Beyond Query Bias: Candidate-Aware Iterative Refinement for Zero-Shot Composed Image Retrieval

Nan Sun[1,2*], **Jing Tang**[2†], **Lei Sun**[2], **Rui Chen**[2], **Yuxing Lu**[3]
Xiangxiang Chu[2], **Hefei Ling**[1§], **Yujun Cai**[4]
[1]Huazhong University of Science and Technology, [2]AMAP, Alibaba Group,
[3]Peking University, [4]University of Queensland
{sunnan, lhefei}@hust.edu.cn, guangyu.tj@alibaba-inc.com
† Project Leader. § Corresponding Author.

Abstract

Zero-Shot Composed Image Retrieval (ZS-CIR) retrieves target images using a reference image and modification text without task-specific training. Existing methods typically rely on MLLMs to generate query vectors with pre-trained models like CLIP. However, those constructed queries suffer from inherent *cognitive bias* due to unknown candidate distribution. We propose CoRR, a training-free framework that reframes ZS-CIR as a self-correcting process through bias-aware query refinement. CoRR uses retrieved results as feedback to perceive the candidate distribution. With carefully designed CoT prompting, the MLLM inspects the retrieved candidates to identify intent misalignments in the query and then corrects them via Historical Query Fusion. We also introduce Retrieval-Driven Caption Optimization to provide context-aligned examples, reducing phrasing and style mismatches. Experiments on public benchmarks show that CoRR significantly outperforms other SOTA methods.

1 Introduction

Composed Image Retrieval (CIR) aims to retrieve a target image that preserves the relevant visual content of a reference image while incorporating the semantic modifications described in a textual query (Vo et al., 2019; Delmas et al., 2022; Huynh et al., 2025; Xing et al., 2025). This fine-grained retrieval task has significant practical value in numerous real-world applications such as web search and e-commerce (Chen et al., 2020; Saito et al., 2023; Bai et al., 2024; Tang et al., 2025b), offering a more intuitive and flexible way for users to interact with visual content.

Zero-Shot Composed Image Retrieval (ZS-CIR) (Saito et al., 2023; Karthik et al., 2024; Yang et al., 2024; Tang et al., 2025a,b; Luo et al., 2025) has emerged as a promising paradigm, due to its cost-effectiveness. A common practice is to employ Multimodal Large Language Models (MLLMs) or Large Language Models (LLMs) to generate composed queries that integrate information from both the reference image and modification text, and then retrieve results with pre-trained multimodal embedding models like CLIP (Radford et al., 2021).

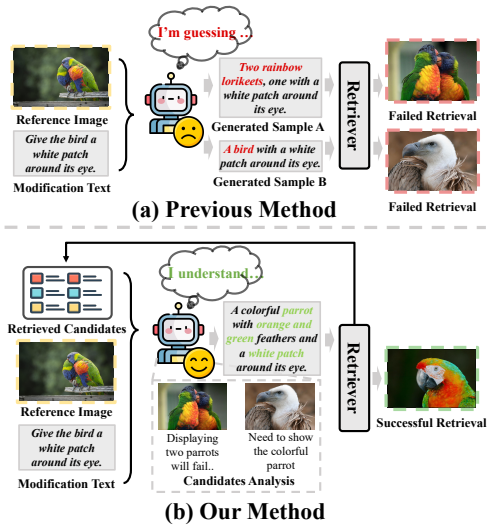

Figure 1: An overview of the previous methods and our method. (a) relies solely on guessing the user's intent based on the query itself, while (b) analyzes and improves the query through feedback obtained from retrieval results.

The effectiveness of these methods heavily relies on the accuracy of the composed queries, which require precise semantic editing of the reference image based on the provided text. However, this task is inherently ambiguous: the semantics of the target image are not strictly defined by the reference image and the modification text (Bordogna and Pasi, 1993; Chen and Wang, 2002; Yang et al., 2024), so any composed query inevitably reflects a biased guess about the unknown candidate space. Recent work seeks to alleviate this issue by enriching the query itself, for example using MLLM-based reasoning or generating diverse captions from multiple perspectives (Tang et al., 2025b; Yang et al., 2024; Sun et al., 2025), with the intuition that a

*Work done during the internship at AMAP, Alibaba.

Findings of the Association for Computational Linguistics: ACL 2026, pages 22318–22329

carefully constructed set of queries can collectively compensate for this bias. However, the candidate distribution remains unseen, even semantically refined or diversified queries cannot guarantee a good match to the actual candidates. As illustrated in Figure 1(a), given the instruction "give the bird a white patch around its eye", the model can generate either a more specific caption ("two rainbow lorikeets, one with a white patch around its eye") or a more generic one ("a bird with a white patch around its eye"). Both queries are semantically reasonable, yet neither matches any image in the candidate set exactly, leading to failed retrieval in both cases. We term this mismatch between seemingly correct queries and the (unseen) candidate distribution as ***cognitive bias***. During the query generation stage, the model has no access to the actual candidates and therefore cannot avoid such biased guesses.

In this paper, we propose CoRR (**C**hain **o**f **R**eflective Composed Image **R**etrieval), a training-free framework designed to correct the cognitive bias caused by blind query generation. Instead of relying only on the reference image and text, CoRR incorporates retrieval feedback, giving the refinement process partial awareness of the candidate distribution and enabling it to adjust biased constraints in the query. As shown in Figure 1(b), an MLLM reflects on the retrieved items to detect which constraints in the current query are unsatisfiable, overly restrictive, or missing, and accordingly relaxes or corrects them. To prevent unstable updates from directly adopting the newly generated query, CoRR updates the query at the representation level via Historical Query Fusion, which smoothly blends the new query with the historical query vector on the embedding space. In addition, CoRR further mitigates bias from language-embedding mismatches through Retrieval-Driven Caption Optimization, which offline constructs embedding-aligned captions of candidate images and feeds them into the MLLM, allowing the model to learn the embedding model's preferences and generate more retrievable queries. Our main contributions are as follows:

- We reinterpret Zero-Shot Composed Image Retrieval through the lens of *cognitive bias*. Based on this perspective, we propose CoRR, a training-free, feedback-driven framework that makes ZS-CIR a dynamic and self-correcting process.
- We design a bias-reduction pipeline that uses MLLM-guided self-reflection over multimodal inputs and retrieval feedback, together with historical query fusion and retrieval-optimized captions, to stably refine queries and better align them with the embedding space.
- Extensive experiments on multiple ZS-CIR benchmarks demonstrate that CoRR consistently improves strong baselines and achieves state-of-the-art performance with only a small number of additional refinement rounds.

2 Related Work

Composed Image Retrieval. Composed Image Retrieval (CIR) retrieves target images using a reference image and textual modifications (Wu et al., 2021; Han et al., 2017; Vo et al., 2019). Classical approaches train on large-scale, manually annotated triplets (Liu et al., 2021; Baldrati et al., 2022; Chen et al., 2020; Chen and Bazzani, 2020; Lee et al., 2021; Anwaar et al., 2021), while zero-shot methods avoid reliance on annotated triplets by either using pseudo-tokens (Saito et al., 2023; Baldrati et al., 2023, 2022; Gu et al., 2024; Tang et al., 2025a; bai et al., 2024; Li et al., 2025, 2026) or leveraging MLLMs to generate captions for text-to-image retrieval (Karthik et al., 2024; Yang et al., 2024; Tang et al., 2025b), enabling retrieval using models like CLIP (Radford et al., 2021). However, these methods refine queries without verifying alignment with actual candidates, leading to cognitive bias and retrieval failures.

Embedding Models and Multimodal Large Language Models. Embedding models, particularly pioneering architectures such as CLIP (Radford et al., 2021) and BLIP (Li et al., 2022), have successfully established a unified semantic space by mapping images and text through training on massive image-text datasets. This breakthrough has enabled diverse applications across image generation (Kim et al., 2022; Rombach et al., 2022), classification (Zhou et al., 2022; Qu et al., 2025), and cross-modal retrieval (Bogolin et al., 2022; Wang et al., 2025). Recently, the field has evolved from simple feature alignment toward deep integration of visual capabilities with Large Language Models (LLMs), resulting in more powerful Multimodal Large Language Models (MLLMs) (Liu et al., 2024; Zhu et al., 2025; Bai et al., 2025a,b; Shahriar et al., 2024; Chen et al., 2025). These models achieve deep visual understanding and complex reasoning capabilities through instruction tuning, enabling tasks such as visual question answering and image description generation.

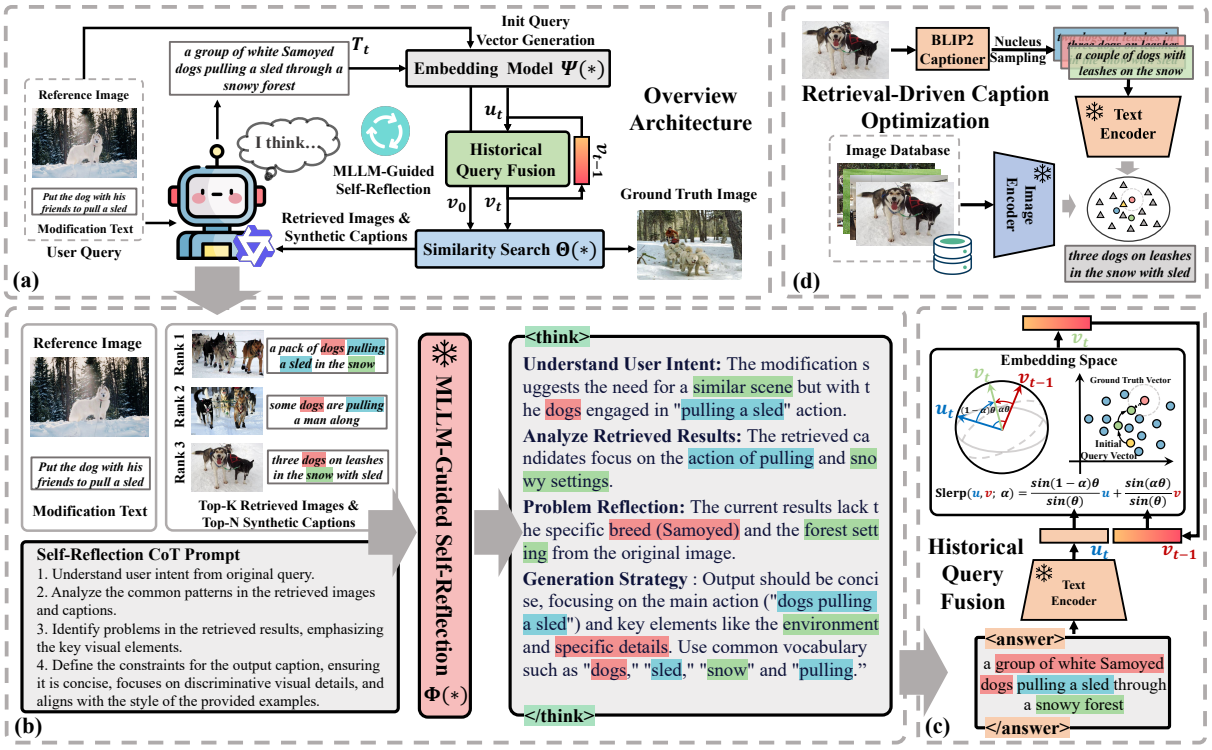

Figure 2: An architecture of our framework. (a) illustrates the simplified pipeline of our proposed method. (b) demonstrates the reasoning and self-reflection process facilitated by the MLLM, where different colors highlight distinct key visual elements. (c) visualizes the Historical Query Fusion process, while (d) showcases our Retrieval-Driven Caption Optimization strategy.

3 Methodology

Before formally presenting our method, we provide a detailed definition of the ZS-CIR task. Given a reference image I_r and a modification text T, Zero-Shot Composed Image Retrieval (ZS-CIR) aims to identify an image from a database $\mathcal{D} = \{I_1, I_2, ..., I_n\}$ that best represents I_r after applying the semantic changes described by T.

ZS-CIR methods employ a multimodal embedding model $\Psi(*)$ that projects images and text into a shared embedding space. The composed query vector is $v_q = \Psi(I_r, T) \in \mathbb{R}^d$, and candidate image vectors are $\Psi_I(I_i) \in \mathbb{R}^d$, where d is the embedding dimension and Ψ_I denotes the image encoder. The retrieval goal is to find the image I^* that maximizes similarity:

$$I^* = \Theta(\mathcal{D}, I_r, T) = \arg\max_{I_i \in \mathcal{D}} \frac{\Psi_I(I_i)^\top \Psi(I_r, T)}{\|\Psi_I(I_i)\| \cdot \|\Psi(I_r, T)\|} \quad (1)$$

3.1 Overview

The pipeline of our proposed framework is illustrated in Figure 2 (a). CoRR follows an iterative retrieval–reflection–refinement process. It begins by generating an initial query vector $v_0 = \Psi(I_r, T)$ from reference image I_r and modification text T using the embedding model to retrieve a set of candidate images $\mathcal{I}_0$. Subsequently, in the t-th round, an MLLM reasoner Φ analyzes the Top-K (default 5) retrieved images $\mathcal{I}_{t-1}$ from the previous retrieval and the generated captions $\mathcal{C}_{t-1}$ of the Top-N (default 10) images, along with the original inputs (I_r, T), to generate a refined text query T'_t. This refined text query is then used to compute an updated query vector v_t via Historical Query Fusion for next round.

Our approach primarily consists of three core modules: (1) MLLM-Guided Self-Reflection, (2) Historical Query Fusion based on Slerp, and (3) Retrieval-Driven Caption Optimization. Algorithm 1 illustrates the overall pipeline. We will elaborate on each in detail.

3.2 MLLM-Guided Self-Reflection

In each retrieval loop, our MLLM Reasoner executes a reasoning process to analyze the retrieved candidates from the previous round, reflect on the results, and generate a refined query for the next iteration. Based on the examples given in the Figure 2 (b), we briefly introduce the process.

Algorithm 1 The Algorithm of CoRR

Require: Reference image I_r, modification text T, database $\mathcal{D}$; embedding model Ψ_T, image encoder Ψ_I; MLLM reasoner Φ; maximum iterations $T_{\max}$; pre-computed optimized captions $\{C^*(I)\}_{I \in \mathcal{D}}$

Ensure: Retrieved target image I^*

Initial query:
$v_0 \leftarrow \Psi(I_r, T)$
$\mathcal{I}_0 \leftarrow \Theta(\mathcal{D}, v_0)$ // Initial retrieval with embedding model
for $t = 1$ **to** $T_{\max}$ **do**
  Context construction:
  $\mathcal{I}_{t-1}^K \leftarrow$Top-$K$ images from $\mathcal{I}_{t-1}$
  $\mathcal{C}_{t-1}^N \leftarrow$Top-$N$ captions from $\{C^*(I) \mid I \in \mathcal{I}_{t-1}\}$ // Look up offline optimized captions
  MLLM-guided reflection:
  $T_t \leftarrow \Phi\big(I_r, T, \mathcal{I}_{t-1}^K, \mathcal{C}_{t-1}^N\big)$ // Reason over retrieved feedback to refine the query
  Query update:
  $u_t \leftarrow \Psi_T(T_t)$ // Encode refined textual query
  $v_t \leftarrow \mathrm{Slerp}(u_t, v_{t-1}; \alpha)$ // Historical query fusion
  Retrieval:
  $\mathcal{I}_t \leftarrow \Theta(\mathcal{D}, v_t)$ // Retrieve images
  $\mathcal{I}_{last} \leftarrow \mathcal{I}_t$
  Stopping criterion (optional):
  if stop_criterion_met **then**
    break
  end if
end for
$I^* \leftarrow$ top-1 image in $\mathcal{I}_{last}$
return I^*

Understand User Intent. First, the MLLM will analyze the reference image and modification text to clarify user intent. As shown in the figure, the MLLM finds that user intent is to retrieve an image of dogs pulling a sled in an environment similar to the original scene.

Analyze Retrieved Results. Then, the MLLM analyzes the retrieved images and synthetic captions from the previous round. This step aims to retain key visual elements that match user intent. As shown in the figure, the MLLM examines the caption of "pulling" and "snow settings".

Problem Reflection. Building upon the analysis of retrieved results, the MLLM synthesizes its findings to identify the problems with the current results relative to the user intent. As shown in the figure, the MLLM finds that the current results lack the specific breed and forest environment.

Generation Strategy. The MLLM reasoner then formulates the next query by keeping the key visuals, correcting the biased parts identified in reflection, and matching the embedding model's style with concise, common vocabulary.

3.3 Historical Query Fusion Based on Slerp

As shown in Figure 2 (c), in each iteration of our framework, the MLLM generates a refined target description T_t, which is then encoded into a new query vector $u_t = \Psi_T(T_t)$. However, directly replacing the previous query vector with this new one can lead to instability. The MLLM's reflection might cause an over-correction or introduce noise, leading to "query drift" (Mitra et al., 1998) where valuable information from the previous state is lost.

To ensure a smooth and stable evolution of the query, we fuse the historical query v_{t-1} from the previous iteration with the new query vector u_t as shown in Figure 2 (c). We employ Spherical Linear Interpolation (Slerp) (Shoemake, 1985) for this task, as it is naturally suited for operating on the hypersphere of unit-normalized embedding models. Slerp ensures a smooth transition between the two vectors, providing a robust update mechanism. The final query vector for the current iteration v_t is computed as follows:

$$v_t \longleftarrow \mathrm{Slerp}(u_t, v_{t-1}; \alpha) \tag{2}$$

The Slerp function is defined as:

$$\mathrm{Slerp}(u, v; \alpha) = \frac{\sin(1-\alpha)\theta}{\sin(\theta)} u + \frac{\sin(\alpha\theta)}{sin(\theta)} v \tag{3}$$

where $\theta = \arccos(u \cdot v)$ is the angle between the two vectors, and $\alpha \in [0, 1]$ (default 0.8) is a fixed hyperparameter that controls the interpolation weight, balancing the influence of the historical query and the new evidence. See Section 4.3 for comparisons of α weight.

This Slerp-based fusion mitigates destructive drift by maintaining momentum from the previous query while still integrating the new insights from the MLLM's reflection. Figure 2 (c) shows the desired smooth trajectory of query vectors evolving in the "Embedding Space" across iterations. The resulting updated vector v_t, is then used to retrieve the image gallery $\mathcal{D}$ with $\Theta(*)$ to produce the candidate set for the next iteration.

3.4 Retrieval-Driven Caption Optimization

Many studies have shown that dense embedding models are highly sensitive to their input (Rafiei Asl et al., 2024; Magomere et al., 2025). To align the MLLM outputs with the preferences of the embedding model, prior methods use fixed, manually-curated caption examples to guide MLLM generation (Karthik et al., 2024; Tang et al., 2025b), but these static examples are disconnected from the current query. Our key insight is that effective examples should be query-relevant. By providing these retrieval-focused captions to the MLLM, we guide it to learn the embedding model's preferences, including semantic granularity and sentence style. Additionally, providing extra captions allows the MLLM to learn the discriminative visual elements within the images, which helps it better analyze the retrieved results. Because the mining depends exclusively on candidate images and not on the query, it is precomputed **offline**, adding no overhead to inference latency.

As shown in Figure 2 (d), our Retrieval-Driven Caption Optimization strategy generates M (default 30) high-quality captions $\{C_i\}_{i=1}^{M}$ for each of the images of database using BLIP2 (Li et al., 2023a). To select the most effective one caption, we employ a retrieval-based validation approach that ranks captions using a two-stage sorting strategy: first by the rank of the source image I when using caption C_i as a query, then by the similarity score between the caption and the source image for captions with identical ranks. Formally, for each caption C_i, we compute its ranking:

$$(r_i, s_i) < (r_j, s_j) \iff (r_i < r_j) \text{ or } (r_i = r_j \text{ and } s_i > s_j) \quad (4)$$

where $r_i = \text{rank}(I, C_i)$ denotes the rank of image I when caption C_i is used as a query, and $s_i = \text{sim}(I, C_i)$ represents the similarity score between caption C_i and image I. As shown in Algorithm 2, this offline procedure takes each candidate image, generates M textual captions, and evaluates each by self-retrieval. Captions are ordered by (r_i, s_i), and the top caption $C^*(I)$ is kept as the embedding-aligned example for that image. Because it depends only on the candidate set, this mining is fully precomputed and adds no inference latency. We provide corresponding ablation studies in Section 4.3.

Algorithm 2 Offline Retrieval-Driven Caption Optimization

Require: Image database $\mathcal{D}$, captioner Γ, image encoder Ψ_I, text encoder Ψ_T, captions M
Ensure: Optimized caption $C^*(I)$ for each image $I \in \mathcal{D}$

for each image $I \in \mathcal{D}$ **do**
  Generate M candidate captions $\{C_i\}_{i=1}^{M} \leftarrow \Gamma(I)$
  for $i = 1$ to M **do**
    $u_i \leftarrow \Psi_T(C_i)$ // Encode caption
    $r_i \leftarrow$ similarity_search(u_i) // rank of image I when using u_i as query for retrieval
    $s_i \leftarrow \text{sim}(\Psi_I(I), u_i)$ // compute similarity between image I and C_i
  end for
  Sort $\{C_i\}$ by the key $(r_i, -s_i)$ // lower rank first; if tied, higher similarity first
  $C^*(I) \leftarrow$ first caption after sorting
end for
return $\{C^*(I)\}_{I \in \mathcal{D}}$

4 Experiments

Implementation Details. Our framework is implemented in PyTorch (Paszke et al., 2019) and all experiments are conducted on a single NVIDIA A6000 GPU with 48GB. For image retrieval, we use a FAISS (Douze et al., 2024) Flat index to perform an exact search, using inner product as the similarity metric. We adopt different existing ZS-CIR models as the embedding models. In init round, retrieval is performed solely by the embedding model. Subsequently, the MLLM executes two additional reflection and refinement rounds to iteratively update the query. Our primary MLLM is Qwen-VL-Max [1]. We leverage BLIP-2 (Li et al., 2023b) with a OPT-6.7b (Zhang et al., 2022) language model as the image captioner to generate optimized caption examples.

Datasets and Evaluation Metrics. We evaluate on three benchmarks: **CIRR** (Liu et al., 2021), **CIRCO** (Baldrati et al., 2023), **FashionIQ** (Wu et al., 2021) and **GeneCIS** (Vaze et al., 2023). CIRCO and CIRR evaluate object modification tasks using mAP@k and Recall@k on hidden test sets. CIRR also reports subset recall (Recall$_{sub}$@k). FashionIQ evaluates attribute adjustment using Recall@k on the validation set for Shirt, Dress, and

[1] https://github.com/QwenLM/Qwen-VL

Toptee categories. GeneCIS evaluates object and attribute composition tasks using Recall@k.

Baseline and Backbone. To evaluate the effectiveness of our module, we conduct comprehensive comparisons against several prior state-of-the-art methods built on pre-trained CLIP (Radford et al., 2021): **SEARLE** (Baldrati et al., 2023), **Pic2Word** (Saito et al., 2023), **Slerp-TAT** (Jang et al., 2024), **LDRE** (Yang et al., 2024), **Context-I2W** (Tang et al., 2024), **CIReVL** (Karthik et al., 2024), **ImageScope** (Luo et al., 2025), **PrediCIR** (Tang et al., 2025a), **OSrCIR** (Tang et al., 2025b), and **MMRet** (Zhou et al., 2025). We group comparisons by CLIP architecture: **CLIP-ViT-B/32**, **CLIP-ViT-L/14** (MMRet-Base is only available for the CLIP-ViT-B/16 variant, our comparison with it is limited to this version). Integrating our module with these backbones (e.g., "Ours+MMRet-large") achieves substantial performance gains.

4.1 Main Results

We present our main quantitative results on the CIRCO and CIRR benchmarks in Table 1, and FashionIQ benchmark in Table 2. The results demonstrate that our proposed paradigm consistently and significantly enhances the performance of various baseline methods.

In Table 1, we evaluate our approach on the CIRCO and CIRR datasets to demonstrate its performance on tasks that require both foreground-background separation and fine-grained modifications. Specifically, on CIRCO, using the CLIP-ViT-L backbone with Slerp ("Ours+Slerp"), we enhances the mAP@5 score from 16.40% to 26.08%, representing a relative improvement of over 59%. Notably, this is achieved in a training-free manner, significantly outperforming Slerp+TAT (Jang et al., 2024), the original training-based method proposed in its paper. Moreover, when applied to a strong baseline such as MMRet-Large, our method still achieves a consistent uplift, improving mAP@5 from 40.2% to 42.7%. Furthermore, on CIRR, when applied to MMRet-Base, our method increases Recall@1 from 35.97% to 41.58% (+5.61). Similarly, with MMRet-Large, Recall@1 rises from 37.95% to 43.21% (+5.26).

On the domain-specific FashionIQ benchmark, which demands precise localization of specific fashion attributes, our method demonstrates strong effectiveness. As illustrated in Table 2, our approach consistently improves the performance of various baseline models, achieving an average increase of 2-5 points in R@10 and R@50 across all three categories. These results highlight our method's ability to capture fine-grained, domain-specific attributes with notable accuracy.

Moreover, we present results on the GeneCIS dataset in Table 3. Our method achieves state-of-the-art performance with an average R@1 score of 19.53, demonstrating the effectiveness of our iterative retrieval approach. Notably, we outperform SOTA methods such as OSrCIR by 1.63 points, highlighting the significant improvement brought by our iterative reflective CoT mechanism and Retrieval-Driven Caption Optimization strategy.

4.2 Qualitative Analysis

To more intuitively demonstrate the effectiveness of our method, we provide some successful retrieval cases in Figure 3. These cases cover a variety of complex modification texts, including: (a) For **action modification**, it relaxes overly specific constraints ("slim, long legs") to retrieve the target by focusing on general attributes ("tan dog"); (b) For **attribute and number modification**, it integrates gradual refinements ("remove one dog, add grass, make one dog black") to incrementally improve retrieval accuracy; (c) For **quantity and scene change**, it resolves semantic conflicts (e.g., "savanna" vs. "hillside") by generalizing constraints ("savanna" to "landscape"); (d) For **holistic replacement**, it handles substantial semantic gaps ("blue pouch" to "yellow shoe") by isolating the query's core intent, enabling successful retrieval. These cases demonstrate that our iterative approach effectively handles complex modifications, significantly improving retrieval through self-correcting.

4.3 Ablation Study

To dissect the contribution of each component and design choice within our framework, we conduct a series of ablation studies.

Impact of Different Components and Prompt Strategies. As shown in the Table 4, we quantify the impact of different components on performance. The "+ reflection" variant performs MLLM-guided reflection on retrieved images and directly uses the newly generated query to replace the previous one, which results in performance degradation due to query drift. In contrast, incorporating historical query fusion ("+query fusion") stabilizes the query update, leading to significant improvement. This highlights the critical role of historical query fusion in preventing query drift and maintaining retrieval

Table 1: **Comparison on CIRCO and CIRR Test Data.** Results are grouped by architecture and sorted by publication year. Training-free methods are marked with "✓". Methods with "†" are our implementations. Green highlighting indicates the best performance, while blue highlighting indicates the second-best performance.

	Architecture	Training-free	CIRCO mAP@k				CIRR Recall@k				CIRR Recall$_{sub}$@k		
			k=5	k=10	k=25	k=50	k=1	k=5	k=10	k=50	k=1	k=2	k=3
CLIP-ViT-B	SEARLE (ICCV'23)		9.35	9.94	11.13	11.84	24.00	53.42	66.82	89.78	54.89	76.60	88.19
	Slerp (ECCV'24)†	✓	6.51	7.05	8.13	8.70	18.12	49.11	63.16	87.59	61.99	80.61	90.94
	Ours+Slerp	✓	14.42	14.99	16.55	17.44	24.77	56.02	70.27	91.66	63.54	82.72	91.59
	Δ *(Ours vs Slerp)*		*(+7.91)*	*(+7.94)*	*(+8.42)*	*(+8.74)*	*(+6.65)*	*(+6.91)*	*(+7.11)*	*(+4.07)*	*(+1.55)*	*(+2.11)*	*(+0.65)*
	Slerp+TAT (ECCV'24)		9.34	10.26	11.65	12.33	28.19	55.88	68.77	88.51	61.13	80.63	90.68
	LDRE (SIGIR'24)	✓	17.96	18.32	20.21	21.11	25.69	55.13	69.04	89.90	60.53	80.65	90.70
	CIReVL (ICLR'24)	✓	14.94	15.42	17.00	17.82	23.94	52.51	66.0	86.95	60.17	80.05	90.19
	ImageScope (WWW'25)	✓	22.36	22.19	23.03	23.83	34.36	60.58	71.40	88.41	74.63	87.93	93.83
	OSrCIR (CVPR'25)	✓	18.04	19.17	20.94	21.85	25.42	54.54	68.19	-	62.31	80.86	91.13
	MMRet-Base (ACL'25)†		34.21	34.78	37.20	38.38	35.97	68.17	79.56	94.72	71.61	87.47	94.46
	Ours+MMRet-Base	✓	**37.22**	**37.94**	**40.4**	**41.55**	**41.58**	**72.31**	**82.48**	**96.12**	**74.77**	**90.1**	**95.66**
	Δ *(Ours vs MMRet-Base)*		*(+3.01)*	*(+3.16)*	*(+3.2)*	*(+3.17)*	*(+5.61)*	*(+4.14)*	*(+2.92)*	*(+1.4)*	*(+3.16)*	*(+2.63)*	*(+1.4)*
CLIP-ViT-L	Pic2Word (CVPR'23)		8.72	9.51	10.64	11.29	23.90	51.70	65.30	87.80	-	-	-
	SEARLE-XL (ICCV'23)		11.68	12.73	14.33	15.12	24.24	52.48	66.29	88.84	53.76	75.01	88.19
	Context-I2W (AAAI'24)		-	-	-	-	25.60	55.10	68.50	89.80	-	-	-
	LinCIR (CVPR'24)		12.59	13.58	15.00	15.85	25.04	53.25	66.68	-	57.11	77.37	88.89
	Slerp (ECCV'24)†	✓	16.40	18.41	20.89	21.97	19.28	48.22	62.24	85.74	58.05	78.05	88.96
	Ours+Slerp	✓	26.08	27.65	30.48	31.74	25.59	56.75	70.12	90.84	62.99	81.64	90.94
	Δ *(Ours vs Slerp)*		*(+9.68)*	*(+9.24)*	*(+9.59)*	*(+9.77)*	*(+6.31)*	*(+8.53)*	*(+7.88)*	*(+5.1)*	*(+4.94)*	*(+3.59)*	*(+1.98)*
	Slerp+TAT (ECCV'24)		18.46	19.41	21.43	22.41	30.94	59.4	70.94	89.18	64.7	82.92	92.31
	LDRE (SIGIR'24)	✓	23.35	24.03	26.44	27.5	26.53	55.57	67.54	88.50	60.43	80.31	89.90
	CIReVL (ICLR'24)	✓	18.57	19.01	20.89	21.8	24.55	52.31	64.92	86.34	59.54	79.88	89.69
	ImageScope (WWW'25)	✓	25.39	25.82	27.07	27.98	34.99	61.35	71.49	88.84	74.94	88.24	94.0
	PrediCIR (CVPR'25)		15.70	17.10	18.60	19.30	27.20	57.00	70.20	-	-	-	-
	OSrCIR (CVPR'25)	✓	23.87	25.33	27.84	28.97	29.45	57.68	69.86	-	62.12	81.92	91.1
	MMRet-Large (ACL'25)†		40.20	41.20	43.80	44.91	37.95	70.36	81.08	94.75	73.23	88.12	94.8
	Ours+MMRet-Large	✓	**42.70**	**44.09**	**46.90**	**48.04**	**43.21**	**73.83**	**83.9**	**95.78**	**76.82**	**90.31**	**96.1**
	Δ *(Ours vs MMRet-Large)*		*(+2.5)*	*(+2.89)*	*(+3.1)*	*(+3.13)*	*(+5.26)*	*(+3.47)*	*(+2.82)*	*(+1.03)*	*(+3.59)*	*(+2.19)*	*(+1.3)*

Table 2: **Comparison on FashionIQ Validation Data.** Results are grouped by architecture and sorted by publication year. Training-free methods are marked with "✓". Methods with "†" are our implementations. Green highlighting indicates the best performance, while blue highlighting indicates the second-best performance.

	Architecture	Training-free	Shirt		Dress		Toptee		Average	
			R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
CLIP-ViT-B	SEARLE (ICCV'23)		24.44	41.61	18.54	39.51	25.70	46.46	22.89	42.53
	Slerp (ECCV'24)†	✓	22.18	39.40	19.98	39.76	26.31	44.01	22.82	41.06
	Ours+Slerp	✓	26.94	45.44	22.16	42.89	29.78	50.28	26.29	46.20
	Δ *(Ours vs Slerp)*		*(+4.76)*	*(+6.04)*	*(+2.18)*	*(+3.13)*	*(+3.47)*	*(+6.27)*	*(+3.47)*	*(+5.14)*
	Slerp+TAT (ECCV'24)		23.06	41.95	19.24	42.14	26.57	47.78	22.96	43.96
	LDRE (SIGIR'24)	✓	27.38	46.27	19.97	41.84	27.07	48.78	24.81	45.63
	CIReVL (ICLR'24)	✓	28.36	47.84	25.29	46.36	31.21	53.85	28.29	49.35
	ImageScope (WWW'25)	✓	24.29	37.49	18.0	35.20	24.99	41.41	22.42	38.03
	OSrCIR (CVPR'25)	✓	31.16	51.13	**29.35**	50.37	36.51	58.71	32.34	53.40
	MMRet-Base (ACL'25)†		33.81	53.14	26.28	49.38	36.1	57.32	32.06	53.28
	Ours+MMRet-Base	✓	**36.85**	**55.74**	27.12	**50.42**	**38.19**	**58.80**	**34.05**	**54.99**
	Δ *(Ours vs MMRet-Base)*		*(+3.04)*	*(+2.6)*	*(+0.84)*	*(+1.04)*	*(+2.09)*	*(+1.48)*	*(+1.99)*	*(+1.71)*
CLIP-ViT-L	Pic2Word (CVPR'23)		26.20	43.60	20.00	40.20	27.90	47.40	24.70	43.73
	SEARLE-XL (ICCV'23)		26.89	45.58	20.48	43.13	29.32	49.97	25.56	46.23
	Context-I2W (AAAI'24)		29.70	48.60	23.10	45.30	30.60	52.90	27.80	48.90
	LinCIR (CVPR'24)		29.10	46.81	20.92	42.44	28.81	50.18	26.28	46.49
	Slerp (ECCV'24)†	✓	27.58	42.89	21.42	41.35	29.22	47.58	26.95	44.62
	Ours+Slerp	✓	32.24	49.12	23.03	45.56	33.76	54.61	29.68	49.76
	Δ *(Ours vs Slerp)*		*(+4.66)*	*(+6.23)*	*(+1.61)*	*(+4.21)*	*(+4.54)*	*(+7.03)*	*(+2.73)*	*(+5.14)*
	Slerp+TAT (ECCV'24)		29.64	46.47	23.35	45.12	31.97	51.20	28.32	47.60
	LDRE (SIGIR'24)	✓	31.04	51.22	22.93	46.76	31.57	53.64	28.51	50.54
	CIReVL (ICLR'24)	✓	29.49	47.40	24.79	44.76	31.36	53.65	28.55	48.57
	ImageScope (WWW'25)	✓	27.82	41.76	20.18	37.48	28.61	44.42	25.54	41.22
	PrediCIR (CVPR'25)		31.80	52.00	25.40	49.50	33.10	55.40	30.10	52.30
	OSrCIR (CVPR'25)	✓	33.17	52.03	29.7	51.81	36.92	59.27	33.26	54.37
	MMRet-Large (ACL'25)†		37.04	56.13	29.84	50.66	37.07	59.01	34.65	55.27
	Ours+MMRet-large	✓	**39.1**	**58.34**	**31.33**	**52.35**	**39.67**	**61.65**	**36.70**	**57.45**
	Δ *(Ours vs MMRet-Large)*		*(+2.06)*	*(+2.21)*	*(+1.49)*	*(+1.69)*	*(+2.6)*	*(+2.64)*	*(+2.05)*	*(+2.18)*

stability. Furthermore, incorporating captions as contextual information ("+ optimized caption") enhances the retrieval performance. Additionally, we separately evaluated performance under conditions where the model is allowed to reason independently without using carefully prepared CoT templates("w/o CoT"), and where it is instructed to output the final answer directly without display-

Table 3: **Evaluation on GeneCIS Test Data.** Methods with "†" are our implementations. Green highlighting indicates the best performance, while blue highlighting indicates the second-best performance.

	Architecture	Focus Attribute			Change Attribute			Focus Object			Change Object			Avg.
		R@1	R@ 2	R@3	R@1	R@ 2	R@3	R@1	R@ 2	R@3	R@1	R@ 2	R@3	R@1
CLIP-ViT-L	SEARLE-XL (ICCV'23)	17.00	29.70	40.70	16.40	25.30	34.10	8.00	16.90	25.60	7.90	16.80	24.80	12.30
	LinCIR (CVPR'24)	16.90	30.00	41.50	16.20	28.00	36.80	8.30	17.40	26.20	7.40	15.70	25.00	12.20
	CIReVL (ICLR'24)	19.50	31.80	42.00	14.40	26.00	35.20	12.30	21.80	30.50	17.20	28.90	37.60	15.90
	OSrCIR (CVPR'25)	20.90	**33.10**	**44.50**	17.20	28.50	**37.90**	15.00	23.60	34.20	18.40	30.60	38.30	17.90
	MMRet-Large (ACL'25)†	18.95	29.94	39.55	14.59	26.75	36.41	16.87	25.82	36.53	18.12	31.08	39.18	17.13
	Ours+MMRet-Large	**21.65**	31.65	41.55	**17.57**	**28.60**	37.17	**19.10**	**27.41**	**37.82**	**19.81**	**32.31**	**41.43**	**19.53**
	Δ (Ours vs MMRet-Large)	*(+2.70)*	*(+1.71)*	*(+2.00)*	*(+2.98)*	*(+1.85)*	*(+0.76)*	*(+2.23)*	*(+1.59)*	*(+1.29)*	*(+1.69)*	*(+1.23)*	*(+2.25)*	*(+2.40)*

Figure 3: Examples from the CIRR validation set where our method retrieves the desired image, in comparison to the MMRet-large baseline. Our approach utilizes two additional rounds (Round 1, Round 2) of iterative synthetic caption generation to refine the retrieval process. The green box indicates the ground truth image. Within the synthetic captions, green highlighting marks correct key visual elements, while pink highlighting denotes elements irrelevant to the target image.

ing its reasoning process("w/o think process"), to demonstrate the validity of our design.

Table 4: Ablation study on CIRCO and FashionIQ data using MMRet-Large as the baseline.

Method	CIRCO				Fashion-IQ	
	k=5	k=10	k=25	k=50	k=10	k=50
Impact of Different Components						
baseline	37.75	38.60	41.22	42.09	34.65	55.27
+ reflection	36.66	37.09	39.69	40.59	30.82	50.71
+ query fusion	40.82	41.26	43.83	44.79	36.26	57.04
+ optimized caption	42.77	43.08	45.64	46.62	36.70	57.45
Impact of Different Prompt Strategies						
w/o CoT	41.45	41.88	44.37	45.36	35.68	56.38
w/o think process	40.77	40.99	43.66	44.54	35.97	56.57
Impact of MLLM Choice						
Qwen-2.5VL-3B	39.40	39.82	42.38	43.32	34.88	55.50
LLaVA-NeXT-7B	38.98	39.93	42.64	43.77	34.82	55.39
Qwen-2.5VL-7B	39.31	40.11	42.54	43.57	34.72	55.22
Qwen-2.5VL-72B	41.50	42.12	44.59	45.52	36.40	56.99

Impact of MLLM Choice. To verify the generalizability of our framework, we evaluate its performance through different open source MLLMs. As shown in the Table 4, our method achieves significant performance gains even with relatively small models like Qwen-2.5VL-3B and LLaVA-NeXT-7B and the gains grow with larger models.

Impact of Stopping Rounds. In Section 4.1 we report the results of adding an additional two rounds. Here we give the results for more rounds in the Figure 4. As can be seen, the gains become marginal when going beyond 2 rounds, and for cost reasons we consider 2 rounds to be the best stopping point. Additionally, since our framework is a multi-round iterative process, we also investigate the temporal cost of our method and more diverse stopping conditions. Each additional iteration adds approximately 3 seconds of latency, yet the overall runtime remains competitive compared to other methods.

Impact of Top-K Images and Top-N Captions. As shown in Figure 5 (a), we investigated the effect of varying the number of retrieved top-K images and top-N captions as the context for MLLM. The results reveal that K=5 and N=10 achieves the highest scores across all metrics. We can find that using too few images and captions leads to a lack of information for effective reasoning, while using too many introduces noise from irrelevant features that

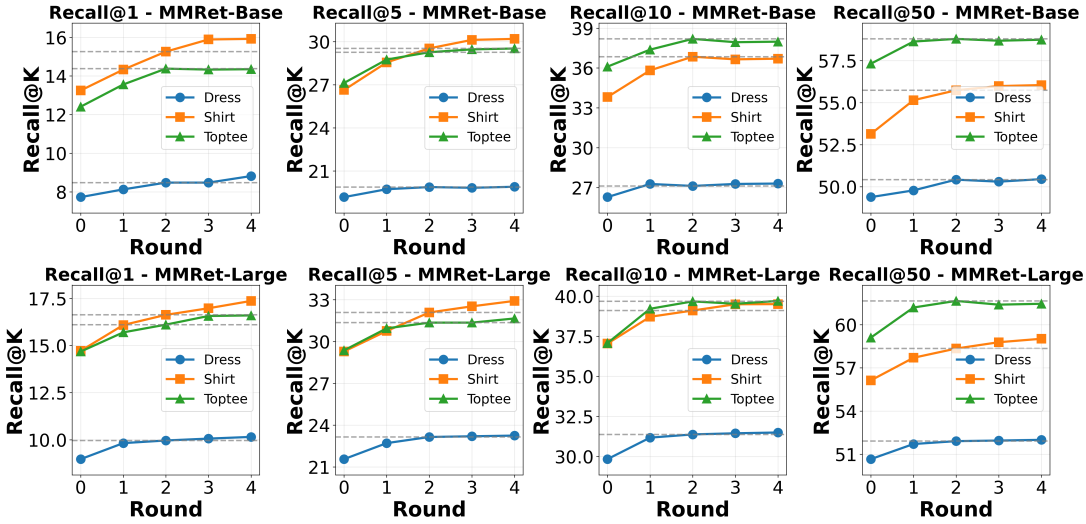

Figure 4: Recall@K Performance Analysis Across Rounds on FashionIQ validation data.

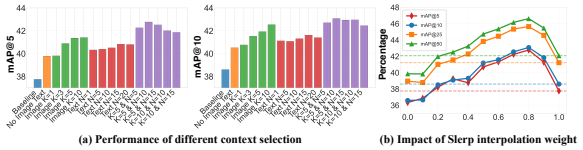

Figure 5: Performance comparison of different context selection strategies and Impact of Slerp interpolation weight on CIRCO validation data.

Table 5: Ablation study on FashionIQ validation dataset across different query fusion strategies.

Method	Dress		Shirt		TopTee	
	R@10	R@50	R@10	R@50	R@10	R@50
baseline	29.84	50.66	37.04	56.13	37.07	59.01
linear interpolation	28.84	50.72	38.17	57.70	37.97	60.08
pseudo-relevance feedback	27.56	46.05	32.82	48.77	33.45	52.12
proposed	31.33	52.35	39.10	58.34	39.67	61.65

can degrade retrieval accuracy.

Impact of Different Slerp Interpolation Weight. As shown in Figure 5 (b), we analyze the impact of the interpolation weight α (as defined in Equation 3) on the performance of our method. The horizontal axis represents the value of α. As α increases, the proportion of the historical query vector increases. It can be seen that the best performance is achieved when $\alpha = 0.8$.

Impact of Different query fusion strategies. Additionally, we employ slerp as the strategy for historical query fusion. We present the results under linear interpolation to demonstrate slerp's effectiveness. In addition, we also update the query representation using the mean embedding of the top-5 retrieved images to simulate a pseudo-relevance feedback strategy. As shown in Figure 5, linear interpolation only yields a marginal performance improvement, whereas a simple pseudo-relevance feedback strategy degrades performance due to the noise it introduces.

5 Conclusion

In this paper, we present CoRR, a novel training-free framework that overcomes a key limitation of existing ZS-CIR methods. CoRR leverages an MLLM to analyze issues in previously retrieved results via Chain-of-Thought reasoning and iteratively refines the query to improve retrieval accuracy. To ensure stable refinement and strong alignment with the retrieval model, we incorporate Historical Query Fusion and Retrieval-Driven Caption Optimization. CoRR achieves state-of-the-art performance across multiple benchmarks, demonstrating the promising result of self-reflection-guided iterative retrieval. In the future, we plan to extend CoRR to broader domains such as video retrieval and multi-step reasoning tasks to further improve its adaptability in complex settings.

6 Limitations

While CoRR achieves significant performance improvements, it has several limitations. First, the iterative refinement process requires multiple MLLM calls, which increases computational time compared to single-pass methods (while still faster than LDRE (Yang et al., 2024) and OSrCIR (Tang et al., 2025b)). When limited to a single iteration, the time overhead (approximately 3s) is comparable to CIReVL (approximately 2s) (Karthik et al., 2024). Second, the framework relies on the reasoning capabilities of the underlying MLLM. The quality of query refinement directly depends on the model's ability to understand multimodal inputs, analyze retrieval feedback, and generate appropriate corrections. Performance may degrade when using less capable models or when dealing with complex scenarios that exceed the MLLM's comprehension limits.

References

Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. 2021. Compositional learning of image-text query for image retrieval. In *WACV*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025a. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. 2025b. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*.

Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. 2024. Sentence-level prompts benefit composed image retrieval. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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

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

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

Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. 2022. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5194–5205.

Gloria Bordogna and Gabriella Pasi. 1993. A fuzzy linguistic approach generalizing boolean information retrieval: A model and its evaluation. *Journal of the American Society for Information Science*, 44(2):70–82.

Rui Chen, Lei Sun, Jing Tang, Geng Li, and Xiangxiang Chu. 2025. Finger: Content aware fine-grained evaluation with reasoning for ai-generated videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3517–3526.

Yanbei Chen and Loris Bazzani. 2020. Learning joint visual semantic matching embeddings for language-guided retrieval. In *ECCV*.

Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020. Image search with text feedback by visiolinguistic attention learning. In *CVPR*.

Yixin Chen and James Ze Wang. 2002. A region-based fuzzy feature matching approach to content-based image retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 24(9):1252–1267.

Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. 2022. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. 2024. Language-only training of zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13225–13234.

Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic spatially-aware fashion concept discovery. In *ICCV*.

Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, Trishul Chilimbi, and Abhinav Shrivastava. 2025. Collm: A large language model for composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3994–4004.

Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. 2024. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. In *European Conference on Computer Vision*, pages 239–254. Springer.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. 2024. Vision-by-language for training-free compositional image retrieval. *International Conference on Learning Representations (ICLR)*.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435.

Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. Cosmo: Content-style modulation for image retrieval with text feedback. In *CVPR*.

Haiwen Li, Zining Chen, Ying Liu, Fei Su, and Zhicheng Zhao. 2025. Slot inversion for asymmetric composed image retrieval. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Haiwen Li, Delong Liu, Zhaohui Hou, Zeliang Ma, Fei Su, and Zhicheng Zhao. 2026. Modality and task adaptation for enhanced zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6100–6108.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

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

Pengfei Luo, Jingbo Zhou, Tong Xu, Yuan Xia, Linli Xu, and Enhong Chen. 2025. Imagescope: Unifying language-guided image retrieval via large multimodal model collective reasoning. In *Proceedings of the ACM on Web Conference 2025*, pages 1666–1682.

Jabez Magomere, Emanuele La Malfa, Manuel Tonneau, Ashkan Kazemi, and Scott A. Hale. 2025. When claims evolve: Evaluating and enhancing the robustness of embedding models against misinformation edits. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22374–22404, Vienna, Austria. Association for Computational Linguistics.

Mandar Mitra, Amit Singhal, and Chris Buckley. 1998. Improving automatic query expansion. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–214.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Xiangyan Qu, Gaopeng Gou, Jiamin Zhuang, Jing Yu, Kun Song, Qihao Wang, Yili Li, and Gang Xiong. 2025. Proapo: Progressively automatic prompt optimization for visual classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25145–25155.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Javad Rafiei Asl, Prajwal Panzade, Eduardo Blanco, Daniel Takabi, and Zhipeng Cai. 2024. RobustSentEmbed: Robust sentence embeddings using adversarial self-supervised contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3795–3809, Mexico City, Mexico. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *CVPR*.

Sakib Shahriar, Brady D Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. 2024. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Applied Sciences*, 14(17):7782.

Ken Shoemake. 1985. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254.

Zelong Sun, Dong Jing, and Zhiwu Lu. 2025. Cotmr: Chain-of-thought multi-scale reasoning for training-free zero-shot composed image retrieval. *arXiv preprint arXiv:2502.20826*.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Gaopeng Gou, and Qi Wu. 2025a. Missing target-relevant information prediction with world model for accurate zero-shot composed image retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24785–24795. Computer Vision Foundation / IEEE.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. 2024. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5180–5188.

Yuanmin Tang, Jue Zhang, Xiaoting Qin, Jing Yu, Gaopeng Gou, Gang Xiong, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Wu. 2025b. Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 14400–14410.

Sagar Vaze, Nicolas Carion, and Ishan Misra. 2023. Genecis: A benchmark for general conditional image similarity. In *CVPR*.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval-an empirical odyssey. In *CVPR*.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. 2025. Cross-modal retrieval: a systematic review of methods and future directions. *Proceedings of the IEEE*.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *CVPR*.

Eric Xing, Pranavi Kolouju, Robert Pless, Abby Stylianou, and Nathan Jacobs. 2025. Context-cir: Learning from concepts in text for composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19638–19648.

Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, and Changsheng Xu. 2024. Ldre: Llm-based divergent reasoning and ensemble for zero-shot composed image retrieval. In *Proceedings of the 47th International ACM SIGIR conference on research and development in information retrieval*, pages 80–90.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. 2025. MegaPairs: Massive data synthesis for universal multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19076–19095, Vienna, Austria. Association for Computational Linguistics.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.