Spherical Linear Interpolation and Text-Anchoring for Zero-shot Composed Image Retrieval

Young Kyun Jang[1], Dat Huynh[1], Ashish Shah[1],
Wen-Kai Chen[2], and Ser-Nam Lim[2]

[1] Meta AI
[2] University of Central Florida

Abstract. Composed Image Retrieval (CIR) is a complex task that retrieves images using a query, which is configured with an image and a caption that describes desired modifications to that image. Supervised CIR approaches have shown strong performance, but their reliance on expensive manually-annotated datasets restricts their scalability and broader applicability. To address these issues, previous studies have proposed pseudo-word token-based Zero-Shot CIR (ZS-CIR) methods, which utilize a projection module to map images to word tokens. However, we conjecture that this approach has a downside: the projection module distorts the original image representation and confines the resulting composed embeddings to the text-side. In order to resolve this, we introduce a novel ZS-CIR method that uses Spherical Linear Interpolation (Slerp) to directly merge image and text representations by identifying an intermediate embedding of both. Furthermore, we introduce Text-Anchored-Tuning (TAT), a method that fine-tunes the image encoder while keeping the text encoder fixed. TAT closes the modality gap between images and text, making the Slerp process much more effective. Notably, the TAT method is not only efficient in terms of the scale of the training dataset and training time, but it also serves as an excellent initial checkpoint for training supervised CIR models, thereby highlighting its wider potential. The integration of the Slerp-based ZS-CIR with a TAT-tuned model enables our approach to deliver state-of-the-art retrieval performance across CIR benchmarks. Code is available at https://github.com/youngkyunJang/SLERP-TAT

Keywords: Zero-Shot Composed Image Retrieval · Slerp-based Search · Text-Anchored-Tuning

1 Introduction

Composed Image Retrieval (CIR) is a vision-language task that utilizes both image and text queries to retrieve images with high precision. It aims to identify images that are visually similar to a reference image, but also incorporate changes specified in a text query, thereby providing users with precise control over the
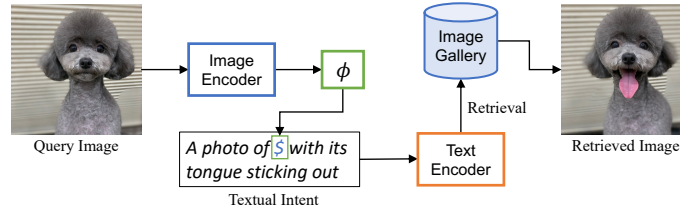
(a) Pseudo-word token-based ZS-CIR method.
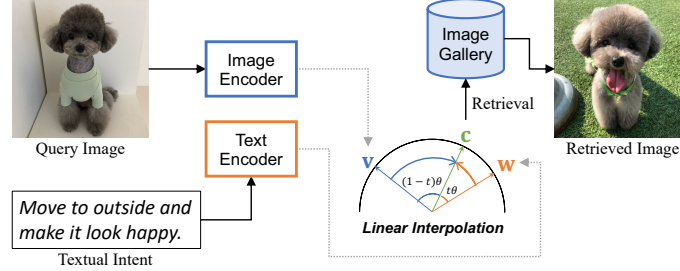
(b) Our linear interpolation based ZS-CIR method.

Fig. 1: Overview of ZS-CIR approaches. The previous works [22,4,1] utilize a projection module, which transforms an image into a textual pseudo-word. This is then combined with text (textual intent) to produce a composed embedding with a text encoder for retrieval purposes. In contrast, we propose a method based on a simple spherical linear interpolation. This method directly combines image ($\mathbf{v}$) and text ($\mathbf{w}$) embeddings to produce a composed embedding ($\mathbf{c}$). We then use $\mathbf{c}$ to perform ZS-CIR.

characteristics of the desired image. This bi-modality of the query allows for a more nuanced search, as some features are better described with language, while others are more effectively expressed visually. Due to its potential in a variety of real-world applications, there has been a surge in attention towards CIR.

Previous works on supervised CIR [2,5,17,7] have proposed solutions to various problems using natural image [18,12,1] and fashion image [30] datasets, configured with triplets of ¡reference image, textual intent, target image¿. However, these solutions face significant challenges, primarily due to the complexity and cost of dataset collection. The process involves creating these triplets, which is not easily automated and requires substantial manual labor. This makes the process time-consuming and resource-intensive, particularly when creating large training sets. Moreover, models trained on these supervised datasets tend to be specialized to specific use-cases, limiting their robustness to diverse unseen domains. Therefore, despite promising results from supervised CIR approaches, their reliance on expensive manually-annotated datasets restricts their scalability and broader applicability.

In response to the challenges faced in supervised CIR, the research community has introduced an alternative approach known as Zero-Shot CIR (ZS-CIR)

[22,4,1,27,12]. ZS-CIR, unlike supervised CIR that relies on triplets, utilizes ¡image, text caption¿ pairs, akin to those used in the Vision-Language Pretraining (VLP) models like CLIP [21] or BLIP [13]. The primary challenge in ZS-CIR is establishing a compositional understanding between image and text. To address this, pseudo-word token-based approaches [22,4,1,8] were proposed as shown in Figure 1a, aiming to utilize a projection module that transforms VLP image encoder output as the pseudo-word token. Then, this pseudo-word token is concatenated with text tokens and forward to VLP text encoder to produce image-text composed representation embedding to perform ZS-CIR.

However, pseudo-word token-based methods may not genuinely "compose image and text" since the representation of the composed embeddings is constrained within the text encoder output embedding space, which may not fully capture the joint image-text information. Moreover, these methods have a limitation in their inference pipeline because the projection module can distort the original, discriminative image representation. This results in a CIR model that is less capable of capturing and representing the complex compositional relationships between image and text samples.

To avoid these problems, we propose a new method called **S**pherical **l**inear Int**erp**olation (Slerp [24])-based Zero-shot CIR. Given that VLP encoders are trained with scaled cosine similarity between image and text embeddings, this results in the distribution of image and text samples on a joint hypersphere with the radius of the scaling factor (temperature). Consequently, Slerp can be applied to find an intermediate embedding of image and text ones, as shown in Figure 1b. Surprisingly, the results from Slerp-based retrieval confirm that simple interpolation between image and text embeddings of VLP can achieve ZS-CIR performance comparable to the best-performing approaches, without any additional training.

Furthermore, we introduce a **T**ext-**A**nchored-**T**uning (TAT) strategy to facilitate the Slerp process by reducing the gap between modalities. Based on the observations in [22,1], which show that the text-only case significantly outperforms the image-only case in CIR benchmarks and sometimes even surpasses the best performing model, we believe that the VLP text embedding itself plays a critical role in CIR. Therefore, we keep the VLP text encoder frozen to maintain its power and allow the text embeddings to serve as an *anchor* for contrastive learning. Specifically, we apply LoRA [11] parameters to the VLP image encoder, which not only preserves the original knowledge of the image encoder but also effectively redistributes image embeddings to align more closely with the corresponding text embeddings. As a result, the modality gap is reduced while the text representations of VLP are retained. Ultimately, by combining Slerp and TAT, we significantly boost the performance of ZS-CIR.

Extensive experimental results on various domain benchmarks, including natural images [18,1] and fashion images [30], underscore the advantages of Slerp-based ZS-CIR. This not only achieves decent retrieval performance, but Slerp also allows for easy adjustment of the contribution of image or text queries based on the user's intent. Our TAT strategy significantly enhances retrieval

performance, even when trained with far fewer image-text training pairs (only 20% of the number used in the CC3M dataset) compared to previous ZS-CIR methods [22,8,27,12]. Moreover, optimal performance of Slerp + TAT method can be achieved with just a single epoch of training, exhibiting superior training efficiency compared to pseudo-word token-based approaches [22,4,1,8], which require tens to hundreds of training epochs. As part of the broader applications of our work, we demonstrate that TAT-trained VLP models can serve as superior initial checkpoints compared to the original VLP models when training supervised CIR models.

2   Related Work

2.1   Supervised Composed Image Retrieval

The field of image retrieval, as surveyed in [20], has attracted significant attention from researchers due to its diverse utilities, such as enhancing search engines and e-commerce platforms. In particular, Composed Image Retrieval (CIR), a method that retrieves images using a pair consisting of a query image and a text that depicts user intent [28], has been widely explored. This expansion from vision-only to vision-language multi-modal applications represents a significant advancement in the field. CIR has been explored in various visual domains, including natural [18,6,1] and fashion [9,10,7,30] images. Notably, CIRR [18] and FashionIQ [30] are widely utilized to train CIR models in a supervised fashion, using human-annotated triplets that contains a reference image, a target image, and a textual description of their difference. Based on this supervision, various supervised CIR methods [5,17,2,18,7] have been proposed, and they have demonstrated decent CIR performances. However, these methods often encounter bottlenecks due to the high cost of labeling and the potential existence of mislabeled samples. Additionally, the current supervised CIR datasets are limited to specific image domains and visual attributes, which restrict their applicability across different domains.

2.2   Zero-shot Composed Image Retrieval

As an alternative to supervised CIR, Zero-Shot CIR (ZS-CIR) methods [22,1,4,8] have been introduced. Instead of using CIR triplets, these zero-shot methods employ image-text caption pairs, similar to those used in the pretraining stage of VLP models [21,13]. ZS-CIR is particularly advantageous for domain generalization as it can utilize simple web-crawled data samples for training, thereby making it easy to scale up the dataset size and incorporate a variety of image domains. This, in turn, can be used to build a robust CIR model. The primary challenge in ZS-CIR is how to achieve composition between image and text. To address this, previous works [22,1,4,8] have used pseudo-word token-based methods, which replace specific characters in text prompts with pseudo-word tokens derived from images. However, this approach does have its limitations. The image representation, which is inherent in the embedding, can become distorted

when it is transformed into a pseudo-word token. Furthermore, this token is then processed through a fixed text encoder. This encoder has only been trained with text inputs, and as such, it may not correctly interpret the information from images. To overcome these issues, we propose a novel ZS-CIR scheme based on Slerp, which simply interpolates between image and text embeddings to obtain a composed representation for retrieval. In this manner, the image embedding directly contributes to building the composed embedding close to its original form. Moreover, we further improve the Slerp-based search with text-anchored fine-tuning to reduce the domain gap in VLP embeddings, thereby facilitating the Slerp process. It's important to note that our TAT differs from the training scheme in [31], which also proposed to freeze a single-modality tower (the image encoder) during the pretraining stage of the VLP model. In contrast, TAT fixes the text encoder and is focused on fine-tuning the existing VLP model to enhance its performance for CIR tasks.

3   Method

In this section, we present our Slerp-based ZS-CIR approach. Then, we introduce the TAT learning method, which aligns image representation closer to a fixed text, thereby enhancing the Slerp-based ZS-CIR performance significantly. We employ the pre-trained vision-language models, CLIP [21] and BLIP [13], as our baseline models. Each of these models is equipped with a separate image and text encoder that generate visual and textual representation embeddings, respectively. During the retrieval stage, the composed representation embedding of the image and text query is obtained using Slerp. This composed embedding is then used to compute the similarity with pre-computed gallery image embeddings, resulting in ranked results.

3.1   Preliminaries

Represented by CLIP [21] and BLIP [13], VLP models are trained on an extensive dataset $\mathcal{D} = \{(x_n, t_n)\}_{n=1}^{N}$ of image-text caption pairs, where x_n is an image and t_n is a tokenized text caption. The VLP model employs a parameterized image encoder E_I and text encoder E_T to produce an image embedding $\mathbf{v} \in \mathbb{R}^d : \mathbf{v} = E_I(x)$, and a text embedding $\mathbf{w} \in \mathbb{R}^d : \mathbf{w} = E_T(t)$ of the same dimensionality d, respectively. Both the image and text embeddings are $l2$-normalized to utilize cosine similarity as a baseline metric.

Then, normalized temperature-scaled cross entropy loss [3] (contrastive loss) is utilized to update trainable parameters of both modality encoders as:

$$\mathcal{L}_{cont.} = \mathcal{L}_{I2T} + \mathcal{L}_{T2I} \tag{1}$$

where each term is described as:

$$\mathcal{L}_{I2T}(\mathcal{B}) = -\frac{1}{N_B} \sum_{i=1}^{N_B} \log \frac{\exp\left(\mathbf{v}_i^T \cdot \mathbf{w}_i / \tau\right)}{\sum_{j=1}^{N_B} \exp\left(\mathbf{v}_i^T \cdot \mathbf{w}_j / \tau\right)}, \tag{2}$$

$$\mathcal{L}_{T2I}(\mathcal{B}) = -\frac{1}{N_B} \sum_{i=1}^{N_B} \log \frac{\exp\left(\mathbf{w}_i^T \cdot \mathbf{v}_i / \tau\right)}{\sum_{j=1}^{N_B} \exp\left(\mathbf{w}_i^T \cdot \mathbf{v}_j / \tau\right)}. \tag{3}$$

Here, $\mathcal{B}$ denotes a training batch consisting of N_B image-text pairs, and τ is the temperature parameter used for scaling. This learning strategy results in an image and its corresponding text caption being aligned, while the unpaired ones are separated.

3.2   Spherical Linear Interpolation-based Retrieval

Due to the cosine similarity-based representation learning in the VLP model, both the image and text embeddings are distributed on a unique hypersphere with a radius of the scaling factor, τ. Moreover, since the VLP model is trained on hundreds of millions of image-text pairs, the embeddings of images and texts are densely distributed on this hypersphere, presumably covering most of the related semantics between the vision and language domains.

Based on this, we assume that intermediate representation of existing image and text embeddings ($\mathbf{v}$ and $\mathbf{w}$) can represent image and text composed representation, and introduce a simple Slerp [24]-based approach for ZS-CIR as:

$$\mathbf{c} : \mathrm{Slerp}(\mathbf{v}, \mathbf{w}; \alpha) = \frac{\sin((1-\alpha)\theta)}{\sin(\theta)}\mathbf{v} + \frac{\sin(\alpha\theta)}{\sin(\theta)}\mathbf{w} \tag{4}$$

where θ is the angle between $\mathbf{v}$ and $\mathbf{w}$ which is obtained as:

$$\theta = \cos^{-1}(\mathbf{v} \cdot \mathbf{w}) \tag{5}$$

and α is a balancing scalar value within the range of $[0,1]$ that determines the scale of the linear combination between $\mathbf{v}$ and $\mathbf{w}$ to produce the composed embedding: $\mathbf{c}$. It's worth noting that with the Slerp-based ZS-CIR method, users have the ability to manually balance the contributions of the query image or text by adjusting the value of α to suit their retrieval purposes.

In line with the experimental results reported in previous works, such as Pic2Word [22] and SEARLE [1], we also observe a similar phenomenon where using a text-only query yields better ZS-CIR performance than an image-only query for both natural and fashion image dataset benchmarks. This observation leads us to hypothesize that text plays a more significant role than images in constructing the composed representation. As a result, we assign more weight ($\alpha \geq 0.8$; *text-weighted*) to $\mathbf{w}$ than to $\mathbf{v}$ when constructing $\mathbf{c}$, which leads to a significant performance gain in ZS-CIR evaluation protocols. That said, as we will show in Section 4, the image still plays an important role in Slerp. When the text is given all the weight, we observed significant performance degradation.
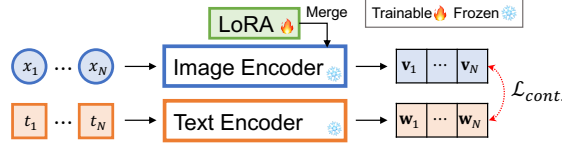
Fig. 2: Workflow of Text-Anchored-Tuning.

3.3 Text-Anchored-Tuning

In reality, the vision (image) and language (text) modality embeddings of the VLP model are distinctly distributed, and the embeddings of paired image and text samples are actually not closely aligned with each other as observed in [14]. This gap between modalities can limit the effectiveness of conducting interpolation between image and text embeddings, which may consequently degrade the Slerp-based ZS-CIR performance. To resolve this issue, we propose Text-Anchored-Tuning (TAT) method as illustrated in Figure 2.

The goal of TAT is to mitigate the performance degradation in the text-weighted Slerp-based search, which is caused by the modality gap, while preserving the powerful original text representation of the VLP model. We aim to achieve this by adjusting the image embeddings to align with the text embedding points that we keep fixed (*text-anchoring*), thereby making images closer to the corresponding texts. This is enabled by fine-tuning the image encoder E_I, while keeping the text encoder E_T unchanged. However, fully fine-tuning all parameters of E_I is not only expensive but can also lead to catastrophic forgetting of the inherent knowledge within E_I. Therefore, we adopt the LoRA technique [11] with additional trainable parameters ($\mathcal{P}_{lora}$), which retains the parameters of E_I and adds a small number of additional parameters to E_I to assist in its adaptation. In this way, the original expressive power of both E_I and E_T is preserved, and the image embeddings can be realigned to follow the text embeddings. For the training objective, we choose the same batch-wise contrastive loss (Equation 1) as used in the pretraining stage of VLP, with a fixed temperature (τ) to ensure stable training. In the end, image embeddings follow the text representation to ease the text-weighted Slerp process.

3.4 Inference

To perform composed image retrieval, we first construct a gallery. This is done by forwarding images (x_g) through the trained image encoder (E_I) to produce image embeddings ($\mathbf{v}_g = E_I(x_g)$), which are then collected to form the retrieval gallery. Next, we take a user's query of an image (x_q) and text (t_q) and forward them into E_I and the text encoder (E_T) separately, to generate the query image embedding ($\mathbf{v}_q = E_I(x_q)$) and text embedding ($\mathbf{w}_q = E_T(t_q)$). The image and text embeddings, $\mathbf{v}_q$ and $\mathbf{w}_q$, are then combined using Slerp with a balancing scale (α) to produce a query composed embedding $\mathbf{c}_q$. Finally, $\mathbf{c}_q$ is used to retrieve relevant images from the gallery by computing the cosine similarity

between $\mathbf{c}_q$ and $\mathbf{v}_g$, producing a ranked list based on the score. Note that, unlike pseudo-token based approaches [22,4,1,8], our Slerp-based method does not require a specific text prompt (*e.g.* "a photo of [\$]", where \$ is replaced with a text token transformed from an image). This requirement can often become a bottleneck for the performance of ZS-CIR. Instead, our Slerp-based search allows the user's query text to be directly applied, providing a more robust solution.

4   Experiments

4.1   Settings

Datasets. For training ZS-CIR models, previous works [22,8] have used datasets configured with web-collected image-text pairs such as CC3M [23]. We adopt three training datasets: (1) CC3M: we attempted to utilize the full scale of the CC3M dataset; however, we were only able to download 2.3M, so we use a subset of the entire package, (2) LLaVA-Align: this dataset contains image-text pairs used in LLaVA [16] to build alignment between image and text modalities, which contains 585K pairs, and (3) Laion-2M: we sub-sample 2M of the Laion-115M image-text pairs used in the pretraining of the BLIP [13] model. Unless otherwise stated, we use this dataset as our baseline for the CIR model training.

For evaluation, there are three standard benchmarks used by ZS-CIR methods [22,1,8]. The first is CIRR [18], which deals with natural images. The second is CIRCO [1], which handles more challenging natural image cases on a larger scale than CIRR and is specifically designed for ZS-CIR models. The third is FashionIQ [30], which focuses on fashion domain images. Each of these benchmarks offers unique challenges and datasets, aiding researchers in expanding the limits of what can be achieved in CIR. Specifically, *CIRR* is configured with 4,351 subgroups (subsets), each containing six similar images, sourced from NLVR2 dataset [25]. We utilize the test split of CIRR, which consists of 503 subgroups of 2,178 images, to evaluate ZS-CIR methods. *CIRCO* contains 800 image-text queries collected from the COCO 2017 unlabeled set [15], with a retrieval gallery size of 123,403 for evaluation. *FashionIQ* is divided into three categories: Dress, Shirt, and Toptee. For this benchmark, we employ the validation split, which consists of triplets of 15,415 images.
Evaluation Metrics. Following the protocols utilized in benchmarks [18,30,1], we report the CIR results with recall scores at the top K retrieval results (R@K) for CIRR and FashionIQ, and more detailed results under the collected subset (R_s@K) for CIRR. In the case of CIRCO, we follow the protocol used in [1,8] and evaluate retrieval results with a ranking-based metric, mean Average Precision (mAP@K) scores.
Implementation Details. The popular CLIP and BLIP models [21,13], which are based on a Transformer [26] backbone, serve as our baseline VLP models. These models are listed with backbone/image patch size as: CLIP-ViT-B/32, CLIP-ViT-L/14, and BLIP-ViT-L/16. For simplicity, we notate each as C-B32, C-L14 and B-L16, respectively. Pretrained weights provided by Hugging-

Table 1: Retrieval results on *CIRR test set*. The best scores are marked in bold, while the second best are underlined.

Backbone	Method	Recall@K				Recall$_{subset}$@K		
		K=1	K=5	K=10	K=50	K=1	K=2	K=3
CLIP-ViT-B/32	Image-only	6.99	23.23	34.31	58.89	20.46	40.46	60.87
	Text-only	19.61	43.90	55.42	78.27	**62.24**	**80.99**	<u>90.36</u>
	PALAVRA† [4]	16.62	43.49	58.51	83.95	41.61	65.30	80.94
	SEARLE† [1]	23.57	<u>52.80</u>	<u>66.48</u>	**89.59**	54.80	76.70	88.10
	Slerp	<u>24.22</u>	50.94	63.49	84.92	57.86	78.27	89.25
	Slerp + TAT	**28.19**	**55.88**	**68.77**	<u>88.51</u>	<u>61.13</u>	<u>80.63</u>	**90.68**
CLIP-ViT-L/14	Image-only	7.33	23.01	33.25	56.24	20.84	41.61	61.04
	Text-only	20.92	43.98	55.42	76.77	<u>60.41</u>	<u>79.74</u>	<u>90.36</u>
	Pic2Word‡ [22]	23.90	51.70	65.30	87.80	-	-	-
	SEARLE‡ [1]	24.22	52.41	66.29	<u>88.63</u>	53.71	74.63	87.61
	LinCIR‡ [8]	<u>25.04</u>	<u>53.25</u>	<u>66.68</u>	-	57.11	77.37	88.89
	Slerp	24.43	49.93	62.29	83.45	57.71	77.59	88.80
	Slerp + TAT	**30.94**	**59.40**	**70.94**	**89.18**	**64.70**	**82.92**	**92.31**
BLIP-ViT-L/16	Image-only	7.18	23.30	33.40	57.49	20.92	41.45	60.58
	Text-only	26.48	51.45	62.10	79.66	<u>66.27</u>	<u>84.53</u>	<u>92.31</u>
	Slerp	<u>28.60</u>	<u>55.37</u>	<u>65.66</u>	<u>84.05</u>	65.16	83.90	92.05
	Slerp + TAT	**33.98**	**61.74**	**72.70**	**88.94**	**68.55**	**85.11**	**93.21**

Face[3] [29] are applied to the CLIP models as: `openai/clip-vit-base-patch32`, `openai/clip-vit-large-patch14`. For the BLIP model, the original checkpoint named `BLIP w/ ViT-L` from the authors[4] is employed.

For TAT learning (refer to Section 3.3), we apply additional LoRA parameters ($\mathcal{P}_{lora}$) configured as follows: LoRA$_{\alpha}$ = 16, rank = 16, and dropout = 0.1. We train the models using eight A100-80GB GPUs. The batch size is set to 1,024, with 128 batches per GPU. During training, we keep the entire set of parameters of E_I and E_T fixed. The temperature hyper-parameter τ is set to 1/0.07. We use the AdamW optimizer [19] with a fixed learning rate of 1e-4 and a weight decay of 0.01. The training, which is conducted in a *single epoch* and where the trainable parameters constitute *less than 0.5% of the total parameters*, is performed in a highly efficient manner, *e.g.*, taking less than 0.5 hour for C-B32 backbone with LLaVA-Align dataset.

During the inference stage, we observe that the CIRR is more heavily influenced by the text than the image, compared to other datasets, a feature also noted in [22,1]. Therefore, we set the Slerp balancing scalar value α to 0.9, 0.8, and 0.8 for CIRR, CIRCO, and FashionIQ, respectively, taking into account the unique features of each dataset.

4.2   Main Results

In this section, we compare our method (Slerp and Slerp + TAT) with existing pseudo-word token based ZS-CIR works [4,1,22,8]. The symbol † denotes retrieval

---

[3] https://huggingface.co/models
[4] https://github.com/salesforce/BLIP

Table 2: Retrieval results on *CIRCO test set*. The best scores are marked in bold, while the second best are underlined.

Backbone	Method	mAP@5	mAP@10	mAP@25	mAP@50
CLIP-ViT-B/32	Image-only	1.51	1.86	2.31	2.64
	Text-only	2.50	2.61	2.91	3.09
	PALAVRA[†] [4]	4.61	5.32	6.33	6.80
	SEARLE[†] [1]	_8.86_	_9.43_	_10.55_	_11.23_
	Slerp	6.35	7.11	8.12	8.75
	Slerp + TAT	**9.34**	**10.26**	**11.65**	**12.33**
CLIP-ViT-L/14	Image-only	2.50	3.09	3.93	4.42
	Text-only	3.30	3.65	4.08	4.38
	Pic2Word[‡] [22]	8.72	9.51	10.64	11.29
	SEARLE[‡] [1]	11.68	12.73	14.33	15.12
	LinCIR[‡] [8]	_12.59_	_13.58_	_15.00_	_15.85_
	Slerp	8.76	9.84	11.30	11.99
	Slerp + TAT	**18.46**	**19.41**	**21.43**	**22.41**
BLIP-ViT-L/16	Image-only	1.52	1.89	2.42	2.77
	Text-only	4.24	4.59	5.20	5.53
	Slerp	_9.61_	_10.11_	_11.10_	_11.66_
	Slerp + TAT	**17.84**	**18.44**	**20.24**	**21.07**

results reported in [1], [‡] denotes results reported in [8], and $-$ indicates that results were not reported.

CIRR. Retrieval results on CIRR dataset are shown in Table 1. Due to the unique property of the CIR triplets in this dataset, where the textual intents of CIR triplets are frequently unrelated as noted in [22,1], text-only retrieval results achieve the highest scores under $\text{Recall}_{\text{subset}}@1, 2$ for the C-B32 backbone. When we compare Slerp with pseudo-word token-based methods, we find that the performance is comparable. The key to this performance is Slerp's ability to adjust the α parameter appropriately, making the retrieval process feasible according to the properties of the retrieval domain. Moreover, except for the C-B32 backbone Recall@50 case (where our method achieves the second-best result), our Slerp + TAT approach significantly outperforms previous works and image/text-only cases for both CLIP and BLIP-based backbones, demonstrating the benefits of both proposals.

CIRCO. For CIRCO dataset, retrieval results are shown in Table 2. Despite the increased challenge posed by CIRCO's larger retrieval gallery size and its greater emphasis on image and relative text (textual intent) in the construction of CIR triplets compared to CIRR, our Slerp + TAT approach still significantly outperforms previous methods and image/text-only cases for both CLIP and BLIP across all mAP score metrics. This performance improvement is even more pronounced when a superior VLP model is used as the backbone, as observed in the results between C-B32 and C-L14. Furthermore, it's noteworthy that simply applying Slerp without any fine-tuning can yield the third-best results with the C-B32 and C-L14 backbones. The only method that outperforms Slerp

Table 3: Retrieval results on *FashionIQ validation set*. The best scores are marked in bold, while the second best are underlined.

Backbone	Method	Dress		Shirt		Toptee		Average	
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
CLIP-ViT-B/32	Image-only	3.87	10.81	7.46	14.03	6.22	13.36	5.85	12.73
	Text-only	13.58	31.78	20.26	35.28	20.19	40.49	18.01	35.85
	PALAVRA[†] [4]	17.25	35.94	21.49	37.05	20.55	38.76	19.76	37.25
	SEARLE[†] [1]	18.15	38.57	**24.83**	41.11	25.60	46.25	22.86	41.98
	Slerp	**20.53**	41.00	23.75	40.92	**26.98**	46.77	**23.75**	42.90
	Slerp + TAT	19.24	**42.14**	23.06	**41.95**	26.57	**47.78**	22.96	**43.96**
CLIP-ViT-L/14	Image-only	4.86	12.99	11.04	20.22	8.67	16.52	8.19	16.58
	Text-only	14.33	32.57	20.46	33.61	21.72	39.32	18.84	35.17
	Pic2Word[‡] [22]	20.00	40.20	26.20	43.60	27.90	47.40	24.70	43.70
	SEARLE[‡] [1]	20.48	43.13	26.89	45.58	29.32	49.97	25.56	46.23
	LinCIR[‡] [8]	20.92	42.44	29.10	**46.81**	28.81	50.18	26.28	46.49
	Slerp	21.96	41.55	28.66	43.96	30.24	48.34	26.95	44.62
	Slerp + TAT	**23.35**	**45.12**	**29.64**	46.47	**31.97**	**51.20**	**28.32**	**47.60**
BLIP-ViT-L/16	Image-only	4.16	11.40	8.78	18.11	7.39	15.45	3.58	9.37
	Text-only	18.05	34.90	21.39	36.36	24.73	42.94	28.16	47.93
	Slerp	22.91	42.39	27.33	45.25	32.33	50.48	31.89	51.60
	Slerp + TAT	**29.15**	**50.62**	**32.14**	**51.62**	**37.02**	**57.73**	**32.77**	**53.32**

is SEARLE, which requires training with 3M image-text pairs and 5.5M text samples to achieve its retrieval results.

FashionIQ. For the fashion domain images from FashionIQ dataset, we present the experimental results in Table 3. Similar to the results with the natural image datasets, our Slerp-based retrieval results are comparable to those of pseudo-token based methods in fashion domain either. Moreover, our Slerp + TAT method achieves the best average scores in most cases for both backbone types, with the exception of C-B32 R@10, where our Slerp method shows the best results. These results demonstrate the effectiveness of our approach across diverse datasets and retrieval tasks, further validating the robustness and versatility of our proposed methods.

4.3   Further Analysis

Ablation Study. In order to validate our method, we conduct an ablation study on the Slerp and TAT schemes. For Slerp, we vary the balancing value α and displayed the results in Figure 3. From these results, we observe that increasing the weight given to text embedding improves both mAP and Recall performance, up to $\alpha = 0.8$. However, the image also plays a significant role, as evidenced by the significant drop in performance when the image has no effect ($\alpha = 1.0$).

For TAT, we perform experiments with diverse setups and display the results in Table 4. Please refer to Table 2 for a comparison with previous works. We experiment with (a, b, c, d) to evaluate the impact of the training dataset by changing it during TAT. For (a), even though TAT is trained with only 0.58M
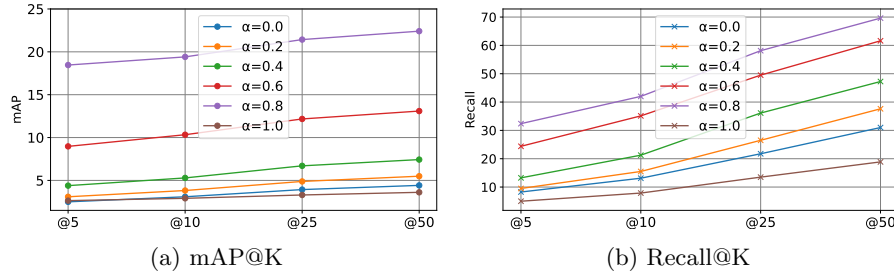
(a) mAP@K   (b) Recall@K

Fig. 3: mAP and Recall scores by varying the α of Slerp with CLIP-ViT-L/14 model on *CIRCO test set*.

Table 4: Ablation study results with CLIP-ViT-L/14 backbone on *CIRCO test set*. Except for (g, h), retrieval results are obtained with Slerp + TAT.

Ablation	mAP@5	mAP@10	mAP@25	mAP@50
(a) LLaVA-Align (0.58M)	17.05	18.23	20.11	21.05
(b) CC3M (2.3M)	16.98	17.82	19.62	20.58
(c) Laion-4M	18.45	19.50	21.53	22.52
(d) Laion-8M	18.23	19.21	21.29	22.26
(e) None-anchoring	8.26	8.90	10.07	10.71
(f) Image-anchoring	7.54	7.73	8.79	9.30
(g) Pic2Word-Laion-2M	8.93	9.96	11.50	12.02
(h) Pic2Word-LoRA	8.89	9.84	11.33	11.60

image-text pairs, which is only 20% of the CC3M used in [22,8], TAT significantly outperforms previous methods. For (b), even when only a subset of CC3M is utilized for training, our TAT still outperforms previous methods by a large margin. We employ (c, d) to test whether scaling the dataset to a larger scale than our baseline (Laion-2M) would affect the performance, and we find that it does not have much impact. For (e, f), we aim to compare text-anchoring used in TAT with no-anchor: applying LoRA to both the image and text encoder for fine-tuning, and image-anchor: applying LoRA to the text encoder and fixing the image encoder. From the results, we demonstrate that text-anchoring is the most effective training method. For (g), we apply the same Laion-2M dataset to Pic2Word [22] to verify that the dataset is not the only factor contributing to our performance, and we observe that the performance gain with Laion-2M in Pic2Word is marginal. Additionally, for (h), we apply LoRA to Pic2Word to show that LoRA is not the powerful factor in our proposed design, which also shows a marginal performance gain compared to the original Pic2Word baseline.

Applying Slerp on Fine-tuned Models. To further demonstrate the advantage of the Slerp-based ZS-CIR, we conduct an experiment with Slerp on the fine-tuned VLP model (BLIP fine-tuned on COCO [15] train set) and show the results in Table 5. It's important to note that, unlike CASE [12] and CoVR [27] which utilize additional visual data samples to improve the BLIP model's performance, our Slerp does not utilize any further training and simply interpolates

Table 5: Retrieval results with fine-tuned BLIP model on *CIRR test set*.

Method	Recall@K				Recall$_{subset}$@K		
	K=1	K=5	K=10	K=50	K=1	K=2	K=3
CASE [12]	35.40	65.78	**78.53**	**94.63**	64.29	82.66	91.61
CoVR [27]	38.48	**66.70**	77.25	91.47	69.28	83.76	91.11
Slerp	**39.08**	65.57	75.45	89.83	**72.96**	**88.10**	**94.87**

Table 6: Supervised trained retrieval results on *CIRR test set*.

Backbone	Pretrained Weight	Recall@K				Recall$_{subset}$@K		
		K=1	K=5	K=10	K=50	K=1	K=2	K=3
CLIP-ViT-B/32	Original CLIP	30.99	61.28	73.69	91.40	59.64	79.95	90.65
	TAT-trained CLIP	**33.98**	**65.06**	**76.87**	**92.63**	**61.78**	**81.52**	**91.18**
CLIP-ViT-L/14	Original CLIP	30.31	61.11	73.52	90.75	60.84	80.22	90.02
	TAT-trained CLIP	**36.58**	**67.71**	**78.19**	**92.72**	**65.52**	**84.07**	**92.89**

Table 7: Supervised trained retrieval results on *FashionIQ validation set*.

Backbone	Pretrained Weight	Dress		Shirt		Toptee		Average	
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
CLIP-ViT-B/32	Original CLIP	25.08	48.79	32.68	52.80	33.35	56.09	30.37	52.56
	TAT-trained CLIP	**25.48**	**49.98**	**33.27**	**53.29**	**34.88**	**57.22**	**31.21**	**53.50**
CLIP-ViT-L/14	Original CLIP	29.65	52.80	39.01	57.85	38.25	59.41	35.64	56.69
	TAT-trained CLIP	**30.09**	**55.73**	**39.55**	**58.93**	**40.44**	**62.47**	**36.69**	**59.04**

the original BLIP's image and text embeddings. Nevertheless, Slerp achieves the best results in Recall@1 and all of the Recall$_{subset}$ cases, verifying its effectiveness and applicability on existing fine-tuned VLP model for retrieval.

Improving Supervised CIR: Replacing VLP with TAT-trained Model. To further assess the utilization of the TAT-tuned VLP model, we apply it to supervised CIR tasks by replacing the original CLIP model in Combiner [2] training pipeline with our TAT-tuned CLIP. From the results listed in Tables 6 and 7, which cover both natural and fashion image domains, we observe a significant improvement in retrieval performance. This confirms that TAT learning of the VLP model not only improves zero-shot performance in the composed image retrieval task, but also serves as an effective initial checkpoint for supervised composed image retrieval scenarios.

Qualitative Results. Figure 4 shows retrieval results using the Slerp + TAT with C-L14 model. From the results, we can confirm that α successfully balances the contribution between the image and text query, enabling user-side adaptation. Moreover, retrieved images are highly relevant, demonstrating the zero-shot capabilities of our proposed method.

4.4   Discussion

Potential Impact and Limitation. The proposed Slerp-based ZS-CIR method and TAT strategy offer significant improvements by enhancing the alignment between image and text embeddings, which streamlines the image retrieval process

(a) Top-5 retrieval results on different α value for Slerp.

(b) Top-1 retrieval results on different α value for Slerp.
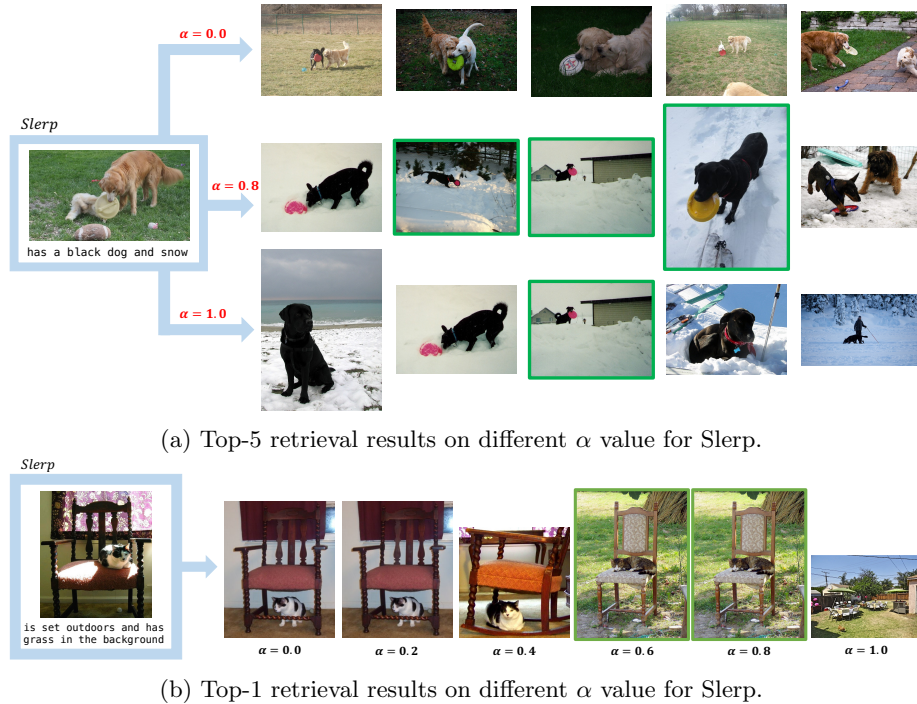
Fig. 4: Qualitative results on *CIRCO validation set*. Green box is ground truth.

and potentially improves the performance of supervised CIR with better initial checkpoints. The TAT strategy is particularly resource-efficient, requiring fewer image-text pairs and only a single training epoch to achieve decent performance. However, the current limitation of ZS-CIR methods is their untested capability in different types of composed retrieval scenarios, such as those involving both image and text in the queries and retrieval galleries. This leaves room for further research and the development of more benchmarks within the community.

5   Conclusion

In this paper, we introduce a Slerp-based zero-shot composed image retrieval method and a text-anchored-tuning strategy. Both present significant advancements in the field of vision-language compositional retrieval tasks. The Slerp-based method provides a simple yet effective solution for integrating image and text, achieving performance on par with existing methods without the need for additional training. The text-anchored-tuning strategy successfully redistributes image samples to closely align with the corresponding text samples in the embedding space, thereby enhancing the performance of the Slerp-based search. Extensive experimental results show that our approach not only achieves superior training efficiency but also demonstrates broader applicability with the state-of-the-art zero-shot composed image retrieval performance.

References

1. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-shot composed image retrieval with textual inversion. In: ICCV (2023)
2. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Effective conditioned and composed image retrieval combining clip-based features. In: CVPR (2022)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
4. Cohen, N., Gal, R., Meirom, E.A., Chechik, G., Atzmon, Y.: "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In: ECCV (2022)
5. Delmas, G., Sampaio de Rezende, R., Csurka, G., Larlus, D.: Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In: ICLR (2022)
6. Forbes, M., Kaeser-Chen, C., Sharma, P., Belongie, S.: Neural naturalist: Generating fine-grained image comparisons. In: EMNLP (2019)
7. Goenka, S., Zheng, Z., Jaiswal, A., Chada, R., Wu, Y., Hedau, V., Natarajan, P.: Fashionvlp: Vision language transformer for fashion retrieval with feedback. In: CVPR (2022)
8. Gu, G., Chun, S., Kim, W., Kang, Y., Yun, S.: Language-only efficient training of zero-shot composed image retrieval. arXiv preprint arXiv:2312.01998 (2023)
9. Guo, X., Wu, H., Cheng, Y., Rennie, S., Tesauro, G., Feris, R.: Dialog-based interactive image retrieval. NeurIPS (2018)
10. Han, X., Wu, Z., Huang, P.X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., Davis, L.S.: Automatic spatially-aware fashion concept discovery. In: ICCV (2017)
11. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2021)
12. Levy, M., Ben-Ari, R., Darshan, N., Lischinski, D.: Data roaming and early fusion for composed image retrieval. arXiv preprint arXiv:2303.09429 (2023)
13. Li, J., et al.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
14. Liang, V.W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.Y.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. NeurIPS (2022)
15. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. Springer (2014)
16. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS (2024)
17. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: ICCV (2021)
18. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: CVPR (2021)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2018)
20. Ma, J., Jiang, X., Fan, A., Jiang, J., Yan, J.: Image matching from handcrafted to deep features: A survey. IJCV **129**, 23–79 (2021)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. PMLR (2021)
22. Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T.: Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In: CVPR (2023)

23. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018)
24. Shoemake, K.: Animating rotation with quaternion curves. In: Proceedings of the 12th annual conference on Computer graphics and interactive techniques (1985)
25. Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., Artzi, Y.: A corpus for reasoning about natural language grounded in photographs. In: ACL (2019)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS (2017)
27. Ventura, L., Yang, A., Schmid, C., Varol, G.: CoVR: Learning composed video retrieval from web video captions. arXiv:2308.14746 (2023)
28. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J.: Composing text and image for image retrieval-an empirical odyssey. In: CVPR (2019)
29. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., 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)
30. Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., Feris, R.: Fashion iq: A new dataset towards retrieving images by natural language feedback. In: CVPR (2021)
31. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: CVPR (2022)