IMAGINE-THEN-REASON FOR TRAINING-FREE ZERO-SHOT COMPOSED IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

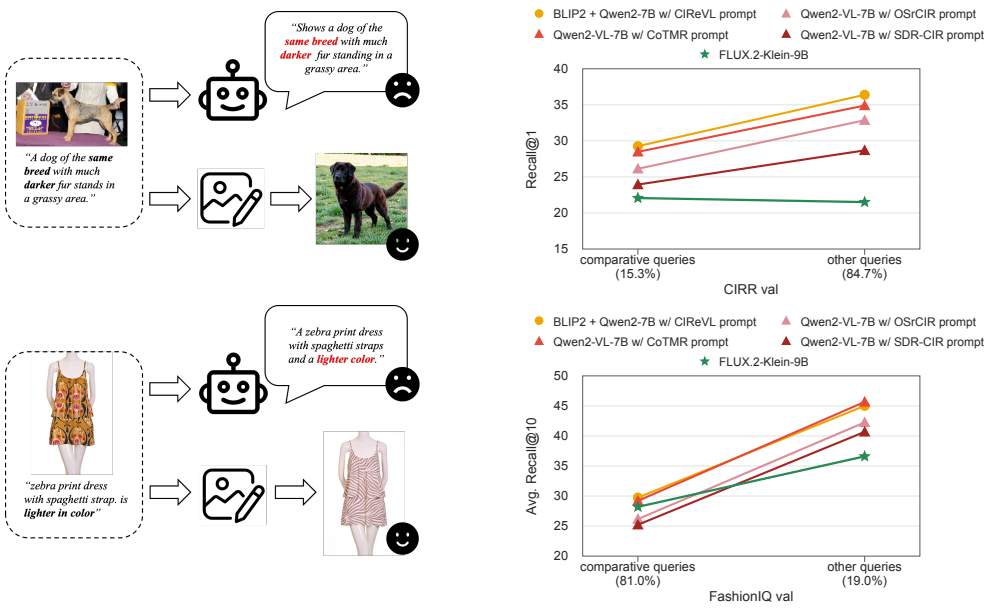
Composed image retrieval (CIR) retrieves gallery images given a reference image and a modification instruction, whose joint interpretation is a key challenge. Prior work identifies CIR as a *fuzzy retrieval* task in which the reference content to preserve, modify, or omit is ambiguous. However, our preliminary analysis of instruction types and retrieval performance highlights an underexplored source of fuzziness: *instruction underdetermination*. In particular, comparative instructions such as "make it darker" specify relative changes rather than concrete appearances, making them difficult to translate into self-contained target captions. Caption-based retrieval consistently underperforms on these queries. To address this challenge, we propose Imagine-Then-Reason for Composed Image Retrieval (ITR-CIR), a training-free zero-shot framework that (1) uses a visual target hypothesis to guide reasoning and generate concrete target captions; (2) fuses multimodal target estimators' evidence for retrieval. ITR-CIR first edits the reference image into a possible target image, then uses an MLLM to assess this hypothesis against the original query and generate grounded target captions while disregarding editing errors. The final ranking combines the captions, the synthetic image, and a direct composition of the original query. Experiments on CIRR, FashionIQ, and CIRCO demonstrate the superior performance of ITR-CIR, achieving the best training-free CIRR and FashionIQ full-gallery recall.

1 INTRODUCTION

Composed image retrieval (CIR) aims to retrieve a target image from a gallery given a reference image and a natural-language instruction specifying the desired change (Vo et al., 2019). At its core, CIR requires interpreting the instruction in the context of the reference image to infer the intended target semantics. Early approaches (Lee et al., 2021; Liu et al., 2021; Baldrati et al., 2022) learn this cross-modal composition from annotated triplets of $\langle$*reference image*, *modification text*, *target image*$\rangle$. Despite their effectiveness, these methods rely on task-specific annotations that are costly to collect at scale and may generalize poorly to unseen scenarios. To alleviate this reliance, zero-shot CIR (ZS-CIR) (Saito et al., 2023) eliminates the need for annotated CIR triplets. A common strategy trains a mapping through a proxy task to project the reference image embedding into the token space of a pretrained text encoder, where the resulting pseudo-token is combined with the modification instruction (Saito et al., 2023; Baldrati et al., 2023). However, such methods leave cross-modal composition implicit in the embedding space, which may limit their ability to infer modification intent beyond the instruction's literal wording. More recently, training-free ZS-CIR (Karthik et al., 2024) has removed the need for proxy-task training by leveraging the reasoning capabilities of frozen large language models (LLMs) or multimodal large language models (MLLMs). CIReVL (Karthik et al., 2024) adopts a two-stage pipeline that first captions the reference image and then prompts an LLM to rewrite the caption according to the modification instruction. Subsequent methods (Tang et al., 2025; Sun et al., 2025; 2026) consolidate this caption-then-reason pipeline into a single stage, using an MLLM to directly generate a target-image caption from the reference image and instruction.

Despite these advances, target-caption-based methods (Karthik et al., 2024; Tang et al., 2025; Sun et al., 2025; 2026) face a fundamental requirement: the generated caption must convey the target semantics jointly specified by the reference image and modification instruction, since it subsequently serves as a standalone query for text-to-image retrieval. This requirement is particularly challeng-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Qualitative examples.

(b) Quantitative results.

Figure 1: **Illustration of our motivation.** (a) Examples of generated target captions and synthetic target images. Comparative statements are repeated in target captions, while edited images realize them in pixel space. (b) Preliminary experimental results. For each dataset, we divide its validation queries into two subsets: queries with comparative statements (horizontal axis, left) and queries without them (horizontal axis, right). For each method, we compute retrieval metrics on these two subsets individually and compare their performance (represented by the slope of a colored line). Performance is consistently worse on comparative queries for caption retrieval, while image retrieval, represented by the green line, shows a different trend.

ing for comparative instructions, such as *"a more formal dress"* or *"an older dog"*, which specify relative changes with respect to the reference rather than an absolute target appearance. Such instructions reflect the underdetermined nature of CIR: a query can admit multiple visually different valid targets, potentially with varying degrees of relevance (Tahani, 1976; Bordogna & Pasi, 1993; Chen & Wang, 2002). Whereas prior work primarily discusses ambiguity about which reference content should be preserved, modified, or omitted (Yang et al., 2024c;b; Li et al., 2025; Lim et al., 2026), comparative instructions pose a different challenge: even when the attribute to change is clear, its desired state remains reference-dependent. A useful target caption must therefore ground the relative change in the reference image without introducing unsupported constraints. However, qualitative inspection reveals that existing methods sometimes carry comparative expressions directly into their target captions without sufficiently grounding their meaning, as illustrated in Figure 1a.

To assess how common comparative instructions are, we audit modification instructions using the semantic-aspect taxonomy introduced with CIRCO (Baldrati et al., 2023; Agnolucci et al., 2025). Comparative statements occur in an estimated 15.3% of CIRR and 81.0% of FashionIQ validation queries; official CIRCO annotations report 22.7%. They are present across all three benchmarks and especially prevalent in FashionIQ. On CIRCO, our auditor achieves an F1 score of 87.5 for this aspect, supporting the reliability of the estimates (see Appendix A for more details of the protocol and results). We then examine retrieval performance by instantiating representative target-caption pipelines with the published prompts of CIReVL (Karthik et al., 2024), OSrCIR (Tang et al., 2025), CoTMR (Sun et al., 2025), and SDR-CIR (Sun et al., 2026), using a shared retrieval backbone. Experimental details are provided in Appendix B. As shown in Figure 1b, all caption-based methods perform worse on comparative queries than on other queries across both CIRR and FashionIQ. Together with the observation that generated target captions often repeat comparative statements without consolidating them (see examples in Figure 1a), these results motivate investigating how to better ground relative modifications in target descriptions.

A promising direction is to instantiate the requested change visually before describing the target. A caption can repeat an expression such as *"more formal"* without explaining how the target should differ from the reference. By contrast, an image editor that follows the instruction must realize the change in a concrete visual candidate. This candidate represents only one possible interpretation of the query, but may expose visual attributes that help an MLLM formulate a more grounded target description. To explore this possibility, we generate synthetic target images and evaluate them through image-to-image retrieval. Figure 1b shows that, unlike caption-based retrieval, synthetic-image retrieval performs similarly across the two query groups on CIRR and exhibits a smaller gap on FashionIQ. Its disadvantage relative to caption-based methods thus narrows on comparative queries; on FashionIQ, it even outperforms some caption-based methods in this group. Although these results do not establish that image editing reliably resolves comparative semantics, they suggest that visual instantiation offers a promising source of evidence for these queries.

Nevertheless, synthetic images do not provide a consistent retrieval advantage. Incorrect edits, unsupported details, and synthetic-to-real discrepancies can make them unreliable as direct retrieval queries (Wang et al., 2026b). Existing approaches combine synthetic images with language-derived target estimates (Li et al., 2025; Wang et al., 2026a). We instead explore a different role for the synthetic image: a visual hypothesis to be examined rather than an independent target estimate to be combined. Specifically, an MLLM can inspect the candidate alongside the reference and instruction, identify which visual changes satisfy the query, and correct or disregard those that do not. This offers a way to exploit concrete visual evidence without requiring the generated image itself to be an accurate target representation.

To this end, we propose ITR-CIR (**I**magine-**T**hen-**R**eason CIR), a training-free, two-stage pipeline. The *Imagine* stage edits the reference image according to the instruction to produce a candidate target image. The *Reason* stage supplies the reference, instruction, and candidate to an MLLM to generate a correct and self-contained target caption for text-to-image retrieval. The original query remains the basis for verification: the MLLM is instructed to use query-supported visual attributes from the candidate while correcting editing errors and excluding unsupported details. Rather than simply describing the synthetic image or retrieving with it directly, ITR-CIR uses visual instantiation to support reasoning about the intended target semantics.

Our main contributions are: **(i)** We investigate the challenge posed by comparative instructions in target-caption-based CIR. A semantic-aspect audit establishes their prevalence, and a preliminary retrieval study reveals consistently weaker performance on these queries. **(ii)** We propose a two-stage pipeline in which the *Imagine* stage generates a synthetic target image as a visual hypothesis to be verified, and the *Reason* stage uses grounded evidence from the hypothesis to support target-semantic reasoning and caption generation . **(iii)** We thoroughly evaluate ITR-CIR across multiple CIR benchmarks, demonstrating its effectiveness in addressing the challenge of comparative queries.

2 Related Work

Zero-shot composed image retrieval. Early supervised methods learn a task-specific composition function from annotated triplets by combining the reference image and the modification in a learnt retrieval space (Vo et al., 2019; Baldrati et al., 2022). Although effective within the training domain, this paradigm depends on costly triplet annotations and can be difficult to transfer. Zero-shot CIR removes explicit triplet supervision by proxy learning. Pic2Word (Saito et al., 2023) learns a pseudo-word token for the reference image and composes it with the modification instruction with a text encoder. SEARLE (Baldrati et al., 2023) obtains better pseudo-word tokens through a textual inversion network. Other works address the task from different perspectives. CompoDiff (Gu et al., 2024a) models composed queries through latent diffusion trained on synthetic triplets, while FlowCIR (He et al., 2026) learns a reference-conditioned semantic transport process through flow matching. Training-free CIR further removes this auxiliary training and utilizes off-the-shelf foundation models at inference time. CIReVL (Karthik et al., 2024) establishes a caption-then-reason pipeline in which an image captioner verbalizes the reference and an LLM infers the target caption, whereas OSrCIR (Tang et al., 2025) performs the composition directly with an MLLM through Chain-of-Thought (CoT) (Wei et al., 2022) reasoning. Recent works introduce a fuzzy retrieval perspective of the task. LDRE (Yang et al., 2024c) and SEIZE (Yang et al., 2024b) identify the ambiguity and address it through sampling multiple reference captions for generating multiple

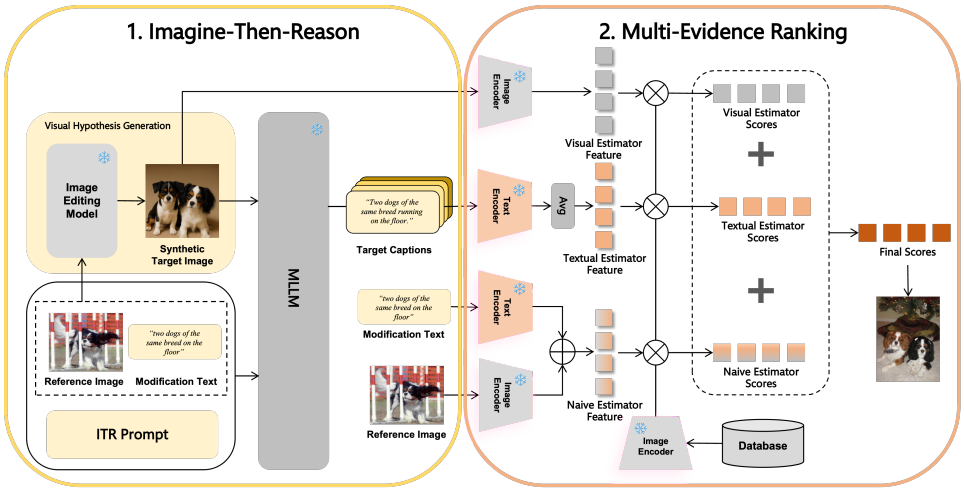


Figure 2: **Overview of ITR-CIR.** The synthetic target image conditions MLLM generation of multiple target captions. A frozen VLM then scores the gallery with a visual estimator, the mean textual estimator, and a naive target estimator, whose scores are fused for ranking.

target captions. These methods refine the language-side bias in estimating target semantics. Our work studies a complementary source of uncertainty: comparative instructions that do not define an absolute target appearance and are therefore difficult to translate into a self-contained target caption.

Synthetic visual proxies for CIR. Image generation has progressed from adversarial learning (Goodfellow et al., 2014) to diffusion models that synthesize images through iterative denoising (Ho et al., 2020). Latent diffusion moves this process into a compressed representation, making generation substantially more efficient (Rombach et al., 2022). Step distillation further reduces sampling cost by training a few-step student to approximate a longer generation trajectory (Salimans & Ho, 2022). Building on these advances, instruction-guided editors can transform a given image according to natural-language commands while preserving content outside the requested change (Brooks et al., 2023). Image editing therefore provides a natural mechanism for synthesizing the target implied by a CIR query. Generative models first enter ZS-CIR as learnt priors and sources of supervision. CompoDiff (Gu et al., 2024a) learns a latent diffusion model from the SynthTriplets18M synthetic triplet dataset to predict target visual representations. Training-free ZS-CIR subsequently introduces off-the-shelf generators at inference time. IP-CIR (Li et al., 2025) creates an imagined proxy and fuses its representation with other estimates. WISER (Wang et al., 2026a) generates a target caption and an edited image in parallel to expand the candidate set. Paracosm (Wang et al., 2026b) further constructs a mental image for each query and synthetic counterparts for the whole gallery to reduce the synthetic-to-real domain gap. However, limitations exist. First, the additional image generation substantially increases the retrieval cost in terms of computation, time and storage. Second, synthetic images are prone to generation artifacts and over-preservation of reference identity, making them a biased target estimator. To address these limitations, we try to mitigate generation cost with a step-distilled model, and we further include the edited image as a visual prior in the reasoning process to generate more convincing and visually-grounded target captions.

3 Methodology

3.1 Preliminaries

Let $\mathcal{I}$ and $\mathcal{T}$ be the pre-defined image and text spaces. Let a composed query be $q = (I_r, m)$, where $I_r \in \mathcal{I}$ is a reference image and $m \in \mathcal{T}$ is a modification instruction. Given a gallery $\mathcal{D} = \{I_1, \ldots, I_N\}$, the goal is to rank images that realize the desired modification while retaining relevant content from I_r. A retrieval system computes a gallery-level score vector

$$S : \mathcal{I} \times \mathcal{T} \times \mathcal{I}^N \to \mathbb{R}^N, \qquad S(I_r, m, \mathcal{D}) = S(q, \mathcal{D}), \tag{1}$$

and ranks the gallery by its entries. In the training-free zero-shot setting, all models introduced, including LLMs, MLLMs, image generation models, and retrieval encoders are frozen.

It is useful to describe the query by its compatible target set

$$\mathcal{Y}(q) = \{I \in \mathcal{I} : I \text{ satisfies } m \text{ relative to } I_r\}. \tag{2}$$

The set is rarely a singleton due to the fuzziness of the multimodal query. First, the instruction usually leaves some reference properties unspecified, creating *preservation uncertainty*. Second, the instruction may itself define only a relation. For an attribute $a(\cdot)$, "make it darker" states

$$a(I_t) > a(I_r) \tag{3}$$

under an appropriate darkness scale, but does not specify $a(I_t)$. We call this *instruction underdetermination*. A target-caption method may nevertheless choose one phrase, such as "a dark shirt" or "a darker shirt", which can erase the relative nature of the request or commit to an unsupported value.

This distinction does not imply that every CIR instruction is fuzzy. Direct replacement, addition, or attribute changes can often be represented by a deterministic positive caption. Comparative statements, in particular, are a source of instruction underdetermination, and are therefore the focus of our diagnostic study. The observations suggest that a target estimator should make an underdetermined instruction concrete and retrieval should avoid relying on a single target estimator. It should combine the evidence from multiple estimators and also retain that of the original query since no estimator is perfect. Our method assigns these roles to a visual target hypothesis, hypothesis-guided target captions, and a naive reference–modification estimator.

3.2 Imagine-then-Reason for Training-Free ZS-CIR

Figure 2 summarizes ITR-CIR. Generating a target image first turns an abstract modification into a visible hypothesis. Reasoning after generation then allows an MLLM to inspect that hypothesis against the reference and instruction before expressing the intended target in language. Because neither the generated image nor caption is uniformly reliable, retrieval preserves them as complementary target estimates and retains a direct estimate of the original query.

Visual target hypothesis. Inferring a target caption directly from (I_r, m) requires a language model to decide both which visual properties should be preserved and what an implicit endpoint such as "darker" should look like. The latter has no unique textual answer. We instead first ask a frozen instruction-based image editor G to instantiate one possible endpoint in the same pixel space as the reference:

$$\hat{I}_t = G(I_r, m). \tag{4}$$

This image grounds the degree, appearance, and spatial consequence of the modification, which a text-only reasoner would otherwise have to imagine internally. However, an editor can omit requested changes, over-preserve the reference, or introduce synthetic artifacts. We therefore treat $\hat{I}_t$ as a *visual target hypothesis*: a possible realization of the target that guides reasoning rather than a confident target estimator.

Hypothesis-guided target captions. Captioning only $\hat{I}_t$ would transfer its errors into the retrieval query, whereas reasoning only over (I_r, m) would discard the visual grounding just introduced. Inspired by works that utilize intermediate synthetic visualizations to complement reasoning and improve multimodal grounding (Xiao et al., 2024; Chang et al., 2026), we design a novel imagine-then-reason (ITR) pipeline for CIR. We give a frozen MLLM R all three inputs under a fixed prompt P and sample a set of K target captions:

$$\mathcal{C}_q = \{\widehat{c}_t^{(k)}\}_{k=1}^{K} = R(P; I_r, m, \hat{I}_t). \tag{5}$$

The reference identifies the content that may need to persist, the instruction defines the desired change, and the visual target hypothesis offers one candidate realization for extra context and visual grounding. The ITR prompt makes these roles explicit: the MLLM first understands the reference image and modification instruction, then verifies and utilizes the evidence provided by the visual target hypothesis to aid generation. A single caption may still omit a valid detail or overcommit to one wording of an underdetermined target. Therefore, following previous works (Yang et al., 2024b; Li et al., 2026), we sample K captions and aggregate their representations, allowing shared target semantics to remain while reducing dependence on any one generation.

Multi-evidence target estimation. The target captions provide the most explicit semantic estimate of the target, but converting an image into language can discard fine appearance (Li et al., 2025). The visual target hypothesis retains such detail, yet carries the generation artifacts and synthetic-to-real domain gap (Wang et al., 2026b). A direct composition of the reference and modification provides a third path that is faithful to the original query. We preserve these different error profiles instead of forcing one estimator to replace the others.

Let Ψ_I and Ψ_T denote the image and text encoders of a frozen contrastive vision–language model, and let $\nu(x) = x/\|x\|_2$. Every gallery image is encoded once, and the resulting features are stacked into a gallery matrix:

$$z_j = \nu(\Psi_I(I_j)), \qquad Z_{\mathcal{D}} = \begin{bmatrix} z_1^\top & \cdots & z_N^\top \end{bmatrix}^\top \in \mathbb{R}^{N \times d}, \tag{6}$$

where N is the gallery size and d is the feature dimension. We derive the visual estimator feature by directly encoding the synthetic target image, $z_v = \nu(\Psi_I(\hat{I}_t))$. For the textual estimator, each caption is encoded independently, after which its features are averaged and normalized:

$$z_t = \nu\left(\frac{1}{K}\sum_{k=1}^{K} \nu(\Psi_T(\hat{c}_t^{(k)}))\right). \tag{7}$$

Furthermore, we hope to ground the target estimators by the original query. Previous works (Lim et al., 2026; Wang et al., 2026b) add the modification instruction into retrieval to reinforce explicit modification semantics. However, following (Jang et al., 2024), we use spherical linear interpolation (SLERP) to fuse the reference and modification features to construct a more comprehensive query feature as a naive target estimator. Specifically, we obtain $z_r = \nu(\Psi_I(I_r))$ and $z_m = \nu(\Psi_T(m))$ and use SLERP with reference weight $\alpha \in [0, 1]$ and $\theta = \arccos(z_r^\top z_m)$. The naive target estimator is

$$z_n = \nu\left(\frac{\sin(\alpha\theta)}{\sin\theta} z_r + \frac{\sin((1-\alpha)\theta)}{\sin\theta} z_m\right). \tag{8}$$

For nearly parallel features, we use the corresponding normalized linear interpolation for numerical stability. Here, α controls how strongly this naive estimate preserves the reference relative to following the modification.

The three branches produce gallery-level cosine-similarity score vectors

$$S_v = Z_{\mathcal{D}} z_v, \qquad S_t = Z_{\mathcal{D}} z_t, \qquad S_n = Z_{\mathcal{D}} z_n, \tag{9}$$

where each score vector lies in $\mathbb{R}^N$. We combine their score vectors with the textual score having unit weight, while λ and $\mu \in (0, 1]$ independently control how much the visual target hypothesis and the naive target estimate affect the ranking:

$$S = S_t + \lambda S_v + \mu S_n. \tag{10}$$

In our setting, the textual estimator is most trustworthy as it is conditioned on both the query and the visual estimator. The visual estimator score is penalized since (1) image-to-image cosine similarity is generally higher than text-to-image cosine similarity; (2) image-to-image retrieval is generally weaker as shown in the preliminary experiments, and the synthetic target image is thus considered as a visual target hypothesis rather than a standalone target estimator, whose score should not dominate the final ranking. Meanwhile, the naive estimator is simple feature fusion in the retrieval space without any reasoning or verification, and its score is also penalized. This late fusion lets the reasoned captions supply reliable semantics, the visual target hypothesis recover visual detail lost in language, and the naive target estimator preserve original query evidence.

4 Experiments

4.1 Experimental Setup

Datasets. Following the evaluation protocols of previous works (Yang et al., 2024c;b), we evaluate our methodology on three standard benchmarks: FashionIQ (Wu et al., 2021), CIRR (Liu et al., 2021), and CIRCO (Baldrati et al., 2023; Agnolucci et al., 2025). FashionIQ evaluates fashion

Table 1: **Comparison on CIRR and CIRCO test sets.** ViT-B/32 and ViT-L/14 use the official OpenAI weights, while ViT-G/14 uses OpenCLIP weights. The top two training-free results are highlighted in **bold** and underline, respectively.

			CIRR						CIRCO			
			Recall@K			$\text{Recall}_{\text{subset}}$@K			mAP@K			
Arch.	Method	Train	K=1	K=5	K=10	K=1	K=2	K=3	K=5	K=10	K=25	K=50
ViT-B/32	SEARLE	✓	24.00	53.42	66.82	54.89	76.60	88.19	9.35	9.94	11.13	11.84
	CIReVL	✗	23.94	52.51	66.00	60.17	80.05	90.19	14.94	15.42	17.00	17.82
	LDRE	✗	25.69	55.13	69.04	60.53	80.65	90.70	17.96	18.32	20.21	21.11
	OSrCIR	✗	25.42	54.54	68.19	62.31	80.86	91.13	18.04	19.17	20.94	21.85
	SEIZE	✗	27.47	57.42	70.17	65.59	**84.48**	92.77	19.04	19.64	21.55	22.49
	STiTch	✗	25.83	55.25	70.20	**65.64**	83.60	**92.80**	**20.26**	**21.01**	**23.01**	**24.04**
	ITR-CIR (Ours)	✗	**30.12**	**60.15**	**72.27**	62.00	81.13	91.16	18.66	19.56	21.53	22.42
ViT-L/14	Pic2Word	✓	23.90	51.70	65.30	53.76	74.46	87.08	8.72	9.51	10.64	11.29
	SEARLE	✓	24.24	52.48	66.29	53.76	75.01	88.19	11.68	12.73	14.33	15.12
	LinCIR	✓	25.04	53.25	66.68	57.11	77.37	88.89	12.59	13.58	15.00	15.85
	CIReVL	✗	24.55	52.31	64.92	59.54	79.88	89.69	18.57	19.01	20.89	21.80
	LDRE	✗	26.53	55.57	67.54	60.43	80.31	89.90	23.35	24.03	26.44	27.50
	OSrCIR	✗	29.45	57.68	69.86	62.12	81.92	91.10	23.87	25.33	27.84	28.97
	SEIZE	✗	28.65	57.16	69.23	62.22	84.05	92.34	24.98	25.82	28.24	28.35
	STiTch	✗	28.87	57.97	69.90	**65.22**	**84.10**	**92.37**	**25.55**	26.27	**28.81**	**29.99**
	ITR-CIR (Ours)	✗	**30.12**	**59.45**	**71.30**	60.70	80.10	90.22	25.39	**26.38**	28.70	29.76
ViT-G/14	Pic2Word	✓	30.41	58.12	69.23	68.92	85.45	93.04	5.54	5.59	6.68	7.12
	SEARLE	✓	34.80	64.07	75.11	68.72	84.70	93.23	13.20	13.85	15.32	16.04
	LinCIR	✓	35.25	64.72	76.05	63.35	82.22	91.98	19.71	21.01	23.13	24.18
	CIReVL	✗	34.65	64.29	75.06	67.95	84.87	93.21	26.77	27.59	29.96	31.03
	LDRE	✗	36.15	66.39	77.25	68.82	85.66	93.76	31.12	32.24	34.95	36.03
	OSrCIR	✗	37.26	67.25	77.33	69.22	85.28	93.55	30.47	31.14	35.03	36.59
	SEIZE	✗	38.87	69.42	79.42	**74.15**	89.23	95.71	32.46	33.77	36.46	37.55
	STiTch	✗	39.23	69.95	79.56	73.56	**89.50**	**95.86**	**34.40**	**35.56**	**38.07**	**40.02**
	ITR-CIR (Ours)	✗	**40.41**	**70.51**	**80.58**	70.31	86.99	94.22	31.76	33.10	35.96	37.03

retrieval over the dress, shirt, and toptee categories; we report Recall@10 and Recall@50 for each category and their mean. CIRR contains open-domain natural images with one annotated target and a six-image subset per query. We report Recall@K for the full gallery and $\text{Recall}_{\text{subset}}$@$K$ within this subset. CIRCO is derived from COCO and annotates multiple valid targets for each query, for which we report mAP@K. Following prior comparisons, the main results use the hidden test sets of CIRR and CIRCO and the validation set of FashionIQ.

Baselines. We compare our method with several state-of-the-art baselines. We group comparison methods according to whether they require additional training. Training-based methods include Pic2Word (Saito et al., 2023), SEARLE (Baldrati et al., 2023), and LinCIR (Gu et al., 2024b). For training-free methods, we compare with CIReVL (Karthik et al., 2024), LDRE (Yang et al., 2024c), OSrCIR (Tang et al., 2025), SEIZE (Yang et al., 2024b), and STiTch (Li et al., 2026). We use published results and compare under matched retrieval backbones.

Implementation details. We use FLUX.2 [Klein] 9B (Black Forest Labs, 2025) as the image editor and Qwen2-VL-7B-Instruct (Wang et al., 2024) as the MLLM by default. We also report results with different MLLMs and image editors in Appendix D.1. The prompt templates are provided in Appendix C. For each query, following previous works (Li et al., 2026), we sample $K = 5$ target captions with temperature 0.7, top-k 50, and top-p 0.8. We set the visual score weight $\lambda = 0.3$, the naive target score weight $\mu = 0.5$, and the reference weight inside SLERP to $\alpha = 0.1$. We provide hyperparameter choices and sensitivity analysis in Appendix D.2. For the retriever, recent works have noticed possible mismatches between reported usage and actual usage of CLIP (Radford et al., 2021) in some prior works (Wang et al., 2026b; Li et al., 2026). To enable fair comparisons, following CIReVL (Karthik et al., 2024), we use CLIP-B/32 and CLIP-L/14 with the official OpenAI weights (Radford et al., 2021), and CLIP-bigG/14 with the OpenCLIP weights (Cherti et al., 2023) for CIRR and CIRCO. For FashionIQ, we use CLIP-B/32, L/14, and bigG/14 all with OpenCLIP weights. All experiments are conducted on a single NVIDIA H100 GPU.

4.2 Comparison with State-of-the-Art Methods

We run all experiments three times with different random seeds and report the mean value. Table 1 and Table 2 compare ITR-CIR with zero-shot methods under matched CLIP retrieval backbones. On CIRR, ITR-CIR attains the best training-free full-gallery Recall@K at all three backbones.

Table 2: **Comparison on the FashionIQ validation set.** All ViT-B/32, ViT-L/14, and ViT-G/14 use OpenCLIP weights. The top two training-free results are highlighted in **bold** and underline, respectively.

			Shirt		Dress		Toptee		Average	
Arch.	Method	Train	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
ViT-B/32	SEARLE	✓	24.44	41.61	18.54	39.51	25.70	46.46	22.89	42.53
	CIReVL	✗	28.36	47.84	25.29	46.36	31.21	53.85	28.29	49.35
	LDRE	✗	27.38	46.27	19.97	41.84	27.07	48.78	24.81	45.63
	OSrCIR	✗	31.16	51.13	29.35	50.37	36.51	58.71	32.34	53.40
	SEIZE	✗	29.38	47.97	25.37	46.84	32.07	54.78	28.94	49.86
	STiTch	✗	25.22	44.16	18.59	40.16	25.97	47.61	23.26	43.98
	ITR-CIR (Ours)	✗	**38.88**	**58.60**	**34.29**	**56.42**	**44.36**	**64.78**	**39.18**	**59.93**
ViT-L/14	Pic2Word	✓	26.20	43.60	20.00	40.20	27.90	47.40	24.70	43.70
	SEARLE	✓	26.89	45.58	20.48	43.13	29.32	49.97	25.56	46.23
	LinCIR	✓	29.10	46.81	20.92	42.44	28.81	50.18	26.28	46.49
	CIReVL	✗	29.49	47.40	24.79	44.76	31.36	53.65	28.55	48.57
	LDRE	✗	31.04	51.22	22.93	46.76	31.57	53.64	28.51	50.54
	OSrCIR	✗	33.17	52.03	29.70	51.81	36.92	59.27	33.26	54.37
	SEIZE	✗	33.04	53.22	30.93	50.76	35.57	58.64	33.18	54.21
	ITR-CIR (Ours)	✗	**41.10**	**59.41**	**32.97**	**56.09**	**43.43**	**63.96**	**39.17**	**59.82**
ViT-G/14	Pic2Word	✓	33.17	50.39	25.43	47.65	35.24	57.62	31.28	51.89
	SEARLE	✓	36.46	55.35	28.16	50.32	39.83	61.45	34.81	55.71
	LinCIR	✓	46.76	65.11	38.08	60.88	50.48	71.09	45.11	65.69
	CIReVL	✗	33.71	51.42	27.07	49.53	35.80	56.14	32.19	52.36
	LDRE	✗	35.94	58.58	26.11	51.12	35.42	56.67	32.49	55.46
	OSrCIR	✗	38.65	54.71	33.02	54.78	41.04	61.83	37.57	57.11
	SEIZE	✗	43.60	**65.42**	**39.61**	**61.02**	45.94	**71.12**	43.05	**65.85**
	STiTch	✗	39.48	56.59	35.04	56.74	42.86	64.95	39.12	59.43
	ITR-CIR (Ours)	✗	**45.13**	63.85	36.46	59.01	**48.16**	68.79	**43.25**	63.89

The margin is largest at the B/32 scale, where Recall@1 improves over SEIZE by 2.65 points, and it persists at L/14 and G/14. Full-gallery retrieval depends on interpreting the modification instruction itself, so this gain is consistent with our preliminary finding that a synthetic target image can help resolve the underdetermination that target captions inherit from comparative instructions. On these metrics ITR-CIR also surpasses the trained Pic2Word, SEARLE, and LinCIR at every backbone, despite requiring no training.

The improvement is more pronounced on FashionIQ, where modification instructions are dominated by comparative phrasings. With B/32, ITR-CIR exceeds the strongest training-free baseline OSrCIR by 6.84 points in average Recall@10, and the same pattern holds at L/14. At G/14, ITR-CIR remains the best training-free method at Recall@10 for shirt, toptee, and the average. This distribution of wins matches the motivation of the paper: queries whose instructions ask for a relative change benefit most from imagine-then-reason.

For CIRR subset recall, where each subset contains 6 near-duplicates of the reference image, ITR-CIR is slightly behind. This is expected since our method includes the visual estimator and naive target estimator which can contain reference identity. These estimators can reduce the model's discrimination ability on images close to the reference. Generally, as a retrieval task, we prioritize the full-gallery retrieval performance. For CIRCO, ITR-CIR trails STiTch. STiTch augments the gallery with multiple local crops, and CIRCO has multiple ground-truth labels for queries, which makes it more likely to match common target semantics among ground-truth images.

4.3 Ablation Studies and Further Analysis

Main component ablation. Table 3 isolates the contribution of each ranking component on the validation sets. The textual retrieval score S_t alone is already a strong baseline, confirming that the generated target captions carry most of the query semantics. The visual retrieval score S_v, obtained by querying the gallery with the imagined image, is clearly weaker alone, aligning with our preliminary study: generation artifacts and the synthetic-to-real gap displace its embedding from the real target. Weak standalone retrieval does not mean it does not provide any useful retrieval signals. Adding S_v to S_t substantially improves full-gallery Recall@1 on CIRR and FashionIQ average Recall@10, the largest per-component gain in the table. The gain in FashionIQ, which has the largest comparative query proportion, further validates our hypothesis that the synthetic target image can provide complementary evidence that target captions do not have. However, we observe performance drops in CIRR subset recall after adding S_v with a 3.11-point drop in $\text{Recall}_{\text{subset}}$@1, a trade-off between gallery retrieval and fine-grained subset discrimination. Furthermore, adding the

Table 3: **Main component ablation on the CIRR, FashionIQ, and CIRCO validation sets.** S_t denotes the averaged target-caption score, S_v the visual-target-hypothesis score, and S_n the naive target score.

Components			CIRR Val		FashionIQ Val		CIRCO Val	
			Recall@K	$\text{Recall}_{\text{subset}}$@K	Average		mAP@K	
S_t	S_v	S_n	K=1	K=1	R@10	R@50	K=5	K=10
✓			38.32	72.09	35.81	57.20	29.88	30.85
	✓		21.78	48.02	29.75	47.60	15.55	16.38
✓	✓		39.87	68.98	41.14	61.63	30.50	31.75
✓		✓	39.46	**72.57**	40.09	61.30	30.26	31.01
✓	✓	✓	**41.38**	70.46	**43.21**	**64.10**	**30.71**	**32.50**

Table 4: **Effectiveness of ITR.** Retrieval uses the generated target captions only on the CIRCO validation set.

	CIRCO Val			
	mAP@K			
Prompt	K=5	K=10	K=25	K=50
OSrCIR	19.99	20.96	23.29	24.17
CoTMR	22.01	23.16	25.59	26.42
SDR-CIR	13.23	13.54	15.03	15.60
Ours w/o ITR	22.76	23.80	26.49	27.32
Ours	**25.29**	**26.74**	**29.02**	**29.84**

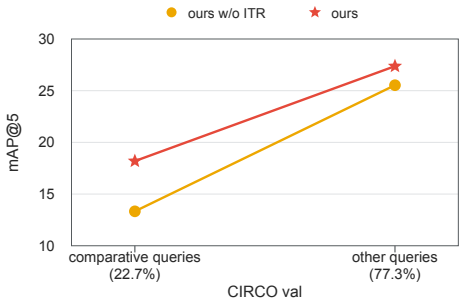


Figure 3: **Effectiveness of ITR.** Decoupled comparison on comparative and other queries.

naive target score S_n instead yields a smaller Recall@1 gain but the best subset Recall@1 of all rows, indicating that preserving original query information, especially the modification instruction, is beneficial for fine-grained subset discrimination since the latter usually states target semantics explicitly. In general, their contributions are additive: the fused score attains the best CIRR Recall@1, the best FashionIQ average Recall@10 and Recall@50, and the best CIRCO mAP@5 and 10, with some regression in CIRR subset recall.

Effectiveness of Imagine-then-Reason. To validate that the visual target hypothesis $\hat{I}_t$ indeed participates in the reasoning process and improves the target caption quality, we compare the retrieval performance of target captions generated by several one-stage methods' prompts with that of target captions generated by our two-stage Imagine-Then-Reason (ITR) method. For a strictly controlled comparison, we also include our method without $\hat{I}_t$ and instructions related to it only. The experiments are conducted with CLIP-bigG/14 on the CIRCO validation set. As shown in Table 4, our prompt without ITR already slightly outperforms the best one-stage prompt. Adding the visual target hypothesis contributes a further 2.53-point gain in mAP@5, and the full method exceeds the best one-stage prompt by 3.28 mAP@5. It shows that conditioning the language reasoning on a concrete visual realization of the instruction produces captions that better match the valid targets. Moreover, Figure 3 shows that ITR yields a larger mAP@5 gain for queries with comparative statements than for other queries, which is consistent with our overall motivation and further supports the intended role of ITR. We also provide some qualitative examples demonstrating how ITR improves target captions and thus retrieval performance in Appendix E.1.

5 Conclusion

In this work, we revisit fuzziness in composed image retrieval and distinguish uncertainty about reference preservation from underdetermination in the modification itself. Comparative instructions expose a persistent limitation of language-only target captions. Synthetic editing offers a concrete visual realization, but its errors make direct image retrieval unreliable. ITR-CIR combines the two capabilities sequentially: it imagines a visual target hypothesis, passes the hypothesis to an MLLM with the original query for generating self-contained and visually grounded target captions, and combines retrieval evidence from both modalities. This simple coupling turns a synthetic image from an unverified estimator into a visual prior for reasoning.

6 AI USE STATEMENT

In this work, we used generative AI tools mainly for editing the paper to improve readability, identifying relevant literature, and polishing writing. We also used generative AI tools to help implement methods. We have not used generative AI tools for proposing hypotheses or designing research methodology or experiments. We have reviewed all AI-assisted work. We have checked all AI-generated code and narratives, and modified or paraphrased them to avoid plagiarism and for better readability. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Lorenzo Agnolucci, Alberto Baldrati, Alberto Del Bimbo, and Marco Bertini. iSEARLE: Improving textual inversion for zero-shot composed image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):10801–10817, 2025.

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining CLIP-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Gloria Bordogna and Gabriella Pasi. 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, 1993.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Boyu Chang, Qi Wang, Xi Guo, Zhixiong Nan, Yazhou Yao, and Tianfei Zhou. AbductiveMLLM: Boosting visual abductive reasoning within MLLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Yixin Chen and James Ze Wang. A region-based fuzzy feature matching approach to content-based image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1252–1267, 2002.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. CompoDiff: Versatile composed image retrieval with latent diffusion. *Transactions on Machine Learning Research*, 2024a.

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

Zhenqi He, Ziqi Jiang, Yuanpei Liu, Yanghao Wang, Teng Wang, and Long Chen. FlowCIR: Semantic transport via flow matching for zero-shot composed image retrieval. In *Proceedings of the European Conference on Computer Vision*, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

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

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

Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, 2023.

Miaoge Li, Dongsheng Wang, Zening Sun, Jinsen Zhang, Wenhan Luo, and Jingcai Guo. STiTch: Semantic transition and transportation in collaboration for training-free zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

You Li, Fan Ma, and Yi Yang. Imagine and seek: Improving composed image retrieval with an imagined proxy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jiyoung Lim, Heejae Yang, and Jee-Hyong Lee. G-MIXER: Geodesic mixup-based implicit semantic expansion and explicit semantic re-ranking for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

OpenAI. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/, 2025.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. GPT-4o technical report. *arXiv preprint arXiv:2303.08774*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2Word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Yi Sun, Jinyu Xu, Qing Xie, Jiachen Li, Yanchun Ma, and Yongjian Liu. SDR-CIR: Semantic debias retrieval framework for training-free zero-shot composed image retrieval. In *Proceedings of the ACM Web Conference 2026*, 2026.

Zelong Sun, Dong Jing, and Zhiwu Lu. CoTMR: Chain-of-thought multi-scale reasoning for training-free zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Valiollah Tahani. A fuzzy model of document retrieval systems. *Information Processing & Management*, 12(3):177–187, 1976.

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

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval—an empirical odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Tianyue Wang, Leigang Qu, Tianyu Yang, Xiangzhao Hao, Yifan Xu, Haiyun Guo, and Jinqiao Wang. WISER: Wider search, deeper thinking, and adaptive fusion for training-free zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

Tong Wang, Yunhan Zhao, and Shu Kong. Generating a paracosm for training-free zero-shot composed image retrieval. In *Proceedings of the European Conference on Computer Vision*, 2026b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Ziyang Xiao, Dongxiang Zhang, Xiongwei Han, Xiaojin Fu, Wing Yin Yu, Tao Zhong, Sai Wu, Yuan Wang, Jianwei Yin, and Gang Chen. Enhancing LLM reasoning via vision-augmented prompting. In *Advances in Neural Information Processing Systems*, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Zhenyu Yang, Shengsheng Qian, Dizhan Xue, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Semantic editing increment benefits zero-shot composed image retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024b.

Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, and Changsheng Xu. 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*, 2024c.

APPENDIX

A DETAILS OF SEMANTIC-ASPECT AUDIT

We annotate each validation set's modification instructions in CIRR and FashionIQ with one or more of the nine semantic aspects introduced with CIRCO (Baldrati et al., 2023; Agnolucci et al., 2025) with Qwen2.5-32B-Instruct (Yang et al., 2025) as the auditor. The LLM examines each query's modification instruction only.

Semantic-aspect definitions. The auditor assigns the following aspects, following the CIRCO taxonomy:

- **cardinality**: the instruction explicitly mentions a specific number of objects, e.g., "one" or "two";
- **addition**: the instruction requires adding an object that is not present in the reference image;
- **negation**: the instruction requests removing an object present in the reference image, e.g., "without", "no";
- **direct addressing**: the instruction explicitly requests a specific change;
- **compare & change**: the instruction changes something while mentioning an attribute of the reference image, e.g., "a cat instead of a dog", "a different color";
- **comparative statement**: the instruction includes a comparison, e.g., "more", "less", "shorter", "darker";
- **conjunction**: the instruction connects statements with ",", "and", or "or";
- **spatial relations & background**: the instruction references the background or a spatial relation among objects;
- **viewpoint**: the instruction mentions a specific viewpoint or perspective, e.g., "shot from the top".

Auditing results on CIRR and FashionIQ. Table 5 reports the resulting aspect distributions. Comparative statements account for 15.3% of CIRR validation queries and 81.0% of FashionIQ validation queries, reflecting the large number of relative attributes in fashion retrieval. Instructions are overwhelmingly multi-aspect: 84.6% of CIRR and 85.7% of FashionIQ queries receive more than one aspect (2.29 and 2.00 aspects per query on average).

Validation against official CIRCO labels. CIRCO is the only benchmark of the three that releases official semantic-aspect labels, so we use it to quantify the reliability of our auditor: we re-audit all 220 CIRCO validation instructions with the identical protocol and compare the predicted labels with the official ones in a multi-label fashion (Table 6). For the comparative-statement aspect that all our analyses rely on, the auditor reaches 91.3% precision, 84.0% recall, and an F1 of 87.5 (Matthews correlation 0.84); the flag disagrees on only 12 of 220 queries. Precision remains high across the whole taxonomy (81.0–97.0%), while recall drops for the two labels, direct addressing (21.0%) and conjunction (54.9%), which the official annotation applies liberally; neither of these labels is used in our analyses.

B PRELIMINARY EXPERIMENTS

Experimental setup. For caption-based methods, we instantiate the two-stage captioner-LLM pipeline following CIReVL (Karthik et al., 2024) with BLIP-2 (Li et al., 2023) captioner and a Qwen2-7B-Instruct (Yang et al., 2024a) LLM, and the one-stage MLLM-based one with Qwen2-VL-7B-Instruct (Wang et al., 2024). The prompts used are from CIReVL, OSrCIR (Tang et al., 2025), CoTMR (Sun et al., 2025), and SDR-CIR (Sun et al., 2026). We generate one target caption for each query and perform text-to-image retrieval. For the synthetic-image-based method, we generate an edited image for each query with FLUX.2 [klein] 9B (Black Forest Labs, 2025), and conduct image-to-image retrieval. All experiments are conducted with CLIP-bigG/14 from OpenCLIP (Cherti et al., 2023) as the retriever. The results are shown in Figure 1b.

Observation 1: caption-based retrieval is substantially weaker on comparative queries. As shown in Figure 1b, for both benchmarks, the performance curves always have positive slopes for

Table 5: **Semantic-aspect distribution of CIRR, FashionIQ, and CIRCO validation sets.** "Audited" columns use our auditor; the CIRCO official column uses the labels released with the dataset. Semantic aspects are not mutually exclusive.

Aspect	CIRR audited	FashionIQ audited	CIRCO audited	CIRCO official
cardinality	22.5	1.6	16.4	16.8
addition	34.0	4.0	29.5	36.4
negation	15.9	9.9	10.5	9.5
direct addressing	36.3	4.5	13.2	54.1
compare & change	27.4	6.5	30.5	39.1
comparative statement	15.3	81.0	20.9	22.7
conjunction	15.4	91.4	42.3	74.5
spatial rel. & background	51.4	1.4	41.8	45.5
viewpoint	10.8	0.0	26.4	24.5
multi-aspect queries (%)	84.6	85.7	83.2	88.2
aspects per query	2.29	2.00	2.31	3.23

Table 6: **Auditor reliability on the CIRCO validation set.** We compute precision, recall, and F1 of the audited labels against the official CIRCO labels.

Aspect	Precision	Recall	F1
cardinality	94.4	91.9	93.2
addition	69.2	56.2	62.1
negation	87.0	95.2	90.9
direct addressing	86.2	21.0	33.8
compare & change	97.0	75.6	85.0
comparative statement	91.3	84.0	87.5
conjunction	96.8	54.9	70.0
spatial rel. & background	95.7	88.0	91.7
viewpoint	81.0	87.0	83.9
micro average	89.6	64.1	74.8
macro average	88.7	72.7	77.6

all caption-based methods from comparative queries to others, with the slope even steeper for FashionIQ. This shows that the retrieval performance is generally weaker on these queries, which could be the actual bottleneck of target-caption-based retrieval. Empirically, for this kind of instruction, a synthetic target image, generated by editing the reference image according to the instruction, can provide a more concrete instance of the target image than the caption by applying comparative changes on the reference image in the pixel space. We test this hypothesis by conducting image-to-image retrieval.

Observation 2: synthetic image retrieval has a different performance distribution from caption-based retrieval. As shown by the green curve in Figure 1b, target-image-based retrieval is even slightly better on comparative queries for CIRR, and the slope is also much flatter for FashionIQ. Although it is insufficient to claim that synthetic target images can better solve comparative queries, they indeed provide a different retrieval signal from target captions.

Observation 3: synthetic image retrieval is weaker overall. Generation artifacts, a synthetic-to-real domain gap, and over-preservation of reference identity can all move its embedding away from the real target (Wang et al., 2026b). Existing approaches therefore use the synthetic image as another target estimator and merge it with a language-derived estimate, sometimes also adapting the gallery representation (Li et al., 2025; Wang et al., 2026b).

Observation 4: one-stage methods are weaker than the two-stage method. With the controlled version and size of reasoners, one-stage MLLM-based methods fall slightly behind the two-stage captioner-LLM-based method. This may suggest that one-stage reasoning is difficult for a small 7B MLLM as it has to understand the reference image and the modification process, and formulate a target caption simultaneously, while the captioner provides textual reference context to the LLM in the two-stage pipeline. Also, the strong performance reported by MLLM-based methods is often based on larger or proprietary models like Qwen2-VL-72B-Instruct (Wang et al., 2024), GPT-4o (OpenAI et al., 2024), and GPT-4.1 (OpenAI, 2025), which hinders deployment and reproducibility.

C Additional Implementation Details

C.1 Details of the Image Editing Prompt

We use the simple prompt for the image editor following Paracosm (Wang et al., 2026b):

```
"Edit the image according to the following instruction:
<modification instruction>."
```

C.2 Details of the ITR Prompt

We provide the full prompt we use for the imagine-then-reason module in Figure 4. First, we make the following refinements over existing one-stage MLLM prompts: (1) Unlike OSrCIR's prompt, we omit the output of intermediate reasoning, which may introduce instability for small MLLMs and distort final target captions. (2) Unlike SDR-CIR's prompt, we still provide examples for in-context learning, which is crucial for small MLLMs to better understand the task and output better target captions for CIR. (3) We rewrite the target caption examples provided by OSrCIR to make them self-contained. Furthermore, we explicitly instruct the MLLM to verify and avoid including incorrect objects or attributes in the synthetic target image, while using useful visual evidence for generating target captions. Following OSrCIR, we use placeholders `<image>` for the reference image and the synthetic target image in the examples to achieve the zero-shot setting.

D Additional Ablation Studies and Analysis

D.1 Impact of MLLM and Image Editor Choice

As our method relies on the MLLM to reason with not only the query but also an additional synthetic target image, following previous works (Tang et al., 2025; Li et al., 2026), we investigate various MLLMs to study their impact on the retrieval performance. Specifically, we further test LLaVA-NeXT-7B (Liu et al., 2024) and Qwen2.5-VL-72B-Instruct (Bai et al., 2025) with the image editor fixed to FLUX.2 [klein] 9B. Table 7 reports the results on the CIRCO validation set with CLIP-bigG/14 as the retriever. As shown in the first three rows, LLaVA-NeXT-7B is weaker overall. We compare the target captions generated by it with those generated by Qwen; the former often ignores the minimality requirement of the target description and emits long captions including unrelated details in the reference image or the synthetic target image. This indicates that MLLMs' instruction following and reasoning capability are crucial for the CIR task. Moreover, Qwen2.5-VL-72B is slightly stronger than Qwen2-VL-7B, further proving the role of the MLLM, and our method's scalability to larger and more advanced MLLMs.

Meanwhile, we also study the impact of different image editing models used for generating synthetic target images as visual target hypotheses for target caption generation and as visual target estimators that contribute to the final retrieval score. We compare FLUX.2 [klein] 9B with FLUX.2 [klein] 4B and Qwen Image Edit (Wu et al., 2025) with the MLLM fixed to Qwen2-VL-7B. Note that the two FLUX.2 models are step-distilled models while Qwen Image Edit is the full model with 50 denoising steps. As shown in rows 4-6, our method is relatively robust to different image editors, indicating the role of the synthetic target image as a visual hypothesis to aid reasoning, and a fallible target estimator that should not dominate retrieval.

D.2 Analysis of hyperparameter choices and sensitivity

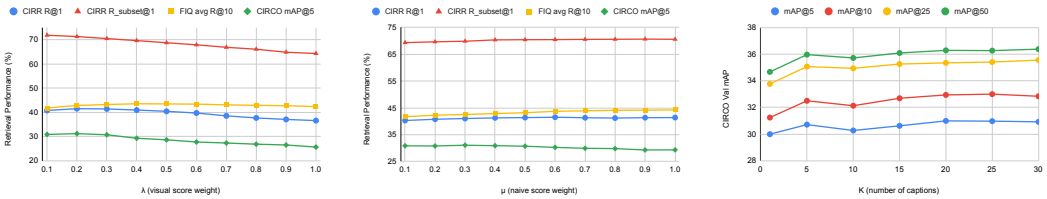
The fusion in Equation 10 introduces two scalar weights, λ and μ; we keep the SLERP reference weight fixed at $\alpha = 0.1$ throughout. Figure 5 sweeps each weight over $[0.1, 1.0]$ and the number of sampled captions K over $[1, 30]$ on the validation sets with the remaining configuration fixed, and the hyperparameters behave differently. The visual weight λ (Figure 5a) favors small values. All metrics except CIRR subset recall improve as soon as the imagined image enters the fusion and peak for λ between 0.2 and 0.3, whereas subset Recall@1 decreases monotonically with λ, the fine-grained discrimination trade-off identified above. Pushing λ further degrades every metric as the

```
I have a reference image. Given an instruction to edit the image, carefully generate a
description of the edited image.
You are also given a possible candidate image of the edited image. Use the candidate image
 to help generate the description.
Verify the candidate image against the instruction, and ignore incorrect objects or
attributes. Use the visual evidence of the candidate image.
I will put the reference image beginning with "Reference Image". The instruction I provide
 will begin with "Modification Text". The candidate image I provide will begin with "
Candidate Image".
The edited description you generate should begin with "Target Image Description". The
description you generate should only talk about the modified image and not the original
image.
Keep the caption as short as possible. Avoid adding imaginary things.

Input format:
<Input>
{
   "Reference Image": <reference image>,
   "Modification Text": <modification text>,
   "Candidate Image": <candidate image>
}

Output format:
<Response>
{
   "Target Image Description": <target image description>
}

Here are some examples for reference:
## Example 1
<Input>
{
   "Reference Image": <image>,
   "Modification Text": "Change to a large fancy white carriage, facing the opposite
direction, must include man in a black suit and hat instead of a woman",
   "Candidate Image": <image>
}
<Response>
{
   "Target Image Description": "A large fancy white carriage facing left with a man
wearing a black suit and black hat driving it."
}

## Example 2
<Input>
{
   "Reference Image": <image>,
   "Modification Text": "is dark and has more words and Desired item is blue and
references Pink Floyd",
   "Candidate Image": <image>
}
<Response>
{
   "Target Image Description": "A dark blue T-shirt with a Pink Floyd graphic and multiple
 words printed."
}

## Example 3
<Input>
{
   "Reference Image": <image>,
   "Modification Text": "has the woman and the man with the roles switched",
   "Candidate Image": <image>
}
<Response>
{
   "Target Image Description": "a woman adjusting a man's tie."
}
```

Figure 4: **Full ITR prompt to generate target captions.**

Table 7: **Impact of MLLM and image editor choice on the CIRCO validation set.** All results use CLIP-bigG/14 as the retriever.

MLLM	Image Editor	mAP@5	mAP@10	mAP@25	mAP@50
LLaVA-NeXT-7B	FLUX.2 [klein] 9B	26.82	28.73	31.39	32.33
Qwen2-VL-7B		30.71	32.50	35.07	35.97
Qwen2.5-VL-72B		31.45	32.95	35.60	36.47
Qwen2-VL-7B	FLUX.2 [klein] 4B	29.81	31.16	33.58	34.64
	FLUX.2 [klein] 9B	30.71	32.50	35.07	35.97
	Qwen Image Edit	30.24	32.19	34.81	35.68



(a) Visual score weight λ. (b) Naive score weight μ. (c) Number of sampled captions K.

Figure 5: **Sensitivity of ITR-CIR hyperparameters.**

visual target hypothesis score accounts for too large a proportion of the final score. The naive target weight μ (Figure 5b) has a common optimum: all four metrics peak near $\mu = 0.5$ with a relatively stable trend that the performance varies by less than 1.5 points. We finally test the influence of the number of sampled captions K on retrieval performance. Figure 5c shows the performance on the CIRCO validation set improves as K increases, and plateaus beyond $K = 10$, with the steepest increase occurring from $K = 1$ to $K = 5$. We therefore fix $\lambda = 0.3$, $\mu = 0.5$, and $K = 5$ as our hyperparameter settings.

E Qualitative Analysis

E.1 Successful cases

Figure 6 visualizes some successful retrieval examples of ITR-CIR on the CIRCO validation set. We deliberately pick these samples whose modification instructions have comparative statements to illustrate the effectiveness of our method. For each query, we present the query (the reference image and the modification instruction), the synthetic target image generated by FLUX.2 [klein] 9B, and the top-5 retrieval results under three retrieval pipelines: (1) "Without ITR": caption-only text-to-image retrieval without Imagine-then-Reason (ITR); (2) "With ITR": caption-only with ITR; (3) "Full Pipeline": the full ITR-CIR pipeline. These examples clearly show a progressive improvement of retrieval performance after introducing ITR and using the full method. For example, as illustrated in the first example, "Without ITR" carries the comparative statement "more organized" to the target caption, retrieving no target images in the top-5. Nevertheless, after including the synthetic target image in the reasoning process, the target caption generated correctly resolves the comparative statement to "a tidy desk", and successfully retrieves two target images in the top-5. Finally, the full pipeline, further combining multiple sources of evidence in ranking, successfully retrieves two target images in the top-2. This example clearly demonstrates and validates the effectiveness of our method's core components.

E.2 Failure cases

Figure 7 shows two failure cases of ITR-CIR on the CIRCO validation set. In the negation example, the synthetic target retains a person wearing a red tie despite the request for a different color and no people, showing a case of incorrect editing. The target caption, however, does not correctly bypass these mistakes. It has a contradictory statement "a red tie with a different color", and copies the negation directly "with no people". The latter is known to be difficult for VLMs like CLIP to model (Alhamoud et al., 2025). As a result, the top-ranked results retain these wrong attributes. In

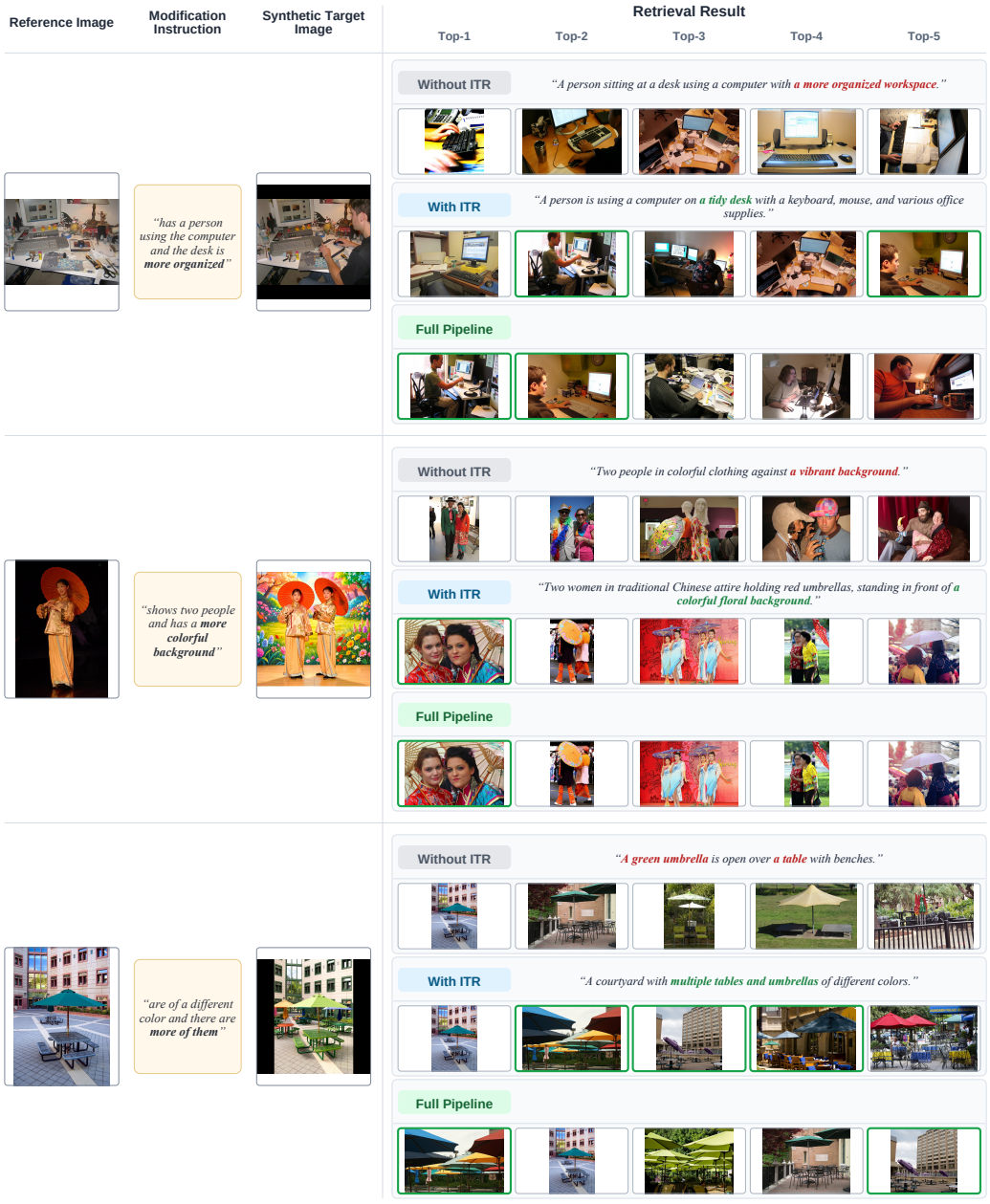


Figure 6: **Visualization of some retrieval results on the CIRCO validation set.** Each example shows the reference image, the modification instruction, the synthesized target image, and the top-5 retrieval results under three variants: caption-only without ITR, caption-only with ITR, and the full ITR-CIR pipeline. Green outlines indicate ground-truth images.

the viewpoint example, although the synthetic target correctly depicts a front view, the target caption omits this critical information. Thus, the top retrieval results do not match the required viewpoint. We attribute this to the fact that the visual estimator score only contributes a relatively small proportion in the final score in ITR-CIR, and therefore, for queries that may require more visual aid like viewpoint change, the retrieval can fail by biasing towards the textual estimator which may not correctly consolidate the required changes. These cases further demonstrate the fact that each estimator can have its own failure modes, while our method, although it combines their evidence, can still fail by biasing towards the failure of one.

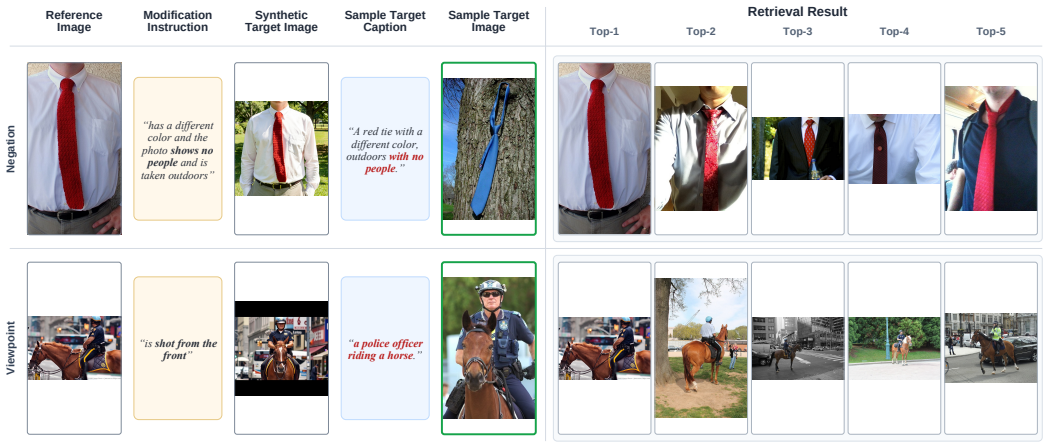


Figure 7: **Visualization of some failure cases on the CIRCO validation set.** Each row shows the reference image, modification instruction, synthetic target image, one target caption, one ground-truth target sample (green outline), and the top-5 retrieval results.

F Limitations and Future Work

Although ITR-CIR achieves promising performance, there are still limitations for future investigation. First, as the negation example in Appendix E.2 illustrates, queries with negations may require extra handling. Existing works (Sun et al., 2025; Lim et al., 2026) utilize powerful MLLMs to extract concepts that should not be present in the target image and penalize the retrieval scores of those concepts explicitly. Although effective, they rely largely on the capability of the MLLM. Therefore, a potential direction is to explore more lightweight yet reliable methods to exclude gallery images that contain negated concepts. Second, as discussed in Appendix E.2, both estimators can fail and distort retrieval results. Although our method already tries to correct potential errors introduced by the synthetic target image, it then considers the target caption, or the textual estimator, more trustworthy and fuses its retrieval score with other estimators with a fixed weight. This can bias the retrieval results towards the textual estimator. As a result, a promising future direction is to explore adaptive fusion strategies based on the accuracy of each estimator. We leave these problems as promising research directions for future work.