# Learning Sample-wise Rank-aware Interpolation Weights for Composed Visual Data Retrieval

Boseung Jeong[1], Taegyu Park[2], Donghyeon Kwon[2], Hyunsouk Cho[3], and Suha Kwak[2,4]

[1]AI Center, Samsung Electronics, Republic of Korea
[2]Dept. of CSE, POSTECH, Republic of Korea
[3]Dept. of AI, Ajou University, Republic of Korea
[4]Graduate School of AI, POSTECH, Republic of Korea
[1]bs0131.jeong@samsung.com, [2]{taegyu.park,kinux98,suha.kwak}@postech.ac.kr,
[3]hyunsouk@ajou.ac.kr

**Abstract.** At the heart of composed visual data retrieval is the fusion of a reference visual input and a textual modification into a single query. While current state-of-the-art methods utilize multimodal large language models for this fusion, their complexity introduces prohibitive query-time latency, limiting their scalability. We instead revisit the efficacy of simple linear interpolation within an embedding space, and introduce SRAIN, the first framework that dynamically predicts query-specific interpolation weights. The key challenge lies in the fact that the quality of an interpolation weight should be measured by the interpolated embedding's discriminability from negatives as well as its proximity to true targets; this makes collecting and predicting optimal weights intractable. We overcome this bottleneck through two key innovations: batch-wise rank-aware weight estimation during training, and a compact memory bank that synthesizes hard negatives during inference. SRAIN achieves the best in composed video retrieval and matches the current state of the art in composed image retrieval, all while substantially reducing query-time latency compared to MLLM-based alternatives.

**Keywords:** Composed Video/Image Retrieval · Multi-Modal Retrieval

## 1 Introduction

Composed image retrieval (CoIR) [44] and composed video retrieval (CoVR) [43] formulate the retrieval task as identifying a target visual instance based on a composite query: a reference visual input and an accompanying textual modification. This approach improves search specificity and expands practical applications in media editing, recommendation systems, and creative content search. Also, its inherent query editing capability has driven growing attention across various applications where modifying existing media is preferred over initiating a completely new search.

Both of CoIR and CoVR demand a principled fusion of the reference and modification text into a single query that reflects the intended edit while preserving the salient content of the reference. A large body of research has been

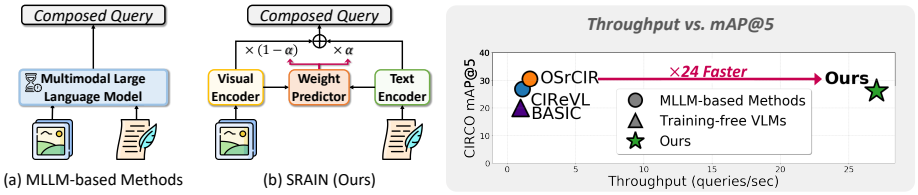


**Fig. 1:** Comparison of SRAIN with MLLM-based methods. (a) MLLM-based methods process reference and modification text together through MLLMs to generate a composed query, introducing substantial computational overhead. (b) SRAIN predicts a sample-specific interpolation weight $\alpha$ and fuses reference and modification embeddings through linear interpolation, enabling efficient retrieval. The right plot shows that SRAIN achieved comparable performance (mAP@5 on CIRCO) while being $\times 24$ faster in throughput compared to MLLM-based methods. Meanwhile, BASIC [26], a training-free method without MLLM, underperformed in throughput as well as retrieval accuracy since it requires numerous text encoder forward passes.

devoted to the multimodal fusion in this context, *e.g.*, via cross-attention between visual and textual embeddings [36, 39, 40, 45, 46, 49, 52, 55] or textual inversion that encodes visual data in the text embedding space [22, 28, 37]. Currently, the best-performing approach leverages multimodal large language models (MLLMs) [1, 6, 21, 23] to rewrite the paired inputs into a single natural language description, which is in turn used for text-based retrieval [12, 15, 39, 40, 50]. However, the use of MLLMs substantially increases query-time latency due to their architectural complexity, limiting the practical utility of this approach.

Meanwhile, it has been recently reported that fusing the reference and modification embeddings by linear interpolation on a unit hyper-sphere can yield effective composed queries without auxiliary fusion modules [14]. This work manually seeks a single interpolation weight for each dataset, which is however highly sub-optimal since the impact of the modification text should be determined by the specific content of the reference and the intent of the composite query.

To overcome this drawback while retaining the efficiency of linear interpolation, we introduce a novel framework, dubbed **S**ample-wise **R**ank-**A**ware **IN**terpolation for multimodal fusion (SRAIN), that for the first time learns to determine how strongly a textual modification should influence the reference in an instance-wise manner. At inference, a lightweight module predicts an interpolation weight for each composite query, and its reference and modification embeddings are linearly interpolated using the predicted weight. The resulting fused embedding is compared with target embeddings through cosine similarity for retrieval. This pipeline enables efficient inference without any dependence on external models such as MLLMs (Fig. 1).

However, training the interpolation weight predictor is challenging, primarily due to the lack of ground-truth supervision. Identifying the optimal interpolation weight is computationally intractable; such a value cannot be determined solely by observing the composite query, but should be estimated by evaluating how


**Fig. 2:** Rank-aware optimal weight estimation. Rather than determining the weight solely based on the similarity to the positive target (a), SRAIN selects the weight based on rank, where the positive consistently outranks nearby hard negatives (b).

effectively the interpolated embedding discriminates true targets from hard negatives. Given the massive scale of the dataset, verifying this discriminative power against all positive and negative examples for every potential weight candidate is computationally intractable. To tackle this problem, we propose a batch-wise rank-aware weight estimation strategy (Fig. 2). This strategy derives a near-optimal weight from ranking behavior within each mini-batch of a sufficiently large size, and offers the estimated weight as the ground-truth label.

Since the ground-truth weight depends on relative ranks with hard negatives in the batch, it is infeasible to infer the optimal weight solely from the reference and modification embeddings. We mitigate this issue by introducing a conditioned weight prediction mechanism, which enables the weight predictor to reference hard negatives in both training and testing (Fig. 3). During training, the weight predictor takes as inputs a positive target and hard negatives within the batch as well as the composite query embeddings. At inference time, it samples synthetic negative embeddings from a compact memory bank that stores representative target embeddings of training data.

Building on these designs, we present a two-stage training strategy. In the first stage, the visual and text encoders are fine-tuned by aligning the interpolated embedding with its corresponding target embedding while pushing apart negatives through a hard-negative contrastive loss [27]. In the second stage, the encoders are frozen, and only the weight predictor is trained to regress the ground-truth interpolation weights. This separation is essential since the joint optimization of the encoders and the weight predictor causes the ground-truth interpolation weights to fluctuate as the encoder representation changes, which results in inconsistent supervision and unstable training.

SRAIN achieved the state of the art on the CoVR benchmark [43], and matched the best model on the CoIR benchmarks [3, 24, 47], all without relying on external models such as LLMs. More importantly, it substantially reduces

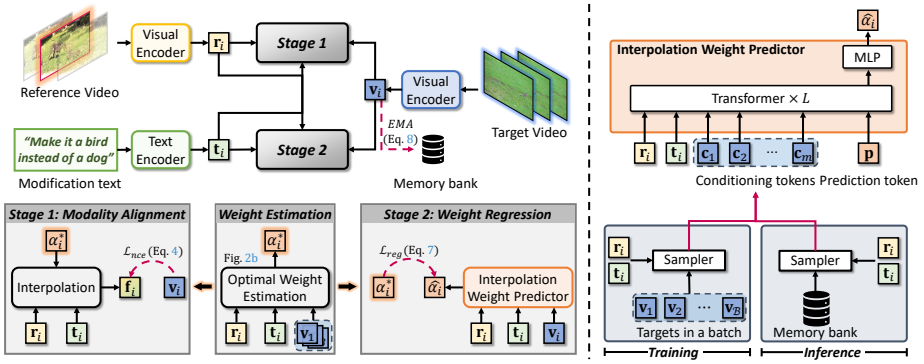


**Fig. 3:** Overall architecture of SRAIN. The reference video, modification text, and target video are encoded using the pretrained encoders. For each batch, optimal interpolation weights are estimated in a rank-aware manner. The model is trained in two stages: In *Stage* 1, the encoders are fine-tuned to align composite query embeddings with target embeddings through hard-negative contrastive learning. In *Stage* 2, the encoders are frozen and a lightweight weight predictor is trained to regress $\alpha_i^*$. A memory bank is updated via exponential moving average (EMA) during training to collect representative target prototypes. At inference, the memory bank supplies top-$n$ prototypes as auxiliary tokens, enabling consistent weight prediction.

query-time latency compared to recent methods using MLLMs across both tasks. The main contribution of this paper is three-fold:

- We propose SRAIN, the first framework to dynamically predict a per-sample interpolation weight to fuse the reference and modification embeddings while preserving computational efficiency.
- We address the lack of ground-truth supervision and the dependence of optimal weights on hard negatives by introducing a rank-aware weight estimation procedure that derives near-optimal weights from batch-wise ranking behavior, and a conditioned weight prediction mechanism that explicitly conditions on hard negatives.
- SRAIN achieves the state of the art on the CoVR benchmark, while matching the best model on the CoIR benchmarks without reliance on external models, reducing query-time latency substantially.

## 2 Related Work

### 2.1 Composed Image Retrieval

Composed Image Retrieval (CoIR) [44] aims to retrieve a target image given a reference image and a textual modification that specifies the desired change. This task extends traditional image-text retrieval [35] to a compositional setting and requires both modality alignment and multimodal fusion. Prior methods [4, 24, 36, 45, 46, 49, 52, 53, 55] have adopted the cross-attention mechanism to integrate visual and textual cues. Another line of studies [15, 38] has leveraged MLLMs

to compose visual and textual semantics without task-specific training. These methods have demonstrated strong generalization capability without training, but they depend on computationally intensive inference of large models. More recently, CoLLM [13] leverages LLMs both for triplet synthesis and for composed query embedding, introducing significant query-time latency. ConText-CIR [48] employs cross-attention with a concept-consistency loss to align noun phrases in the modification text with relevant image regions. Zero-shot CoIR methods [14, 22, 26, 28, 37] have shown that manipulating representations in the embedding space can effectively perform CoIR without explicitly merging cross-modal data. Pic2Word [28], Context-I2W [37], and FTI4CIR [22] introduced textual inversion to map an image into text tokens for later text-to-image retrieval. BASIC [26] introduces score-level late fusion of image-to-image and text-to-image similarities in a training-free manner, without relying on MLLMs at inference time. However, BASIC shows low throughput due to the numerous text encoder forward passes. Meanwhile, Slerp [14] employed spherical linear interpolation between visual and textual embeddings, showing that simple interpolation can yield competitive performance. However, it employs a single dataset-level interpolation weight that overlooks sample-level diversity. We address this issue by adaptively predicting a sample-wise, rank-aware interpolation weight.

### 2.2 Composed Video Retrieval

Composed Video Retrieval (CoVR) [43] aims to retrieve a target video given a reference video and a textual description of the intended modification. [43] introduced WebVid-CoVR, a large-scale dataset with video-text-video triplets derived from web video captions, and CoVR, a BLIP-based [20] model trained with contrastive learning [27]. CoVR-2 [42] enhanced CoVR by leveraging BLIP-2 [19] as a more capable multimodal backbone and introducing a caption-retrieval objective to provide auxiliary supervision. [40] and [39] proposed frameworks that augment WebVid-CoVR with additional captions describing the difference between reference and target videos generated by an LLM, and use cross-attention mechanism to integrate video and textual representations within a unified grounding encoder. TFR-CVR [12] adopted a training-free pipeline that combines LLM-generated captions with visual candidate filtering and text-based re-ranking, focusing on zero-shot generalization rather than end-to-end fusion learning.

While these approaches have advanced compositional understanding, they depend on large language models to fuse the reference and textual cues, which substantially increases query-time latency. To overcome this limitation, our method adopts a lightweight and interpolation-based fusion that avoids introducing a heavy inference step or cross-modal modules.

## 3 Proposed Method

This section presents details of our framework, SRAIN; its overall architecture is depicted in Fig. 3. We first describe the embedding extraction process for the

reference and target visual data and the modification text in Sec. 3.1, and then elaborate on the rank-aware interpolation weight estimation strategy in Sec. 3.2. The two-stage training strategy of SRAIN is then introduced in Sec. 3.3. Finally, Sec. 3.4 describes the weight prediction procedure at inference time.

### 3.1 Embedding Extraction

Following CoVR-2 [42], SRAIN employs BLIP-2 [19] as the pretrained vision-language model, which consists of a ViT image encoder [8] and a Q-Former [19] that projects visual and textual representations into a shared embedding space. In contrast to CoVR-2 that jointly processes the reference visual input and the modification text within BLIP-2, SRAIN encodes each modality independently. Extracting separate embeddings for the reference and modification allows efficient linear interpolation in the shared embedding space without introducing additional fusion modules. More architectural details of our embedding networks are provided in the supplement.

**Reference Visual Embedding.** A reference image is used as-is, and for a reference video, we use its middle frame as input following prior work [42]. The image is encoded by the frozen ViT encoder, producing patch embeddings that are subsequently queried by the Q-Former using $N_q$ learnable tokens. The resulting token outputs are passed through the BLIP-2 projection layer, averaged across tokens, and $\ell_2$-normalized to obtain the reference visual embedding $\mathbf{r} \in \mathbb{R}^d$.

**Modification Text Embedding.** The modification text is tokenized and processed by the Q-Former without visual conditioning. The hidden state of the first token is projected by the BLIP-2 text projection layer and $\ell_2$-normalized to yield the modification text embedding $\mathbf{t} \in \mathbb{R}^d$.

**Target Visual Embedding.** For a target image, we directly encode the image using the frozen ViT and Q-Former, producing $N_q$ token embeddings. These tokens are averaged and $\ell_2$-normalized to obtain the target visual embedding $\mathbf{v} \in \mathbb{R}^d$. For a target video, we uniformly sample $N$ frames and encode each frame using the frozen ViT encoder. The resulting patch embeddings are queried by the frozen Q-Former with $N_q$ learnable tokens, producing $N_q$ token embeddings per frame ($N \times N_q$ tokens in total). To aggregate temporal information, query scoring [2,42] is applied separately for each query token: for each token, per-frame weights are computed based on the similarity between its embeddings across frames and the modification text embedding, and a weighted average over the frames produces a single temporally-aggregated token. This yields $N_q$ aggregated tokens, which are then averaged and $\ell_2$-normalized to obtain the target visual embedding $\mathbf{v} \in \mathbb{R}^d$. Since the BLIP-2 encoders remain frozen, these weights can be precomputed and cached.[1]

[1] Following CoVR-2 [42], we freeze BLIP-2 for target visual data to reduce the computational cost for both training and inference.

### 3.2 Interpolation Weight Estimation

SRAIN learns to predict sample-specific linear interpolation weights between reference visual embeddings and modification text embeddings to form composed query embeddings. The major challenge in this direction is the absence of ground-truth interpolation weights. The optimal interpolation weight must prioritize the discriminative power of the interpolated query embedding against negative samples, instead of merely maximizing its similarity to the true target embeddings. In other words, the optimal weight is the one that achieves the most desirable ranking behavior, where positive targets are closer to the composed query than negative ones. Hence, the search for the optimal interpolation weight involves comparison with all training data (or all data within the search database during testing), which results in a complexity proportional to the square of the dataset size, and thus determining the optimal interpolation weights is impractical.

To address this, we approximate the optimal weights within each mini-batch of a sufficiently large size. By leveraging hard negatives present in the batch, we can effectively estimate near-optimal interpolation weights that reflect the rank-aware objective in practice. The weight estimation procedure is illustrated in Fig. 2. Given a batch of $B$ triplets, we consider a discrete set of $K$ interpolation candidates $\{\alpha^k\}_{k=1}^K$ uniformly spaced in $[0, 1]$. Inspired by previous work [14,33], we adopt spherical linear interpolation to fuse the reference visual and modification text embeddings. For each triplet $i$ and candidate weight $\alpha^k$, the composed query embedding $\mathbf{f}_i^k$ is computed by

$$\mathbf{f}_i^k = \frac{\sin((1-\alpha^k)\theta_i)}{\sin(\theta_i)}\mathbf{r}_i + \frac{\sin(\alpha^k\theta_i)}{\sin(\theta_i)}\mathbf{t}_i, \tag{1}$$

where $\theta_i$ is the angle between $\mathbf{r}_i$ and $\mathbf{t}_i$. The fused query $\mathbf{f}_i^k$ is compared with all target embeddings $\{\mathbf{v}_j\}_{j\in B}$ in the batch using cosine similarity as follows:

$$s_{i,j,k} = \frac{\mathbf{f}_i^{k\top}\mathbf{v}_j}{\|\mathbf{f}_i^k\|_2 \, \|\mathbf{v}_j\|_2}. \tag{2}$$

For the $i$-th query, we construct a score matrix $\mathbf{S}_i = \left[s_{i,j,k}\right]_{B\times K}$ where each column corresponds to one interpolation candidate.

For each column $k$ of $\mathbf{S}_i$, which corresponds to the similarity scores between the query fused with $\alpha^k$ and all targets in the batch, $\text{rank}_i^k$ denote the rank of the positive target ($j = i$) among the $B$ similarity scores after sorting them in descending order. In other words, $\text{rank}_i^k \in \{1, \ldots, B\}$ denotes the index of the positive target in the sorted list, where a smaller value indicates better retrieval performance. The approximate optimal interpolation weight for the $i$-th query is then selected as

$$\alpha_i^* := \alpha^z, \text{where } z = \underset{k\in[1,K]}{\text{argmin}} \, \text{rank}_i^k. \tag{3}$$

In cases where multiple candidates achieve the minimum rank, we set $\alpha_i^*$ to their median. The selected $\alpha_i^*$ then serves as the ground-truth label.

### 3.3 Two-stage Training Strategy

Training SRAIN is conducted in two stages. In the first stage, the Q-Former [19] is fine-tuned with the ground-truth weights to achieve *modality alignment*, *i.e.*, aligning the fused query embeddings with their corresponding target embeddings while pushing apart negatives. In the second stage, the learned Q-Former is frozen, and a lightweight module for interpolation weight prediction is trained to regress the ground-truth weights. This two-stage design is motivated by the need to freeze the encoder during the weight predictor training. Since the ground-truth weights are computed based on ranking behavior within the feature space of the encoder, updating the encoder simultaneously with the weight predictor causes the estimated weights to vary even for the same query pair, making it impossible to provide consistent supervision to the weight predictor.

Following Eq. (1), the fused query embedding $\mathbf{f}_i \in \mathbb{R}^d$ for each triplet $i$ is obtained by spherical linear interpolation between the reference visual embedding $\mathbf{r}_i$ and the modification text embedding $\mathbf{t}_i$ using the ground-truth interpolation weight $\alpha_i^*$. The goal of the first stage is to align $\mathbf{f}_i$ with its corresponding target visual embedding $\mathbf{v}_i \in \mathbb{R}^d$. To this end, we employ the hard-negative contrastive loss [27], which encourages higher similarity for positive pairs while suppressing hard negatives through adaptive weighting based on their hardness. Given a batch with $B$ triplets, the loss is formulated as:

$$\begin{aligned}\mathcal{L}(\mathcal{B}) = &-\sum_{i=1}^{B}\left(\log\frac{e^{s_{ii}/\tau}}{\gamma\cdot e^{s_{ii}/\tau}+\sum_{j=1,j\neq i}^{B} e^{s_{ij}/\tau} w_{ij}}\right)\\ &-\sum_{i=1}^{B}\left(\log\frac{e^{s_{ii}/\tau}}{\gamma\cdot e^{s_{ii}/\tau}+\sum_{j=1,j\neq i}^{B} e^{s_{ji}/\tau} w_{ji}}\right),\end{aligned} \tag{4}$$

where $s_{ij}$ denotes the cosine similarity between the fused query embedding $\mathbf{f}_i$ and the target visual embedding $\mathbf{v}_j$.

In the second stage, the interpolation weight predictor is trained to predict the ground-truth interpolation weight $\alpha_i^*$ given the query inputs while the pre-trained encoders are fixed. Predicting the interpolation weight solely from the reference visual embedding and the modification text embedding is challenging since the ground-truth weight is estimated in a rank-aware manner that explicitly depends on hard negatives within the batch. To address this, as illustrated in Fig. 3, we introduce a conditioned weight prediction mechanism that enables the weight predictor to infer more accurate weights by conditioning on informative negative samples. Specifically, the weight predictor takes as auxiliary input a conditioning token set $C_i$ that consists of the positive target embedding $\mathbf{v}_i$ and $m-1$ hardest negative target embeddings, *i.e.*, those of negative targets most similar with the fused query embedding $\mathbf{f}_i$.

The interpolation weight predictor consists of a transformer encoder [41] with $L$ layers followed by a three-layer MLP. It takes as input the reference embedding $\mathbf{r}_i^\top \in \mathbb{R}^{1\times d}$, the modification embedding $\mathbf{t}_i^\top \in \mathbb{R}^{1\times d}$, the conditioning token set $C_i \in \mathbb{R}^{m\times d}$, and a learnable prediction token $\mathbf{p} \in \mathbb{R}^{1\times d}$. The prediction

token aggregates information from all inputs through the transformer encoder to produce an output embedding:

$$\mathbf{p}' = \texttt{Transformer}\big([\mathbf{r}_i^\top, \mathbf{t}_i^\top, C_i, \mathbf{p}]; \theta_t\big), \tag{5}$$

where $\theta_t$ denotes the encoder parameters and $\mathbf{p}' \in \mathbb{R}^{1\times d}$ is the output embedding. This embedding is then fed into the MLP, which projects it to a scalar and applies a sigmoid function to produce the predicted weight:

$$\hat{\alpha}_i = \sigma\big(\texttt{MLP}(\mathbf{p}')\big). \tag{6}$$

The details of the architecture of the predictor are presented in the supplementary material. The predictor is optimized by minimizing the mean squared error between the predicted and the ground-truth interpolation weights as

$$\mathcal{L}_{\text{reg}} = \frac{1}{B}\sum_{i=1}^{B}(\hat{\alpha}_i - \alpha_i^*)^2. \tag{7}$$

### 3.4 Weight Prediction at Inference

At inference time, the positive and negative target embeddings are unavailable, preventing direct construction of the conditioning token set. To address this, we introduce a memory bank $\mathcal{M} = \{\mathbf{m}_k\}_{k=1}^{|\mathcal{M}|}$ of prototype embeddings $\mathbf{m}_k \in \mathbb{R}^d$ that represent the distribution of target visual data for training. During training, each prototype is updated by exponential moving average with momentum $\lambda$:

$$\mathbf{m}_z \leftarrow \lambda \cdot \mathbf{m}_z + (1-\lambda)\cdot \mathbf{v}_i,\ z = \underset{k\in[1,|\mathcal{M}|]}{\operatorname{argmax}} \frac{\mathbf{m}_k^\top \mathbf{v}_i}{\|\mathbf{m}_k\|_2 \|\mathbf{v}_i\|_2}. \tag{8}$$

After training, the memory bank is frozen and used during inference to provide conditioning tokens for the weight predictor. To select informative prototypes regardless of the fusion weight, we fuse the query with multiple candidate interpolation weights uniformly sampled from $[0, 1]$, similar to Sec. 3.2. For each candidate weight, we retrieve the top-$k$ most similar prototypes from the memory bank. We then select the $n$ prototypes that are most frequently retrieved across all candidate weights, which are likely to be hard negatives. This approach maintains consistency between training and inference by providing informative reference points that enable accurate weight prediction without access to batch negatives. The justification for this approximation and detailed analysis of the memory bank's capacity are provided in the supplement.

## 4 Experiments

This section first describes the experimental setup, including datasets, evaluation metrics, and implementation details in Sec. 4.1. Then Sec. 4.2 provides quantitative results to demonstrate the effectiveness of SRAIN compared to previous work. Next, we include the qualitative results of SRAIN in Sec. 4.3, and finally analyze the impact of SRAIN's components through ablation studies in Sec. 4.4.

**Table 1:** Composed video retrieval results on the WebVid-CoVR-Test, where all compared methods are fine-tuned on the WebVid-CoVR training set. Compared methods differ in fusion strategy: Avg indicates element-wise averaging of reference and modification embeddings, MLP denotes a multi-layer perceptron-based fusion module, CA employs cross-attention, and LI refers to linear interpolation. External Knowledge denotes the use of external models for additional video descriptions generation.

| Methods | External Knowledge | Modality | Fusion | Backbone | Evaluation Metric | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | R@1 | R@5 | R@10 | R@50 |
| Text-only ($\alpha = 1$) | ✗ | Text | - | BLIP-2 | 23.32 | 46.21 | 56.22 | 78.36 |
| Visual-only ($\alpha = 0$) | ✗ | Visual | - | BLIP-2 | 36.03 | 64.24 | 74.77 | 92.64 |
| Average ($\alpha = 0.5$) | ✗ | Visual+Text | Avg | BLIP-2 | 58.69 | 83.88 | 90.49 | 98.21 |
| CoVR [43] | ✗ | Visual+Text | MLP | BLIP | 50.59 | 74.65 | 83.57 | 95.46 |
| CoVR-2 [42] | ✗ | Visual+Text | MLP | BLIP-2 | 51.88 | 79.38 | 86.46 | 97.42 |
| CoVR [43] | ✗ | Visual+Text | CA | BLIP | 55.95 | 81.22 | 89.05 | 98.08 |
| CoVR-2 [42] | ✗ | Visual+Text | CA | BLIP-2 | 59.82 | 83.84 | **91.28** | 98.24 |
| Thawakar *et al.* [40] | MiniGPT4 [56] | Visual+Text | CA | BLIP | 60.12 | 84.32 | 91.27 | **98.72** |
| SRAIN (Ours) | ✗ | Visual+Text | LI | BLIP-2 | **61.07** | **84.90** | 91.20 | 98.40 |

### 4.1 Experimental Setup

**Datasets.** We evaluate and compare the performance of our method against previous work on one CoVR benchmark, WebVid-CoVR [43], and three CoIR benchmarks, FashionIQ [47], CIRCO [3] and CIRR [24]. **WebVid-CoVR** is a large-scale CoVR dataset containing 1.64M video-text-video triplets for training and 2,556 triplets for testing, derived from web video captions. Each triplet consists of a reference video, a textual modification describing the desired change, and a target video. **FashionIQ** is a CoIR dataset in the fashion domain, containing 30,134 triplets from 77,684 images across three categories: Dress, Shirt, and Toptee. The images are split into 6:2:2 for training, validation, and test. Each query pairs a reference image with natural language feedback describing desired modifications such as changes in style, color, or pattern. We report the validation recalls because the labels of the evaluation split have not been publicly released. **CIRCO** is a zero-shot CoIR benchmark built on COCO images, containing 800 test queries with a retrieval gallery size of 123,403 for evaluation. The benchmark is specifically designed to assess model generalization capability with test queries. We follow the standard zero-shot evaluation protocol, where models are not fine-tuned on CIRCO. **CIRR** is a CoIR dataset sharing the same triplet structure as FashionIQ, but covering open-domain scenes rather than a specific category. It is built on real-life images sourced from the NLVR2 dataset, containing 21,552 images and 36,554 triplets split in an 8:1:1 ratio for training, validation, and test.

**Evaluation Metrics.** We employ the standard metric of recall at $K$ (R@$K$) to evaluate retrieval performance for both composed video retrieval and composed image retrieval. Specifically, we report R@$K$ with $K$=1, 5, 10, 50 for WebVid-CoVR, $K$=10, 50 for FashionIQ and $K$=1, 5, 10 for CIRR. In all datasets, samples are ranked based on their similarities to the query. For CIRCO, we follow the protocol of Baldrati *et al.* [3] and evaluate our method with a ranking-based metric, mean average precision at $K$ (mAP@$K$, with $K$=5, 10, 25, 50).

**Table 2:** Composed image retrieval results on the FashionIQ validation set. CC3M [32], LAION2B (L2B) [29], LAION2M (L2M), and LLaVA-Align (L-A) [23] contain 3M, 2.3B, 2M, and 585k image-text pairs, respectively, where LAION2M is a subset of LAION400M [30]. MagicLens36M [53] comprises 36M (query image, instruction, target image) triplets. SDP [51] provides 2.47M text prompts, while ST18M [10] and LaSCo [18] comprise 18.8M and 389k triplets for training, respectively.

| Methods | ViT | External Knowledge | Pretraining Data | Dress | | Shirt | | Toptee | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | R@10 | R@50 | R@10 | R@50 | R@10 | R@50 | R@10 | R@50 |
| ***Zero-shot*** | | | | | | | | | | | |
| CIReVL [15] | G | GPT-4 [1] | ✗ | 27.07 | 49.53 | 33.71 | 51.42 | 35.80 | 56.14 | 32.19 | 52.36 |
| OSrCIR [38] | G | GPT-4o [1] | ✗ | 33.02 | 54.78 | 38.65 | 54.71 | 41.04 | 61.83 | 37.57 | 57.11 |
| Pic2Word [28] | L | ✗ | CC3M | 20.00 | 40.20 | 26.20 | 43.60 | 27.90 | 47.40 | 24.70 | 43.70 |
| SEARLE-XL [3] | L | ✗ | CC3M | 26.89 | 45.58 | 20.48 | 43.13 | 29.32 | 49.97 | 25.56 | 46.23 |
| TAT [14] | L | ✗ | CC3M+L2M+L-A | 29.15 | 50.62 | 32.14 | 51.62 | 37.02 | 57.73 | 32.77 | 53.32 |
| MagicLens [53] | L | ✗ | MagicLens36M | 25.50 | 46.10 | 32.70 | 53.80 | 34.00 | 57.70 | 30.70 | 52.50 |
| LinCIR [11] | G | ✗ | CC3M+SDP | 38.08 | 60.88 | 46.76 | 65.11 | 50.48 | 71.09 | 45.11 | 65.69 |
| CoVR [43] | L | ✗ | WebVid-CoVR | 21.95 | 39.05 | 30.37 | 46.12 | 30.78 | 48.73 | 27.70 | 44.63 |
| CoVR-2 [42] | G | ✗ | WebVid-CoVR | - | - | - | - | - | - | 27.78 | - |
| SRAIN (Ours) | G | ✗ | WebVid-CoVR | **23.22** | **43.33** | **33.14** | **52.01** | **31.91** | **49.03** | **29.42** | **48.12** |
| ***Fine-tuning on FashionIQ*** | | | | | | | | | | | |
| CompoDiff [10] | G | ✗ | ST18M+L2B | 38.39 | 51.03 | 41.68 | 56.02 | 45.70 | 57.32 | 39.81 | 51.90 |
| CoVR [43] | L | ✗ | WebVid-CoVR | 44.55 | 69.03 | 48.43 | 67.42 | 52.60 | 74.31 | 48.53 | 70.25 |
| CASE [18] | G | ✗ | LaSCo | 47.44 | 69.36 | 48.48 | 70.23 | 50.18 | 72.24 | 48.79 | 70.68 |
| MAAF [7] | L | ✗ | ✗ | 23.80 | 48.60 | 21.30 | 44.20 | 27.90 | 53.60 | 24.30 | 48.80 |
| LF-BLIP [18] | L | ✗ | ✗ | 25.31 | 44.05 | 25.39 | 43.57 | 26.54 | 44.48 | 25.75 | 43.98 |
| CoSMo [17] | L | ✗ | ✗ | 25.64 | 50.30 | 24.90 | 49.18 | 29.21 | 57.46 | 26.58 | 52.31 |
| DCNet [16] | L | ✗ | ✗ | 28.95 | 56.07 | 23.95 | 47.30 | 30.44 | 58.29 | 27.78 | 53.89 |
| FashionVLP [9] | L | ✗ | ✗ | 32.42 | 60.29 | 31.89 | 58.44 | 38.51 | 68.79 | 34.27 | 62.51 |
| CLIP4CIR [4] | L | ✗ | ✗ | 31.63 | 56.67 | 36.36 | 58.00 | 38.19 | 62.42 | 35.39 | 59.03 |
| CoVR-2 [42] | G | ✗ | ✗ | 45.25 | 68.86 | 49.95 | 69.95 | **51.37** | **72.56** | 48.86 | **70.46** |
| SRAIN (Ours) | G | ✗ | ✗ | **46.25** | **69.07** | **51.23** | **70.00** | 50.30 | 71.75 | **49.26** | 70.27 |

**Implementation Details.** We use pretrained BLIP-2 [19] with ViT-G/14 [8] and Q-Former [19] as our vision-language backbone following CoVR-2 [42]. SRAIN is trained for 5 epochs for each stage with batch size 512 using AdamW optimizer [25] with learning rates of 2e−5 for Q-Former and 1e−3 for other parameters. The interpolation weight predictor comprises a 2-layer Transformer encoder and a 3-layer MLP, and contains 1.63M parameters, which corresponds to only 0.14% of the total model size. During training, it takes 50 conditioning tokens consisting of 1 positive and 49 batch-level hard negatives. Since the hard negatives are only processed by this lightweight predictor, they introduce a negligible computational overhead of merely 0.03 ms per query, accounting for 0.08% of the total wall-clock time. We set $|\mathcal{A}| = 101$ in Sec. 3.2. The memory bank has 1,024 prototypes and is updated by $\lambda = 0.99$ in Eq. (8). For the hard-negative contrastive loss in Eq. (4), we set $\gamma = 1$, $\tau = 0.07$, and compute $w_{ij}$ following [27] with $\beta = 0.5$. These settings apply to WebVid-CoVR [43], FashionIQ [47] and CIRR [24]. More implementation details are presented in the supplement.

### 4.2 Quantitative Results

**Composed Video Retrieval.** We evaluate SRAIN on the WebVid-CoVR-Test benchmark and compare it with state-of-the-art methods in Tab. 1. SRAIN sur-

**Table 3:** Zero-shot composed image retrieval results on CIRCO.

| Methods | ViT | External Knowledge | mAP@$K$ | | | |
|---|---|---|---|---|---|---|
| | | | $K$=5 | $K$=10 | $K$=25 | $K$=50 |
| TFCIR [34] | G | ✓ | 26.52 | 28.25 | 31.23 | 31.99 |
| CIReVL [15] | G | ✓ | 26.77 | 27.59 | 29.96 | 31.03 |
| OSrCIR [38] | G | ✓ | 30.47 | 31.14 | 35.03 | 36.59 |
| SEARLE-XL [3] | L | ✗ | 11.68 | 12.73 | 14.33 | 15.12 |
| CompoDiff [10] | G | ✗ | 15.33 | 17.71 | 19.45 | 21.01 |
| LinCIR [11] | G | ✗ | 19.71 | 21.01 | 23.13 | 24.18 |
| BASIC [26] | G | ✗ | 19.98 | 20.88 | 22.85 | 23.71 |
| CoVR [43] | L | ✗ | 21.43 | 22.33 | 24.47 | 25.48 |
| CoVR-2 [42] | G | ✗ | 25.06 | - | - | - |
| SRAIN (Ours) | G | ✗ | **26.16** | **27.34** | **29.94** | **30.81** |

passes all existing methods in terms of R@1. Notably, SRAIN outperforms [40], which uses cross-attention fusion and external MLLM-generated captions, by 0.95% in R@1. Compared to CoVR [43] and CoVR-2 [42] that adopt cross-attention fusion, SRAIN achieves significant improvements of 5.12% and 1.25% in R@1, respectively, using a simple but effective linear interpolation strategy. These results demonstrate that SRAIN enables effective multimodal fusion without relying on computationally expensive cross-attention or external models.

**Composed Image Retrieval.** We further assess SRAIN on FashionIQ [47], CIRCO [3] and CIRR [24], as summarized in Tab. 2, Tab. 3 and Tab. 4, respectively. In all tables, we explicitly report the ViT backbone size for each method, where L and G denote ViT-L and ViT-G, respectively. For methods appearing in both tables, the backbone and pretraining data in Tab. 3 are identical to those reported in Tab. 2. On FashionIQ, SRAIN is evaluated in both the zero-shot and fine-tuned settings. Our model trained solely on WebVid-CoVR achieves competitive performance in the zero-shot setting. While most existing methods show strong results, their zero-shot performance is largely attributed to pretraining on significantly larger image–text corpora or leveraging external knowledge. In contrast, our zero-shot evaluation is based on a model trained with the same pretraining data used for composed video retrieval. We note that WebVid-CoVR consists solely of web video triplets with limited diversity and does not inherently favor out-of-domain image benchmarks. The strong zero-shot results reflect the genuine generalization capability of SRAIN rather than data-related confounders. For fine-tuning, SRAIN outperforms previous methods in average R@10. On CIRCO, SRAIN outperforms all methods that do not use external knowledge in all metrics, and remains competitive with approaches that rely on external knowledge and incur heavy computational cost.

**Table 4:** Composed image retrieval results on CIRR.

| Methods | ViT | R@$K$ | | |
|---|---|---|---|---|
| | | $K$=1 | $K$=5 | $K$=10 |
| CLIP4CIR [4] | L | 33.59 | 65.35 | 77.35 |
| CompoDiff [10] | G | 32.39 | 57.61 | 77.25 |
| CoVR [43] | L | 49.69 | 78.60 | 86.77 |
| CoVR-2 [42] | G | 50.87 | 80.80 | 88.84 |
| SRAIN (Ours) | G | **50.95** | **81.21** | **89.60** |

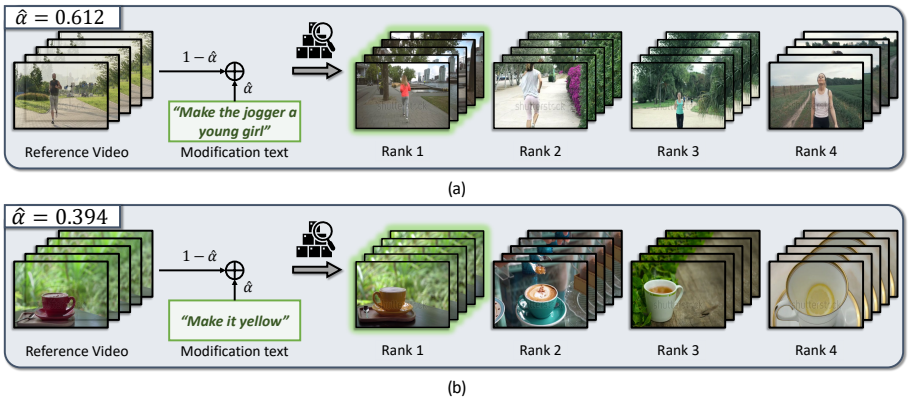


**Fig. 4:** Top-4 retrieval results on the WebVid-CoVR test set. The ground-truth target video is highlighted in green.

On CIRR, where all methods are evaluated under the supervised setting, SRAIN surpasses all methods across all retrieval metrics.

### 4.3 Qualitative Results

Fig. 4 illustrates the Top-4 retrieved video results from our method for the given query. In Fig. 4(a), the modification text "Make the jogger a young girl" is highly discriminative and sufficient to retrieve the correct target, where the predicted weight $\hat{\alpha}$ is relatively high, focusing more on the modification. In contrast, Fig. 4(b) presents a case where the text "Make it yellow" is visually ambiguous, leading the model to assign a lower $\hat{\alpha}$ and rely more on the reference video to retrieve the correct target. These results highlight the benefit of our framework in adaptively controlling the influence of visual and textual cues, enabling precise and context-aware fusion tailored to each query.

### 4.4 Ablation Studies

We evaluate the effectiveness of the proposed components in SRAIN through comprehensive experiments. For the ablation study, we report composed video retrieval results on the WebVid-CoVR [43].

**Effect of two-stage training strategy.** We examine the effect of training strategy, as presented in Tab. 5 (a). With the frozen BLIP-2, training the weight predictor alone ("Stage 2 only") notably improves the performance over averaging ("Average"), demonstrating the effectiveness of our sample-wise interpolation approach even without encoder fine-tuning, while exhibiting lower performance overall compared to our full two-stage approach; it highlights the necessity of stage 1 for learning accurate cross-modal alignment. On the other hand, end-to-end training underperforms the proposed two-stage strategy despite jointly optimizing the encoders and the predictor. This result supports our design choice,

**Table 5:** Ablation studies on key components of our method.

| **Methods** | R@1 | R@5 | R@10 | R@50 |
|---|---|---|---|---|
| **(a)** ***Training Strategy*** | | | | |
| Average (BLIP-2 frozen) | 45.66 | 71.71 | 81.30 | 94.80 |
| Stage 2 only (BLIP-2 frozen) | 48.08 | 73.40 | 83.22 | 95.74 |
| End-to-end training | 57.36 | 82.71 | 89.20 | 97.77 |
| Stage 1 only | 60.50 | 83.90 | 91.09 | 98.40 |
| **(b)** ***Effect of*** $C_i$ ***in Eq.*** **(5)** | | | | |
| w/o $C_i$ (Train & Test) | 60.33 | 84.66 | 90.92 | 98.36 |
| w/o Memory $\mathcal{M}$ (Train only) | 60.62 | 84.65 | 91.09 | 98.40 |
| **(c)** ***Oracle (Upper bound)*** | | | | |
| Stage 1 + Optimal $\alpha^*$ | 71.95 | 89.95 | 94.41 | 99.10 |
| SRAIN (Ours) | 61.07 | 84.90 | 91.20 | 98.40 |

as end-to-end training leads to unstable training since the ground-truth interpolation weight changes as the encoder updates, degrading the consistency of the supervision. Meanwhile, we also include a "Stage 1 only" baseline, where the Q-former is fine-tuned with estimated optimal weights ($\alpha^*$) and a fixed weight ($\alpha = 0.5$) is used for evaluation without the weight predictor, since $\alpha = 0.5$ is empirically identified as the globally optimal fixed interpolation weight. Note that this comparison favors "Stage 1 only" because it uses the per-dataset globally optimal weight, whereas SRAIN predicts weights autonomously at test time. Since the best global $\alpha$ is highly dataset-dependent [14] and performance is highly sensitive to the choice of $\alpha$, the gains from our weight predictor are substantial. Moreover, the "Stage 2 only" variant provides a clear boost over the frozen baseline, underscoring SRAIN's value for zero-shot or data-constrained applications. The full ablation across all benchmarks and both BLIP-2 and BLIP backbones is provided in the supplement, confirming that these improvements are consistent.

**Effect of conditioned weight prediction mechanism.** We further evaluate the impact of the conditioned prediction mechanism in Tab. 5 (b), showing that using $C_i$ during training improves performance over not using it at all. Introducing the memory bank at test time to maintain train–test consistency further improves the performance.

**Upper bound analysis.** After stage 1, we report an oracle performance by assigning the optimal interpolation weights computed via the rank-aware weight estimation procedure in Sec. 3.2 across the entire test set (Tab. 5(c)). It achieves remarkable performance, suggesting that sample-aware weighting can drive significant performance gains, and that further improvement is possible through better prediction of instance-specific interpolation weights. The gap to the upper bound is mainly attributed to the predictor's difficulty in approximating extreme values in the tails of the optimal weight distribution. While the mean prediction error is near zero, indicating no systematic bias, the nontrivial standard deviation reveals that errors concentrate in these extreme cases. A detailed analysis of the oracle weight distribution over the test set is included in the supplement.

**Table 6:** Ablation studies on the memory bank for zero-shot CoIR.

| | CIRCO | FIQ-D | FIQ-S | FIQ-T | FIQ-A |
|---|---|---|---|---|---|
| **Methods** | mAP@5 | R@10 | R@10 | R@10 | R@10 |
| w/o Memory | 23.79 | 21.49 | 30.98 | 29.63 | 27.37 |
| Random Vector | 23.72 | 21.17 | 30.96 | 29.57 | 27.23 |
| SRAIN (Ours) | 26.16 | 23.22 | 33.14 | 31.91 | 29.42 |

**Robustness of memory bank under domain shift.** To verify that the memory bank provides useful conditioning even under domain shift in zero-shot CoIR, we compare SRAIN against a variant without the memory bank and one using random vectors as conditioning tokens in Tab. 6. On CIRCO and FashionIQ, SRAIN consistently outperformed the w/o Memory variant and the random vector variant. These results suggest that the memory bank provides meaningful context even when the target domain differs from the training domain, consistent with prior findings that memory-based prototypes [31], pairwise affinity [54], and relative similarity [5] remain reliable under distribution shift.

**Other backbone.** We also validated SRAIN with the BLIP [20] backbone on WebVid-CoVR. As shown in Tab. 7, SRAIN achieved 56.91% R@1, outperforming CoVR [43] with the same backbone. The component-wise ablation with BLIP, provided in the supplement, shows consistent improvements from each proposed component, confirming that our framework generalizes across different vision-language backbones.

**Table 7:** Performance on WebVid-CoVR.

| **Methods** | R@1 | R@5 | R@10 |
|---|---|---|---|
| CoVR | 55.95 | 81.22 | 89.05 |
| Ours (BLIP) | 56.91 | 81.53 | 88.90 |

## 5 Conclusion

We presented SRAIN, a novel framework for efficient multimodal fusion in composed image and video retrieval. To this end, we proposed a sample-wise rank-aware interpolation approach that predicts instance-specific weights for fusing reference and modification embeddings. SRAIN introduces a rank-aware weight estimation strategy based on ranking behavior within mini-batches and a conditioned weight prediction mechanism that references hard negatives to ensure train-test consistency. The two-stage training strategy prevents unstable optimization by separating encoder fine-tuning from weight predictor training. The method achieved state-of-the-art performance on WebVid-CoVR and competitive results on CIRCO and FashionIQ, while significantly reducing query-time latency. Our work demonstrates that lightweight, sample-wise linear interpolation can achieve superior retrieval performance without relying on computationally expensive architectures.

**Acknowledgments.** This work was supported by AI Center, Samsung Electronics Co., Ltd., and the IITP grants (RS-2022-II220290, RS-2022-II220926, RS-2024-00509258, RS-2024-00469482, RS-2026-25518317, RS-2019-II191906) funded by Ministry of Science and ICT, Korea.

## References


1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: A clip-hitchhiker's guide to long video retrieval. arXiv preprint arXiv:2205.08508 (2022)
3. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-shot composed image retrieval with textual inversion. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2023)
4. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Effective conditioned and composed image retrieval combining clip-based features. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
5. Brown, D., Shapiro, M.R., Bittner, A., Warley, J., Kvinge, H.: Wild comparisons: A study of how representation similarity changes when input data is drawn from a shifted distribution. In: ICLR 2024 Workshop on Representational Alignment (2024), https://openreview.net/forum?id=xapkeSwqf0
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Proc. Neural Information Processing Systems (NeurIPS) (2020)
7. Dodds, E., Culpepper, J., Herdade, S., Zhang, Y., Boakye, K.: Modality-agnostic attention fusion for visual search with text feedback. arXiv preprint arXiv:2007.00145 (2020)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. International Conference on Learning Representations (ICLR) (2021)
9. Goenka, S., Zheng, Z., Jaiswal, A., Chada, R., Wu, Y., Hedau, V., Natarajan, P.: Fashionvlp: Vision language transformer for fashion retrieval with feedback. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
10. Gu, G., Chun, S., Kim, W., Jun, H., Kang, Y., Yun, S.: Compodiff: Versatile composed image retrieval with latent diffusion. Transactions on Machine Learning Research (TMLR) (2024)
11. Gu, G., Chun, S., Kim, W., Kang, Y., Yun, S.: Language-only training of zero-shot composed image retrieval. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
12. Hummel, T., Karthik, S., Georgescu, M.I., Akata, Z.: Egocvr: An egocentric benchmark for fine-grained composed video retrieval. In: Proc. European Conference on Computer Vision (ECCV). Springer (2024)
13. Huynh, C., Yang, J., Tawari, A., Shah, M., Tran, S., Hamid, R., Chilimbi, T., Shrivastava, A.: Collm: A large language model for composed image retrieval. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3994–4004 (2025). https://doi.org/10.1109/CVPR52734.2025.00378
14. Jang, Y.K., Huynh, D., Shah, A., Chen, W.K., Lim, S.N.: Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. In: Proc. European Conference on Computer Vision (ECCV). Springer (2024)
15. Karthik, S., Roth, K., Mancini, M., Akata, Z.: Vision-by-language for training-free compositional image retrieval. In: Proc. International Conference on Learning Representations (ICLR) (2024)

16. Kim, J., Yu, Y., Kim, H., Kim, G.: Dual compositional learning in interactive image retrieval. In: Proc. AAAI Conference on Artificial Intelligence (AAAI) (2021)
17. Lee, S., Kim, D., Han, B.: Cosmo: Content-style modulation for image retrieval with text feedback. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
18. Levy, M., Ben-Ari, R., Darshan, N., Lischinski, D.: Data roaming and quality assessment for composed image retrieval. In: Proc. AAAI Conference on Artificial Intelligence (AAAI) (2024)
19. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proc. International Conference on Machine Learning (ICML). PMLR (2023)
20. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proc. International Conference on Machine Learning (ICML). PMLR (2022)
21. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP) (2024)
22. Lin, H., Wen, H., Song, X., Liu, M., Hu, Y., Nie, L.: Fine-grained textual inversion network for zero-shot composed image retrieval. In: Proc. International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) (2024)
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Proc. Neural Information Processing Systems (NeurIPS) (2023)
24. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2021)
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proc. International Conference on Learning Representations (ICLR) (2019)
26. Psomas, B., Retsinas, G., Efthymiadis, N., Filntisis, P., Avrithis, Y., Maragos, P., Chum, O., Tolias, G.: Instance-level composed image retrieval. In: Proc. Neural Information Processing Systems (NeurIPS) (2025)
27. Radenovic, F., Dubey, A., Kadian, A., Mihaylov, T., Vandenhende, S., Patel, Y., Wen, Y., Ramanathan, V., Mahajan, D.: Filtering, distillation, and hard negatives for vision-language pre-training. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
28. 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: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
29. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Proc. Neural Information Processing Systems (NeurIPS) (2022)
30. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
31. Sharma, A., Kalluri, T., Chandraker, M.: Instance level affinity-based transfer for unsupervised domain adaptation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5357–5367 (2021). https://doi.org/10.1109/CVPR46437.2021.00532

32. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proc. Annual Meeting of the Association for Computational Linguistics (ACL) (2018)
33. Shoemake, K.: Animating rotation with quaternion curves. In: Proc. Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH) (1985)
34. Sun, S., Ye, F., Gong, S.: Training-free zero-shot composed image retrieval with local concept reranking. arXiv preprint arXiv:2312.08924 (2023)
35. Sun, S., Chen, Y.C., Li, L., Wang, S., Fang, Y., Liu, J.: Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In: Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (2021)
36. Sun, Z., Yang, G., Lu, Z., Jiang, H., Zhu, G., Cao, Z.: Image retrieval with composed query by multi-scale multi-modal fusion. In: IEEE international conference on acoustics, speech and signal processing (ICASSP) (2024)
37. Tang, Y., Yu, J., Gai, K., Zhuang, J., Xiong, G., Hu, Y., Wu, Q.: Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In: Proc. AAAI Conference on Artificial Intelligence (AAAI) (2024)
38. Tang, Y., Zhang, J., Qin, X., Yu, J., Gou, G., Xiong, G., Lin, Q., Rajmohan, S., Zhang, D., Wu, Q.: Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
39. Thawakar, O., Demidov, D., Thawkar, R., Anwer, R.M., Shah, M., Khan, F.S., Khan, S.: Beyond simple edits: Composed video retrieval with dense modifications. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2025)
40. Thawakar, O., Naseer, M., Anwer, R.M., Khan, S., Felsberg, M., Shah, M., Khan, F.S.: Composed video retrieval via enriched context and discriminative embeddings. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Proc. Neural Information Processing Systems (NeurIPS) (2017)
42. Ventura, L., Yang, A., Schmid, C., Varol, G.: Covr-2: Automatic data construction for composed video retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2024)
43. Ventura, L., Yang, A., Schmid, C., Varol, G.: Covr: Learning composed video retrieval from web video captions. In: Proc. AAAI Conference on Artificial Intelligence (AAAI) (2024)
44. 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: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
45. Wan, Y., Wang, W., Zou, G., Zhang, B.: Cross-modal feature alignment and fusion for composed image retrieval. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2024)
46. Wen, H., Zhang, X., Song, X., Wei, Y., Nie, L.: Target-guided composed image retrieval. In: Proc. ACM International Conference on Multimedia (ACM Multimedia) (2023)
47. 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: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

48. Xing, E., Kolouju, P., Pless, R., Stylianou, A., Jacobs, N.: Context-cir: Learning from concepts in text for composed image retrieval. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19638–19648 (2025). https://doi.org/10.1109/CVPR52734.2025.01829
49. Xu, Y., Bin, Y., Wei, J., Yang, Y., Wang, G., Shen, H.T.: Multi-modal transformer with global-local alignment for composed query image retrieval. IEEE Transactions on Multimedia (2023)
50. Yang, Z., Xue, D., Qian, S., Dong, W., Xu, C.: Ldre: Llm-based divergent reasoning and ensemble for zero-shot composed image retrieval. In: Proc. International ACM SIGIR conference on research and development in information retrieval (SIGIR) (2024)
51. Zhang, F.: Stable diffusion prompts 2.47m. https://huggingface.co/datasets/FredZhang7/stable-diffusion-prompts-2.47M (2023)
52. Zhang, G., Wei, S., Pang, H., Zhao, Y.: Heterogeneous feature fusion and cross-modal alignment for composed image retrieval. In: Proc. ACM International Conference on Multimedia (ACM Multimedia) (2021)
53. Zhang, K., Luan, Y., Hu, H., Lee, K., Qiao, S., Chen, W., Su, Y., Chang, M.W.: Magiclens: Self-supervised image retrieval with open-ended instructions. In: Proc. International Conference on Machine Learning (ICML) (2024)
54. Zhang, Z., Chen, W., Cheng, H., Li, Z., Li, S., Lin, L., Li, G.: Divide and contrast: Source-free domain adaptation via adaptive contrastive learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Proc. Neural Information Processing Systems (NeurIPS). vol. 35, pp. 5137–5149. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/215aeb07b5996c969c0123c3c6ee8f54-Paper-Conference.pdf
55. Zhao, S., Xu, H.: Neucore: neural concept reasoning for composed image retrieval. In: Proc. UniReps: the First Workshop on Unifying Representations in Neural Models. PMLR (2024)
56. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)