrieval results on CIRR dataset are presented in Table 3. Despite the unified training of both image-caption and video-caption pairs, where image and video may not necessarily support each other and could potentially degrade each other's performance, Slerp$^+$ still manages to achieve outstanding performance. This holds true even when compared with text-only Gu et al. (2023),

Table 4: Zero-shot CoIR results on *FashionIQ validation set*.

Method	Dress		Shirt		Toptee		Average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Pic2Word‡ Saito et al. (2023)	20.00	40.20	26.20	43.60	27.90	47.40	24.70	43.70
SEARLE‡ Baldrati et al. (2023)	20.48	43.13	26.89	45.58	29.32	49.97	25.56	46.23
LinCIR‡ Gu et al. (2023)	20.92	42.44	29.10	46.81	28.81	50.18	26.28	46.49
Image2Sentence‡ Du et al. (2024)	25.33	46.26	30.03	48.58	33.45	53.80	29.60	49.54
Slerp + TAT‡ Jang et al. (2024a)	<u>29.15</u>	<u>50.62</u>	<u>32.14</u>	<u>51.62</u>	<u>37.02</u>	<u>57.73</u>	<u>32.77</u>	<u>53.32</u>
CoVR‡ Ventura et al. (2023)	21.95	39.05	30.37	46.12	30.78	48.73	27.70	44.63
Slerp$^+$	**31.78**	**54.04**	**37.73**	**56.82**	**41.36**	**62.37**	**36.96**	**57.74**

Table 5: Ablation study results on CoVR and CoIR benchmarks.

Ablation	*Activitynet-CoVR*				*CIRR test*			
	R@1	R@5	R@10	R@50	R@1	R@5	R@10	R@50
Baseline	43.00	60.62	68.37	84.50	39.74	67.74	77.40	91.55
(a) Number of frames = 4	42.37	60.12	66.75	83.70	38.48	67.08	77.32	91.27
(b) Number of frames = 16	43.50	60.86	68.88	84.90	39.81	68.07	77.52	91.68
(c) Fusion-Avg	41.87	58.13	63.62	80.37	-	-	-	-
(d) Without VTC	41.85	57.82	66.78	82.20	37.68	65.22	75.52	90.58
(e) Without VTM	42.06	58.84	67.45	83.63	38.48	66.08	76.10	90.78
(f) Image-only	40.98	57.98	66.10	81.52	37.34	64.72	75.10	90.38
(g) Video-only	41.20	58.10	66.68	81.88	37.08	64.37	75.02	90.11

image-caption only Saito et al. (2023); Baldrati et al. (2023); Du et al. (2024); Jang et al. (2024a) or video-caption-only Ventura et al. (2023) methods. When compared with Slerp + TAT, Slerp$^+$ proves to be a superior model for applying Slerp-based composed retrieval, demonstrating its advantages in the image domain. This is not only applicable to general retrieval scenarios but also effective for subset cases, as evidenced by Slerp$^+$ achieving the best scores for all K values.

FashionIQ. The experimental results for specific visual domain images, specifically fashion images from the FashionIQ dataset, are presented in Table 4. Consistent with the results from the natural image and video datasets, our Slerp$^+$ significantly outperforms other methods across all recall scores. This demonstrates the versatility and robustness of the Slerp$^+$ model, as it excels not only in general image and video retrieval tasks but also in specialized domains like fashion. The model's ability to effectively utilize both image, video with text modalities contributes to its superior performance, making it a comprehensive solution for various retrieval scenarios.

4.4 Analysis

Ablation Study. To validate our approach, we perform an ablation study on Slerp$^+$ regarding training schemes and report the results in Table 5. In (a, b), we vary the number of frames extracted from videos during training. Compared to the baseline, which uses 8 frames per video, we find that using more frames slightly improves performance. However, the increase is marginal, so we choose to use 8 frames per video by default for efficiency. In (c), we try averaging all frame embeddings with text embedding instead of using the Slerp to obtain composed embedding, but this result in poorer performance. In (d, e), we examine the impact of each training loss. We observe that both have a decent effect on performance improvement, and they complement each other to achieve the best performance when combined. Lastly, in (f, g), we examine the joint training of image-video-text by excluding video in (f) and image in (g). The results show that utilizing both image and video samples simultaneously with the proposed training scheme improves both image and video composed retrieval accuracy.

Qualitative Results. Figures 3 and 4 show the retrieval results on CoVR and CoIR tasks, respectively. Our Slerp$^+$ model effectively combines video or image and text modifications to successfully retrieve relevant results.

Figure 3: Retrieval results of Slerp$^+$ on WebVid-CoVR (above) and Activitynet-CoVR (below).

Figure 4: Retrieval results of Slerp$^+$ on CIRR (above) and FashionIQ (below).

5 Discussion & Conclusion

Potential Impact. By combining visual data with textual modifications, our Slerp$^+$ system addresses the limitations of both supervised and zero-shot learning approaches, providing a unified framework that improves retrieval accuracy without relying on costly supervised triplets. The potential societal impact includes significantly improved search functionalities across various applications, from e-commerce to multimedia management, by providing accurate and contextually relevant results.

Limitation. A potential limitation of Slerp$^+$ system is its dependency on the quality and diversity of the image-caption and video-caption pairs used during training. While the zero-shot learning approach alleviates the need for supervised triplets, the model's performance may still be constrained by the representational richness and variety of the training data.

Conclusion. In this paper, we introduced the CVR task and Slerp$^+$, a novel unified framework for composed image and video retrieval that seamlessly integrates visual and textual modalities. By employing a vision and text encoder which are fine-tuned with both image-caption and video-caption pairs, Slerp$^+$ overcomes the limitations of traditional supervised learning and zero-shot approaches. Our method's effectiveness is demonstrated through significant improvements on existing CoIR and CoVR benchmarks and a new and more complex video evaluation benchmark that we introduce in this paper. The results underscore the potential of Slerp$^+$ to advance compositional vision-language retrieval, offering a robust and scalable solution for diverse real-world applications.

REFERENCES

Meta llama3, https://llama.meta.com/llama3/, 2024.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *CVPR*, 2022.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *ICCV*, 2023.

Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *ECCV*, 2022.

Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *ICLR*, 2022.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yongchao Du, Min Wang, Wengang Zhou, Shuping Hui, and Houqiang Li. Image2sentence based asymmetrical zero-shot composed image retrieval. *arXiv preprint arXiv:2403.01431*, 2024.

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 *CVPR*, 2022.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval. *arXiv preprint arXiv:2312.01998*, 2023.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021.

Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. *arXiv preprint arXiv:2405.00571*, 2024a.

Young Kyun Jang, Donghyun Kim, Zihang Meng, Dat Huynh, and Ser-Nam Lim. Visual delta generator with large multi-modal models for semi-supervised composed image retrieval. In *CVPR*, 2024b.

Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual compositional learning in interactive image retrieval. In *AAAI*, 2021.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017.

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.

Junnan Li and et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*. PMLR, 2023.

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 *ECCV*. Springer, 2014.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021.

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 *CVPR*, 2023.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, 1985.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*, 35, 2022.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. CoVR: Learning composed video retrieval from web video captions. *arXiv:2308.14746*, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020.

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 *CVPR*, 2021.

Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, Chun-Mei Feng, et al. Sentence-level prompts benefit composed image retrieval. In *ICLR*, 2024.