Modality and Task Adaptation for Enhanced Zero-shot Composed Image Retrieval

Haiwen Li[1*], Delong Liu[1*], Zhaohui Hou[4], Zeliang Ma[4], Fei Su[1,2,3], Zhicheng Zhao[1,2,3†]

[1]Beijing University of Posts and Telecommunications
[2]Beijing Key Laboratory of Network System and Network Culture
[3]Key Laboratory of Intereactive Technology and Experience System
[4]SenseTime

Abstract

As a challenging vision-language task, Zero-Shot Composed Image Retrieval (ZS-CIR) is designed to retrieve target images using bi-modal (image+text) queries. Typical ZS-CIR methods employ an inversion network to generate pseudo-word tokens that effectively represent the input semantics. However, the inversion-based methods suffer from two inherent issues: First, the task discrepancy exists because inversion training and CIR inference involve different objectives. Second, the modality discrepancy arises from the input feature distribution mismatch between training and inference. To this end, we propose a lightweight post-hoc framework, consisting of two components: (1) A new text-anchored triplet construction pipeline leverages a large language model (LLM) to transform a standard image-text dataset into a triplet dataset, where a textual description serves as the target of each triplet. (2) The MoTa-Adapter, a novel parameter-efficient fine-tuning method, adapts the dual encoder to the CIR task using our constructed triplet data. Specifically, on the text side, multiple sets of learnable task prompts are integrated via a Mixture-of-Experts (MoE) layer to capture task-specific priors and handle different types of modifications. On the image side, MoTa-Adapter modulates the inversion network's input to better match the downstream text encoder. In addition, an entropy-based optimization strategy is proposed to assign greater weight to challenging samples, thus improving adaptation efficiency. Experiments show that, with the incorporation of our proposed components, inversion-based methods achieve significant improvements, reaching state-of-the-art performance across four widely-used benchmarks.

Code — https://github.com/JThuge/MoTa-Adapter

1 Introduction

Composed Image Retrieval (CIR) (Vo et al. 2019) has gained increasing attention. It retrieves target images matching a bi-modal query (reference image + relative caption), offering more flexible searches than traditional image retrieval (Liu et al. 2016) by integrating visual and textual inputs. Benefiting from large-scale vision-language pretraining (VLP) models (Radford et al. 2021; Jia et al. 2021; Li et al. 2022), CIR has advanced significantly and found applications in e-commerce, web search, and other real-world scenarios.

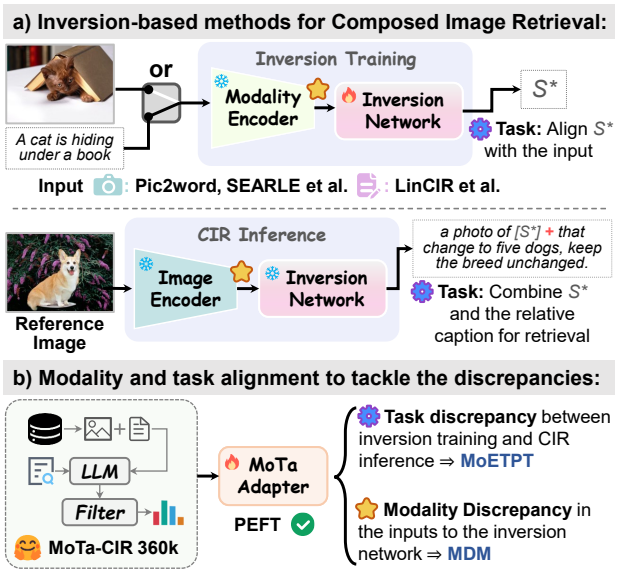


Figure 1: Motivation of our work. a) Existing inversion-based approaches, encompassing both image-based methods such as Pic2Word (Saito et al. 2023) and SEARLE (Baldrati et al. 2023), and text-based methods like LinCIR (Gu et al. 2024b). b) Our proposed lightweight post-hoc framework aims at addressing the task and modality discrepancies inherent in inversion-based CIR methods.

Supervised CIR approaches (Baldrati et al. 2022; Liu et al. 2024b; Xu et al. 2024) utilize labeled triplets (I_r, T_c, I_t), where I_r represents the reference image, T_c denotes the relative caption, and I_t is the target image. However, manual collection and annotation of such triplets are both labor-intensive and time-consuming, leading to relatively small-scale datasets (Liu et al. 2021; Wu et al. 2021) that limit generalization. Zero-Shot CIR (ZS-CIR) (Saito et al. 2023; Baldrati et al. 2023) addresses this by leveraging larger and more diverse publicly available datasets (Sharma et al. 2018; Schuhmann et al. 2022) covering a wide range of domains and semantics, enabling stronger generalization without reliance on such annotated triplets.

*Equal Contribution.
†Correspondence Author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Mainstream ZS-CIR approaches are primarily based on textual inversion (Cohen et al. 2022; Gal et al. 2022), and can be categorized into image-based (Saito et al. 2023; Baldrati et al. 2023) and text-based methods (Gu et al. 2024b). These methods train an inversion network that maps the features of an input image (or text) into a pseudo-word token, denoted as S^*. This token is subsequently incorporated into a prompt template (e.g., a photo of S^*) to represent the input semantics faithfully. However, as illustrated in Figure 1(a), the inversion-based methods face two key issues between training and inference: (1) ***Task discrepancy*** (Byun et al. 2024). The task of inversion training is to align the pseudo-word token S^* with the input image (or text), such that S^* serves as a good representation of the input. In contrast, the task during CIR inference is to compose a query from I_r and T_c to retrieve target images I_t, which is not tackled during training. (2) ***Modality discrepancy***. This issue is particularly prominent in text-based inversion training, where the image modality is absent during training but required at inference. Although LinCIR (Gu et al. 2024b) introduces random perturbations at the feature level to mitigate this discrepancy, we demonstrate that this strategy is far from sufficient.

To tackle the aforementioned two discrepancies and improve inversion-based approaches, we propose a lightweight post-hoc framework, as illustrated in Figure 1(b). It introduces an additional adaptation stage that adopts a parameter-efficient tuning strategy on automatically generated text-anchored triplets, comprising two parts: (1) ***MoTa-Adapter***. The task discrepancy arises on the query-side, as the inversion training does not involve the integration of I_r and T_c. Building on this, we perform task adaptation using the input in the form of (I_r, T_c) to learn the integration process on the query side. Specifically, we insert several learnable task prompts into the input of the text encoder and integrate them through a Mixture-of-Experts (MoE) (Jacobs et al. 1991) mechanism to capture CIR task priors, and meanwhile handle diverse types of modifications associated with T_c. In addition, since the modality discrepancy exists between the inversion network's inputs during inversion training and CIR inference. To deal with it, we propose Modality Distribution Modulation (MDM), which adaptively shifts and aligns the feature distribution fed into the inversion network to mitigate the modality discrepancy and better support the downstream text encoder. (2) ***MoTa-CIR***. Considering that task and modality adaptation require triplet data involving I_r for training, we propose a pipeline to automatically generate text-anchored triplets in the form of (I_r, T_c, T_t). Specifically, given a standard image-text dataset, we guide an LLM to expand each image-text pair (I_r, T_r) into a triplet by generating T_c and T_t based on T_r. Compared with using I_t as the target, T_t offers lower computational cost and better training efficiency. Moreover, since CIR resembles a fuzzy retrieval task (Bordogna and Pasi 1993; Chen and Wang 2002) and there may be many valid target images, using a textual target T_t for training facilitates robust learning. Finally, after filtering, we obtain a high-quality text-anchored triplet dataset, MoTa-CIR, comprising approximately 360k samples. In summary, our contributions are fourfold:

- We propose a lightweight post-hoc framework that effectively mitigates the two inherent and interrelated task and modality discrepancies in inversion-based methods.
- We introduce the MoTa-Adapter, a novel parameter-efficient fine-tuning approach for CIR, which optimizes only a small set of learnable task prompts on the text side and a modulation layer on the image side, addressing the two discrepancies simultaneously.
- To facilitate task and modality adaptation, we propose a scalable pipeline for the automatic construction of text-anchored triplets, resulting in a diverse and high-quality dataset, MoTa-CIR. Additionally, we introduce a novel entropy-based loss weighting strategy for efficient training on the text-anchored triplets.
- When our proposed modules are integrated into the existing inversion-based methods, the performance is significantly enhanced across four widely-used benchmarks.

2 Related Work

Composed Image Retrieval (CIR) is primarily evaluated in the fashion (Wu et al. 2021) and real-world (Liu et al. 2021; Baldrati et al. 2023) domains. Mainstream supervised approaches (Baldrati et al. 2022; Liu et al. 2024b; Xu et al. 2024) leverage the cross-modal alignment capabilities of the VLP models and adopt either early or late fusion to integrate the two modalities in composed queries. Recent work explores zero-shot CIR, with textual inversion (Gal et al. 2022; Cohen et al. 2022) emerging as a key technique. Representative methods, including image-based approaches such as Pic2Word (Saito et al. 2023) and SEARLE (Baldrati et al. 2023), as well as the text-based methods like LinCIR (Gu et al. 2024b), train an inversion network to generate a pseudo-word token that effectively captures and represents the input semantics. Additionally, some attempts (Tang et al. 2024; Byun et al. 2024; Wang et al. 2025; Tang et al. 2025) have improved upon inversion training, in which most related to ours is RTD (Byun et al. 2024). Unlike RTD, which only fine-tunes the text encoder to reduce the task discrepancy, our approach introduces a lightweight adapter that enables multi-modal adaptation of the dual encoder, effectively addressing the inherent limitations of inversion-based methods. Other works explore automatic CIR triplet construction (Ventura et al. 2024; Levy et al. 2024; Gu et al. 2024a) and training-free approaches based on LLM reasoning (Karthik et al. 2024; Yang et al. 2024). While both have shown promising results, they are limited by the quality of constructed triplets and model complexity, respectively.

Parameter-Efficient Fine-Tuning (PEFT) aims to adapt large-scale pretrained VLP models by updating only a small amount of task-specific parameters, reducing overhead while preserving performance. Methods such as Adapter Tuning (Houlsby et al. 2019), Prompt Tuning (Lester, Al-Rfou, and Constant 2021), and LoRA (Hu et al. 2021) have shown strong results in NLP. In the vision-language domain, approaches like CoOp (Zhou et al. 2022b), CoCoOp (Zhou et al. 2022a) and Maple (Khattak et al. 2023) adapt CLIP using context prompts or prompt modules for better generalization and robustness. Our work builds on this line by

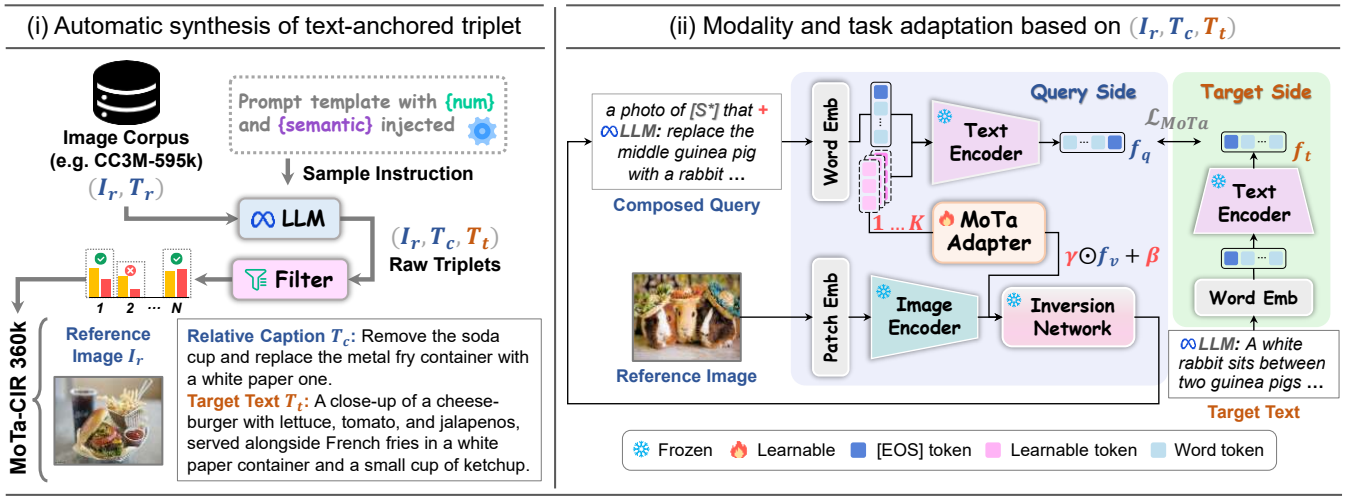


Figure 2: Overview of our proposed framework. (i) A standard image-text dataset containing (I_r, T_r) is first selected. An LLM is guided to generate diverse (T_c, T_t) based on T_r. Subsequently, the raw triplets (I_r, T_c, T_t) are scored and filtered by an MLLM to form the MoTa-CIR dataset. (ii) Based on the text-anchored triplets constructed in (i), the lightweight MoTa-Adapter is incorporated to explicitly reduce task and modality discrepancies. It operates on both the text and image sides, enhancing the model's capacity to interpret composed queries, thereby improving overall CIR performance.

introducing a PEFT-based strategy tailored for CIR, jointly addressing both task and modality discrepancies.

3 Methodology

In this section, we introduce our proposed approach in detail. 3.1 presents the pipeline for automatically constructing text-anchored triplets, including diversified triplet expansion and data filtering. 3.2 describes the modality and task adaptation, including the design of the MoTa-Adapter and its role in model optimization and the inference workflow.

3.1 Automatic Synthesis of Text-Anchored Triplet

As supervised CIR relies on manually annotated triplets, several studies have proposed using cheaply and automatically collected triplets for training. Among these, the most closely related to our work is RTD (Byun et al. 2024), but it adopts text-only triplets in the form of (T_r, T_c, T_t).

Diversified Triplet Expansion. Unlike RTD, which attributes the task discrepancy only to the text encoder, we argue that the discrepancy lies in the dual-encoder structure. In particular, the integration of the reference image I_r and the relative caption T_c on the query side is essential for effectively bridging the task gap. To this end, one or several publicly available image-text datasets are selected, and their union is denoted as $\mathcal{D} = \{I_r^i, T_r^i\}_{i=1}^N$. For each T_r^i, a template $\mathcal{P}(num, semantic)$, as illustrated below, guides an LLM to generate a relative caption T_c^i and a target text T_t^i, resulting in a triplet of the form (I_r^i, T_c^i, T_t^i).

> Given the reference caption $\{\boldsymbol{T_r}\}$, please generate an editing instruction involving $\{\boldsymbol{num}\}$ modifications on the $\{\boldsymbol{semantic}\}$ aspect, along with a target text.

Note that num is sampled from 1 to 3, and $semantic$ is drawn from a predefined set covering operations such as attribute modification, substitution, and removal. This guarantees that T_c includes varying types and quantities of modifications, thereby achieving satisfactory semantic diversity.

Data Filtering. To enhance the overall quality of the generated triplets, we use a multi-modal large language model (MLLM) as a filtering module to score each triplet on two dimensions, ranging from 1 to 10: (1) Triplet alignment: how well the composition of I_r and T_c semantically aligns with T_t, and (2) Variety: whether T_t involves meaningful and diverse modifications relative to T_r. For each dimension, the bottom 30% of triplets are discarded based on their scores.

Using this pipeline, we construct a diverse and high-quality text-anchored triplet dataset, MoTa-CIR-360k. This dataset is used to support the modality and task adaptation.

3.2 Modality and Task Adaptation

A brief introduction of CLIP and inversion-based methods is first given, and then our proposed MoTa-Adapter and entropy-based optimization are introduced in detail.

Preliminary. CIR models are mainly built on CLIP (Radford et al. 2021), which is trained on massive image-text pairs and exhibits strong cross-modal alignment capabilities. On the text side, CLIP employs Byte-Pair Encoding (BPE) (Shin et al. 2020) to tokenize the input into a sequence of tokens $E = \{e_i \in \mathbb{R}^{d_e}\}_{i=1}^M$, which is then fed into the text encoder $\mathcal{T}$ to obtain the text feature $f_t = \mathcal{T}(E) \in \mathbb{R}^d$. On the image side, it follows a similar process by extracting image features $f_v \in \mathbb{R}^d$ using the image encoder $\mathcal{I}$. The resulting features f_t and f_v are aligned in a joint embedding space via a contrastive loss (He et al. 2020).

Inversion-based approaches utilize an inversion network ϕ to project an input image (or text) feature $f \in \mathbb{R}^d$ into

the textual embedding space, producing a pseudo-word token $e_{s^*} = \phi(f) \in \mathbb{R}^{d_e}$ that is inserted into the prompt "a photo of S^*". After tokenization, the resulting sequence is fed into $\mathcal{T}$ to obtain the inverse feature $f^* \in \mathbb{R}^d$, which is trained to be close to the original feature f by contrastive learning. This allows "a photo of S^*" to serve as an effective representation of the input image (or text).

MoTa-Adapter. However, we notice that the lack of integration between I_r and T_c during inversion training is the primary cause of the task and modality discrepancies. Therefore, it is necessary to adapt the dual encoder.

The modality discrepancy arises because of the misalignment in the feature distributions input to the ϕ network during inversion training and CIR inference. Although VLMs like CLIP have performed cross-modal alignment, internal covariance and scale differences still exist within the distributions of visual and textual modalities. Therefore, as shown in Figure 2(ii), MoTa-Adapter incorporates Modality Distribution Modulation (MDM) on the image side, where a modulation layer after the image encoder $\mathcal{I}$ is built to adaptively shift the distribution of f_v using a learnable scale parameter $\gamma \in \mathbb{R}^d$ and a learnable shift parameter $\beta \in \mathbb{R}^d$:

$$\tilde{f}_v = \gamma \odot f_v + \beta, \tag{1}$$

where $\odot$ represents the inner product, and the modulated image features are then mapped to the text side using the inversion network, i.e., $e_{s^*} = \phi(\tilde{f}_v) \in \mathbb{R}^{d_e}$, in which d_e is the dimension of the textual embedding space. MDM not only reduces the modality discrepancy but also integrates the internal information of the image features to better resolve task discrepancy. In other words, the two discrepancies are interrelated and require simultaneous adaptation.

Meanwhile, the task adaptation should enable the text encoder $\mathcal{T}$ to understand "a photo of S^* that T_c" rather than just "a photo of S^*". Building on this, our MoTa-Adapter introduces Mixture-of-Experts Task Prompt Tuning (MoE-TPT) on the text side, which inserts K experts, each containing N learnable task prompts $P_k = \{p_i \in \mathbb{R}^{d_e}\}_{i=1}^N, k = 1, 2, ..., K$, into the input token sequence of $\mathcal{T}$. The K experts are fused through a routing function $\mathcal{R}$ and are conditioned on the pseudo-word representation e_{s^*}:

$$\begin{aligned} \mathcal{R}(x)_k &= \text{Softmax}(\mathcal{W}x)_k \\ P &= \Sigma_{k=1}^K P_k \cdot \mathcal{R}(e_{s^*})_k, \end{aligned} \tag{2}$$

where $\mathcal{W} \in \mathbb{R}^{d_e \times K}$ is a linear layer, generating the weights of K experts based on e_{s^*}. Subsequently, P is concatenated with other textual tokens and input into $\mathcal{T}$ to obtain the composed feature $\tilde{f}_c = \mathcal{T}([P, E])$, which is then optimized to inject the task-specific priors into the CIR model. The MoE structure provides two main advantages for the CIR task: (1) ***Task Specialization***, each expert focus on different types of modifications in the relative captions, such as attribute manipulation, substitution, or removal. (2) ***Sample Customization***, the routing function $\mathcal{R}$ generates weights based on the pseudo-word representation e_{s^*}, assigning the optimal mixture of experts for each sample.

Entropy-based Optimization. Given that the inversion-trained models already possess a certain level of CIR capabilities, we propose an entropy-based optimization strategy that assigns greater attention to challenging samples. Specifically, for each composed feature f_c^i produced by the base model (w/o MoTa-Adapter), the cosine similarities with all target text features $\{f_t^j\}_{\mathcal{B}}$ in a batch are computed. Based on the predicted probabilities, the entropy can be quantified:

$$\begin{aligned} p_{c2t}^{(i,j)} &= \frac{e^{sim(f_c^i, f_t^j)}}{\Sigma_{j \in \mathcal{B}} e^{sim(f_c^i, f_t^j)}} \\ H_{c2t}^i &= -\sum_j p_{c2t}^{(i,j)} \log p_{c2t}^{(i,j)}, \end{aligned} \tag{3}$$

where $sim(\cdot, \cdot)$ denotes the cosine similarity. A larger entropy H_{c2t}^i indicates higher predictive uncertainty of the base model for f_c^i, suggesting it should be assigned greater weight during the adaptation. We normalize and apply exponential smoothing to compute the sample-wise weight w_{c2t}^i:

$$\hat{H}_{c2t}^i = \frac{H_{c2t}^i}{\log |\mathcal{B}|} \quad w_{c2t}^i = e^{\beta \hat{H}_{c2t}^i}, \tag{4}$$

the division is based on Jensen's inequality, yielding $H_{c2t}^i \leq \log |\mathcal{B}|$, where $|\mathcal{B}|$ denotes the batch size. β is a hyperparameter controlling the sharpness of the exponential smoothing function. Similarly, the weight w_{t2c}^i for each target text feature f_t^i can be obtained. Finally, the tuned model (w/ MoTa-Adapter) generates $\tilde{f}_c$ and $\tilde{f}_t$ that are optimized using a weighted contrastive loss, i.e. $\mathcal{L}_{MoTa}$, in which τ is a learnable temperature parameter,

$$\begin{aligned} \mathcal{L}_{MoTa}(\mathcal{B}) = &- \sum_{i=1}^{|\mathcal{B}|} w_{c2t}^i \cdot \log[\frac{e^{sim(\tilde{f}_c^i, \tilde{f}_t^i)/\tau}}{\Sigma_{j \in \mathcal{B}} e^{sim(\tilde{f}_c^i, \tilde{f}_t^j)/\tau}}] \\ &- \sum_{i=1}^{|\mathcal{B}|} w_{t2c}^i \cdot \log[\frac{e^{sim(\tilde{f}_t^i, \tilde{f}_c^i)/\tau}}{\Sigma_{j \in \mathcal{B}} e^{sim(\tilde{f}_t^i, \tilde{f}_c^j)/\tau}}]. \end{aligned} \tag{5}$$

The lower bound occurs when all samples have identical predicted entropy, resulting in equal weights. In this case, the loss degenerates into the standard contrastive loss.

3.3 Inference Workflow

During inference, the image encoder is used to extract features of an image gallery, resulting in a set $\mathcal{S} = \{f_v^i\}_{i=1}^N$. After that, given an input composed query (I_r, T_c), the composed feature $\tilde{f}_c$ is obtained following the procedure described in Section 3.2. The cosine similarity between $\tilde{f}_c$ and each target image feature in $\mathcal{S}$ is computed, with the top-K most similar items returned as the retrieval results.

4 Experiments

4.1 Experimental Setup

Evaluation benchmarks. FashionIQ (Wu et al. 2021) simulates a realistic online shopping environment, featuring images in fashion domain. It comprises 30,134 triplets derived from 77,684 images. CIRR (Liu et al. 2021) is the first open-domain dataset in CIR, collecting 21,552 real-life images, with human-annotated relative captions. CIRCO (Baldrati et al. 2023) utilizes real-world images from the COCO

Method	Backbone	Zero-shot Composed Image Retrieval							
		FashionIQ		CIRR			CIRCO		GeneCIS
		R@10	R@50	R@1	R@5	R@10	mAP@5	mAP@10	R@1
CoVR-BLIP (Ventura et al. 2024)	BLIP	27.70	44.63	38.48	66.70	77.25	21.43	22.33	-
CASE (Levy et al. 2024)	BERT+ViT	-	-	35.40	65.78	78.53	-	-	-
CIReVL† (Karthik et al. 2024)	CLIP-G	32.19	34.65	64.29	67.95	75.06	26.77	27.59	17.4
CompoDiff† (Gu et al. 2024a)	CLIP-G	39.02	51.71	26.71	55.14	74.52	15.33	17.71	15.5
Comparison with methods based on textual inversion									
ContextI2W (Tang et al. 2024)	CLIP-L	27.80	48.90	25.60	55.10	68.50	13.00	14.60	12.7
KEDs (Suo et al. 2024)		26.80	47.90	26.40	54.80	67.20	-	-	-
PrediCIR (Tang et al. 2025)		30.10	52.30	27.20	57.00	70.20	15.70	17.10	16.6
Image2Sentence (Du et al. 2024)	BLIP	29.79	49.19	29.68	58.72	70.79	9.67	10.32	-
Slerp+TAT (Jang et al. 2024)	BLIP	32.77	53.32	33.98	61.74	72.70	17.84	18.44	-
PrediCIR (Tang et al. 2025)	CLIP-G	47.20	67.80	37.00	66.10	77.90	23.70	24.60	18.7
Pic2Word (Saito et al. 2023)	CLIP-L	24.70	43.70	23.90	51.70	65.30	8.72	9.51	11.2
+ CIG† (Wang et al. 2025)		25.16	44.85	24.63	52.75	65.28	-	-	-
+ RTD (Byun et al. 2024)		27.59	**48.90**	27.86	56.24	68.48	9.13	9.63	11.9
+ MoTa-Adapter (Ours)		**27.73**	48.15	**28.06**	**57.22**	**69.37**	**9.79**	**10.18**	**12.1**
SEARLE (Baldrati et al. 2023)		25.56	46.23	24.24	52.48	66.29	11.68	12.73	12.3
+ CIG† (Wang et al. 2025)		25.66	46.50	26.72	55.52	68.10	12.84	13.64	-
+ RTD (Byun et al. 2024)		**29.34**	**50.73**	26.63	56.17	68.96	16.53	17.89	12.4
+ MoTa-Adapter (Ours)		27.78	48.51	**28.19**	**57.95**	**69.98**	**16.77**	**17.94**	**14.7**
LinCIR (Gu et al. 2024b)		26.28	46.49	25.04	53.25	66.68	12.59	13.58	12.2
+ CIG† (Wang et al. 2025)		26.60	47.22	26.17	54.94	67.64	12.84	13.77	12.2
+ RTD (Byun et al. 2024)		**30.24**	**51.08**	26.63	56.17	68.96	17.11	18.11	13.2
+ MoTa-Adapter (Ours)		28.76	49.53	**28.02**	**58.62**	**71.06**	**18.22**	**19.46**	**16.7**
LinCIR (Gu et al. 2024b)	CLIP-G	45.11	65.69	35.25	64.72	76.05	19.71	21.01	13.6
+ CIG† (Wang et al. 2025)		45.80	66.35	35.47	66.00	76.89	20.62	21.82	13.8
+ RTD (Byun et al. 2024)		46.21	67.26	36.31	67.47	78.31	21.08	22.29	-
+ MoTa-Adapter (Ours)		**47.06**	**67.37**	**38.39**	**69.47**	**80.05**	**25.82**	**27.06**	**19.1**

Table 1: Performance comparison with existing zero-shot CIR methods. The best results are marked in **bold**. † indicates the method that, in addition to using a retrieval backbone, incorporates complex auxiliary network structures such as Diffusion Models (DMs) or Large Language Models (LLMs), which results in lower inference efficiency.

Method	CIRR MeanR	FashionIQ MeanR	CIRCO mAP@5
LinCIR(Gu et al.)	58.67	55.40	19.71
+ Full Fine-tuning	59.76	55.97	22.43
+ CoOp(Zhou et al.)	61.48	56.55	24.89
+ CoCoOp(Zhou et al.)	60.19	57.04	21.42
+ Maple(Khattak et al.)	61.50	56.87	25.21
+ MoTa-Adapter (Ours)	**62.64**	**57.22**	**25.82**

Table 2: Results with different tuning strategies. Experiments are conducted based on the CLIP-G backbone.

Dataset	CIRR MeanR	FashionIQ MeanR
LinCIR (Gu et al. 2024b)	58.67	55.40
+ST18M (Gu et al. 2024a)	50.18	52.76
+LaSCo (Levy et al. 2024)	53.74	55.80
+WebVid-CoVR (Ventura et al. 2024)	58.25	56.18
+MoTa-CIR (Ours)	**62.64**	**57.22**

Table 3: Results with different datasets. Experiments are conducted based on the MoTa-Adapter, using CLIP-G.

dataset (Lin et al. 2014) to develop a benchmark tailored for ZS-CIR with multiple ground truths. GeneCIS (Vaze, Carion, and Misra 2023) assesses the models' ability to adapt to various notions of visual similarity given different text prompts.

Evaluation metrics. Evaluation of the model performance primarily employs the Rank-K (R@K) metric, which measures the probability of finding at least one target image within the top-K candidates. Specifically for CIRCO, the mean Average Precision (mAP) is the main criterion. Higher values in R@K and mAP indicate better performance.

Implementation details. (1) For the triplet expansion in Section 3.1, we select LLaMA3-8B (Dubey et al. 2024) and convert the CC3M-595k (Liu et al. 2024a) dataset into a text-anchored triplet dataset. We then apply Qwen2.5-VL-32B (Bai et al. 2025) for filtering, resulting in our

Model	Filter	CIRR MeanR	FashionIQ MeanR
LLaMA3-8B(Dubey et al.)	✗	62.26	56.81
	✓	**62.64**	57.22
Qwen2.5-7B(Yang et al.)	✗	62.07	56.68
	✓	62.30	56.87
Qwen2.5-32B(Yang et al.)	✗	62.03	56.74
	✓	62.45	**57.26**

Table 4: Results with different LLMs for the triplet expansion. Experiments are conducted based on LinCIR+MoTa-Adapter, using CLIP-G as the backbone.

Components	CIRR MeanR	CIRCO mAP@5
LinCIR(Gu et al.)	58.67	19.71
w/ TPT (CoOp)	61.48	24.89
w/ MoE-TPT	61.76	25.09
w/ MDM	59.80	22.16
w/ MoE-TPT+MDM	62.12	25.43
w/ MoE-TPT+MDM+EBO	**62.64**	**25.82**

Table 5: Ablation study. Experiments are conducted based on the CLIP-G backbone.

MoTa-CIR-360k. (2) During the adaptation stage, for the MoTa-Adapter, we select $K = 4$ experts, each associated with $N = 8$ task prompts. We employ the AdamW optimizer (Loshchilov and Hutter 2019) with a learning rate of $2e-3$, a weight decay of 0.01, and a batch size of 256. The model is trained for $1,000$ steps including 100 warm-up steps on a single NVIDIA A100 GPU.

4.2 Quantitative Results

Comparison with State-of-the-Art Methods. As shown in Table 1, we conduct experiments based on three baselines: two image-based methods, i.e., Pic2Word and SEARLE, and a text-based method LinCIR. (1) ***Consistent Improvements.*** MoTa-Adapter consistently improves model performance across all evaluation metrics, no matter which baseline is used. This demonstrates its effectiveness in addressing the inherent limitations of inversion-based methods. (2) ***Essence of Discrepancy.*** The task discrepancy lies in the model's inability to effectively integrate I_r and T_c on the query side. RTD fine-tunes only the text encoder on pure text triplets, thus failing to fully resolve this issue. In contrast, our MoTa-Adapter adapts the dual encoder, reducing both task and modality discrepancies at a deeper level. This is further supported by empirical results, where MoTa-Adapter significantly outperforms both RTD and CIG on CIRR, CIRCO, and GeneCIS. On FashionIQ, MoTa-Adapter slightly underperforms RTD, which can be attributed to the domain gap between the real-world text-anchored triplets we construct and the fashion-related data in FashionIQ. (3) ***Efficient Optimization.*** MoTa-Adapter is more efficient and resource-friendly. For example, CIG introduces T2I models (Rombach et al. 2022), which incur

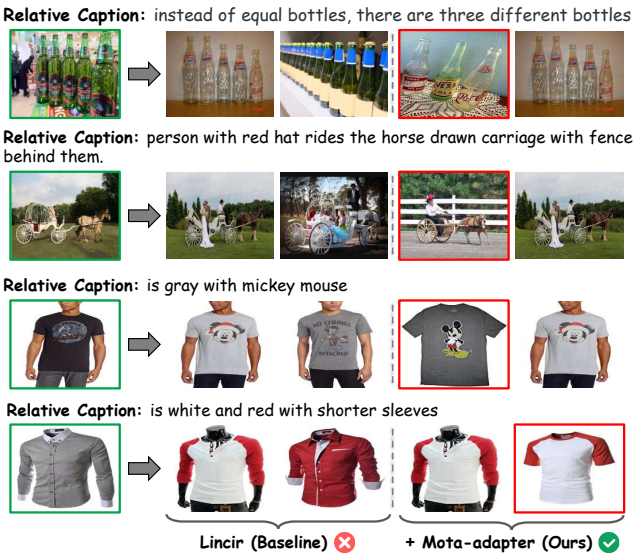


Figure 3: Qualitative results on CIRR and FashionIQ. The reference image and its corresponding target image are highlighted with green and red outline.

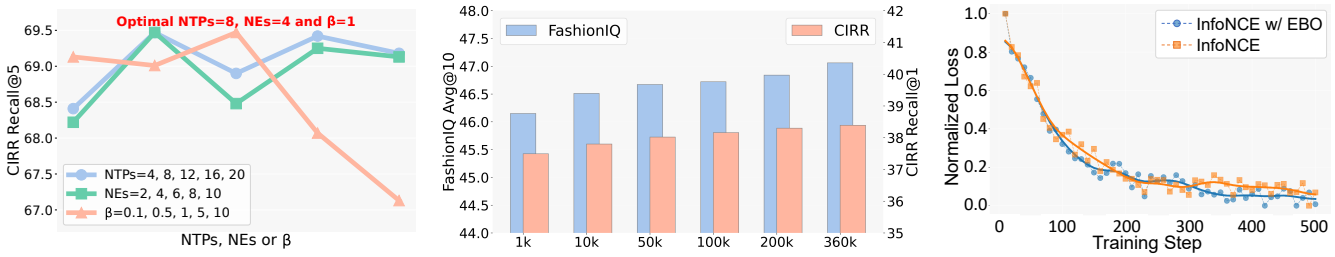
substantial computational costs during both training and inference. Training-free approaches such as CIReVL leverage LLMs for reasoning, leading to high resource consumption and slow inference speed. In contrast, our method introduces minimal learnable parameters and completes training in just 2 hours, making it both efficient and convenient. (4) ***Inference Efficiency.*** We evaluate the per-query inference time on a single A100 GPU. Built on CLIP-G, LinCIR requires **0.011s**, and with the MoTa-Adapter, the time increases to **0.013s**. In contrast, CIReVL takes over **1s**, indicating that MoTa-Adapter introduces minimal inference overhead.

Comparison with Different Tuning Strategies. Table 2 compares several tuning strategies on the MoTa-CIR dataset. (1) PEFT-based methods outperform full fine-tuning while being more lightweight. For example, MoTa-Adapter uses fewer than **1M** trainable parameters, whereas full fine-tuning updates the entire CLIP dual encoder (**441M**). (2) Multi-modal fine-tuning (Maple and MoTa-Adapter) is more effective than uni-modal approaches (CoOp and CoCoOp), and our MoTa-Adapter, specifically designed for the CIR task, achieves the best results.

Comparison with Different Training Datasets. Table 3 shows that only our MoTa-CIR consistently improves performance, while the other three datasets lead to either degradation or no clear gain. This suggests that MoTa-Adapter, as a PEFT method, is sensitive to data quality. Our pipeline generates high-quality, well-aligned text-anchored triplets, whereas the others may introduce noise due to low-quality images, repetitive relative captions, and poor alignment.

4.3 Qualitative Results

Figure 3 presents qualitative predictions from the baseline LinCIR (Gu et al. 2024b) as well as the results after integrating the MoTa-Adapter, which further support the effective-



(a) Impact of NTPs, NEs and β. (b) Data efficiency of MoTa-Adapter. (c) Visualization of training loss.

Figure 4: Hyperparameter analysis. Experiments are built upon LinCIR+MoTa-Adapter, using CLIP-G as the backbone.

ness of our method. MoTa-Adapter is capable of handling fine-grained semantic relationships across various scenarios, such as object composition (row 2-3), attribute manipulation (row 4) and quantity change (row 1), in which LinCIR fails.

4.4 Ablation Study

Architectural Design. Table 5 summarizes several model variants. TPT, which only introduces a single set of learnable task prompts (equivalent to CoOp (Zhou et al. 2022b)), improves performance, indicating its role in reducing task discrepancy. Adding the MoE mechanism (MoE-TPT) yields further gains, suggesting its effectiveness in capturing different types of modifications. While MDM also brings performance gains, its impact is smaller than that of MoE-TPT, implying that task discrepancy is the more dominant factor. Combining MoE-TPT with MDM leads to additional enhancements, highlighting the presence of two intertwined discrepancies that need to be addressed simultaneously. Finally, applying entropy-based optimization (EBO) achieves the overall best performance.

Impact of Different LLMs. We adopt different LLMs for text-anchored triplet construction, and the results, as shown in Table 4, are quite similar. Considering both performance and efficiency, we finally select LLaMA3-8B (Dubey et al. 2024) as our model of choice. Additionally, applying data filtering leads to further performance gains, demonstrating the effectiveness of our proposed filtering strategy.

Sensitivity Analysis. Figure 4(a) illustrates the impact of different hyperparameters, including the number of task prompts (NTPs), the number of experts (NEs), and the exponential smoothing hyperparameter β in EBO. The performance generally shows an initial increase followed by a decline, with the optimal hyperparameter combination, i.e., $NTPs = 8, NEs = 4, \beta = 1$, identified.

Data efficiency of MoTa-Adapter. As shown in Figure 4(b), the performance reaches satisfactory levels with tens of thousands of samples, highlighting the data efficiency of our method. Further increases in data size result in marginal improvements; therefore, the full set of 360k generated triplets is released to support future research efforts.

Visualization of Training Loss. Figure 4(c) visualizes the normalized training loss. After applying entropy-based optimization (EBO), the loss decreases more rapidly and stabilizes, maintaining a lower value towards the end, which contributes to improved model performance.

5 Discussion

Although our method greatly enhances the performance of inversion-based methods, it still has certain limitations. (1) The first issue relates to generalization, to be specific, in the construction of text-anchored triplets as discussed in Section 3.1. We expand the real-world dataset CC3M-595k (Liu et al. 2024a) into a text-anchored triplet dataset, i.e., MoTa-CIR-360k. This enables our method to achieve strong performance on real-domain benchmarks, including CIRR, CIRCO, and GeneCIS. Although it also leads to a notable improvement on FashionIQ, our approach slightly underperforms RTD (Byun et al. 2024), which is trained exclusively with textual supervision. Therefore, constructing multi-domain triplet datasets to improve generalization is a promising direction. (2) The generated relative captions tend to exhibit somewhat repetitive syntactic structures. It is worthwhile to explore whether supervised fine-tuning (SFT) of the LLM, or alternative strategies, can enhance the diversity of sentence structures and improve linguistic variation.

6 Conclusion

In this paper, we propose a lightweight post-hoc framework to improve inversion-based zero-shot CIR methods. Our framework includes MoTa-Adapter, a parameter-efficient tuning strategy, and a scalable text-anchored triplet construction pipeline. MoTa-Adapter adapts the dual encoder on the generated text-anchored triplets to reduce task and modality discrepancies by modulating the inversion network's input on the image side and introducing learnable task prompts on the text side. These prompts are integrated via an MoE mechanism to capture task-specific priors and handle different modifications in the relative caption. Notably, to avoid the high computational cost of full fine-tuning, we are the first to introduce a lightweight adapter for zero-shot CIR, significantly enhancing performance while reducing training overhead. Experiments on four widely used benchmarks demonstrate the effectiveness of our approach.

Acknowledgments

This work was supported by the Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing (GJJ-24-021) and the BUPT Innovation and Entrepreneurship Support Program (2025-YC-T032).

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Baldrati, A.; Agnolucci, L.; Bertini, M.; and Del Bimbo, A. 2023. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15338–15347.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Del Bimbo, A. 2022. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21466–21474.

Bordogna, G.; and Pasi, G. 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.

Byun, J.; Jeong, S.; Kim, W.; Chun, S.; and Moon, T. 2024. Reducing Task Discrepancy of Text Encoders for Zero-Shot Composed Image Retrieval. arXiv:2406.09188.

Chen, Y.; and Wang, J. Z. 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.

Cohen, N.; Gal, R.; Meirom, E. A.; Chechik, G.; and Atzmon, Y. 2022. "This is my unicorn, Fluffy": Personalizing frozen vision-language representations. In *European conference on computer vision*, 558–577. Springer.

Du, Y.; Wang, M.; Zhou, W.; Hui, S.; and Li, H. 2024. Image2sentence based asymmetrical zero-shot composed image retrieval. *arXiv preprint arXiv:2403.01431*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. arXiv:2208.01618.

Gu, G.; Chun, S.; Kim, W.; Jun, H.; Kang, Y.; and Yun, S. 2024a. CompoDiff: Versatile Composed Image Retrieval With Latent Diffusion. arXiv:2303.11916.

Gu, G.; Chun, S.; Kim, W.; Kang, Y.; and Yun, S. 2024b. Language-only training of zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13225–13234.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Jang, Y. K.; Huynh, D.; Shah, A.; Chen, W.-K.; and Lim, S.-N. 2024. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. In *European Conference on Computer Vision*, 239–254. Springer.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Karthik, S.; Roth, K.; Mancini, M.; and Akata, Z. 2024. Vision-by-Language for Training-Free Compositional Image Retrieval. arXiv:2310.09291.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045–3059.

Levy, M.; Ben-Ari, R.; Darshan, N.; and Lischinski, D. 2024. Data roaming and quality assessment for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2991–2999.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1096–1104.

Liu, Z.; Rodriguez-Opazo, C.; Teney, D.; and Gould, S. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2125–2134.

Liu, Z.; Sun, W.; Hong, Y.; Teney, D.; and Gould, S. 2024b. Bi-Directional Training for Composed Image Retrieval via Text Prompt Learning. In *Proceedings of the*

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 5753–5762.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saito, K.; Sohn, K.; Zhang, X.; Li, C.-L.; Lee, C.-Y.; Saenko, K.; and Pfister, T. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19305–19314.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565.

Shin, T.; Razeghi, Y.; au2, R. L. L. I.; Wallace, E.; and Singh, S. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. arXiv:2010.15980.

Suo, Y.; Ma, F.; Zhu, L.; and Yang, Y. 2024. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26951–26962.

Tang, Y.; Yu, J.; Gai, K.; Zhuang, J.; Xiong, G.; Gou, G.; and Wu, Q. 2025. Missing target-relevant information prediction with world model for accurate zero-shot composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24785–24795.

Tang, Y.; Yu, J.; Gai, K.; Zhuang, J.; Xiong, G.; Hu, Y.; and Wu, Q. 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, 5180–5188.

Vaze, S.; Carion, N.; and Misra, I. 2023. Genecis: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6862–6872.

Ventura, L.; Yang, A.; Schmid, C.; and Varol, G. 2024. CoVR: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5270–5279.

Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.-J.; Fei-Fei, L.; and Hays, J. 2019. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6439–6448.

Wang, L.; Ao, W.; Boddeti, V. N.; and Lim, S.-N. 2025. Generative Zero-Shot Composed Image Retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29690–29700.

Wu, H.; Gao, Y.; Guo, X.; Al-Halah, Z.; Rennie, S.; Grauman, K.; and Feris, R. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 11307–11317.

Xu, X.; Liu, Y.; Khan, S.; Khan, F.; Zuo, W.; Goh, R. S. M.; Feng, C.-M.; et al. 2024. Sentence-level Prompts Benefit Composed Image Retrieval. In *The Twelfth International Conference on Learning Representations*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; et al. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Yang, Z.; Xue, D.; Qian, S.; Dong, W.; and Xu, C. 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*, 80–90.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.