SETR: A Two-Stage Semantic-Enhanced Framework for Zero-Shot Composed Image Retrieval

Yuqi Xiao, Yingying Zhu*,

College of Computer Science and Software Engineering, Shenzhen University, China
zhuyy@szu.edu.cn

Abstract

Zero-shot Composed Image Retrieval (ZS-CIR) aims to retrieve a target image given a reference image and a relative text, without relying on costly triplet annotations. Existing CLIP-based methods face two core challenges: (1) union-based feature fusion indiscriminately aggregates all visual cues, carrying over irrelevant background details that dilute the intended modification, and (2) global cosine similarity from CLIP embeddings lacks the ability to resolve fine-grained semantic relations. To address these issues, we propose SETR (Semantic-enhanced Two-Stage Retrieval). In the coarse retrieval stage, SETR introduces an intersection-driven strategy that retains only the overlapping semantics between the reference image and relative text, thereby filtering out distractors inherent to union-based fusion and producing a cleaner, high-precision candidate set. In the fine-grained re-ranking stage, we adapt a pretrained multimodal LLM with LoRA to conduct binary semantic relevance judgments ("Yes/No"), which goes beyond CLIP's global feature matching by explicitly verifying relational and attribute-level consistency. Together, these two stages form a complementary pipeline: coarse retrieval narrows the candidate pool with high recall, while re-ranking ensures precise alignment with nuanced textual modifications. Experiments on CIRR, Fashion-IQ, and CIRCO show that SETR achieves new state-of-the-art performance, improving Recall@1 on CIRR by up to 15.15 points. Our results establish two-stage reasoning as a general paradigm for robust and portable ZS-CIR.

1 Introduction

Composed Image Retrieval (CIR) seeks to retrieve a target image I_t from a large gallery, given a reference image I_r and a relative text T. This task has significant applications in e-commerce, visual search, and content creation. However, traditional CIR methods require densely annotated triplets (I_r, T, I_t), whose collection is both labor-intensive and domain-specific (Baldrati et al. 2022; Wen et al. 2023; Zhang et al. 2021). To circumvent this, Zero-Shot CIR (ZS-CIR) has emerged as a promising paradigm, leveraging powerful pretrained vision-language models (VLMs) like CLIP (Radford et al. 2021) to eliminate the need for task-specific supervision.

A dominant design choice in existing ZS-CIR is the union-based fusion strategy, where the visual representation of the reference image is combined with the semantics of the relative text to construct a query (Saito et al. 2023; Li et al. 2025; Gu et al. 2024; Yang et al. 2024b; Tang et al. 2024b,a). This approach aims to maximize coverage by retaining all cues from the reference image and relative text. However, its inclusiveness is also its weakness: irrelevant background or contextual details from the reference image are indiscriminately merged, leading to semantic interference that often overwhelms the intended modification. Figure 1 highlights this limitation: Given the query "remove all but one dog and add a woman hugging it," union-based strategy latches more onto the unrelated semantic snow background rather than the directive "hugging" and fails to retrieve the correct target.

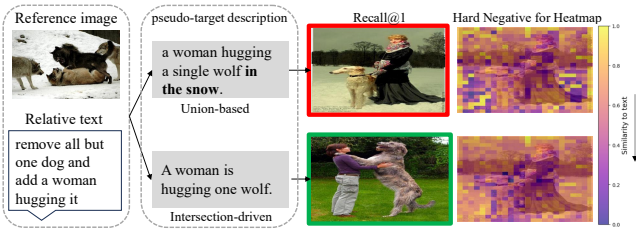

Figure 1: Comparison of union-based vs. intersection-driven strategies. We show attention heatmaps on the same hard negative image retrieved by the union-based strategy (red border). Top: the union-based strategy retains the irrelevant "in the snow" cue, ranks the wrong image at recall@1 (red border), and its similarity attention focuses on the snowy background. Bottom: our intersection-driven strategy correctly retrieves the target image (green border) with a clean pseudo-target description, and its similarity attention highlights the precise clue "hugging."

We argue that this union paradigm fundamentally misaligns with the objective of CIR. The task is not to collect all semantics from the reference image but to selectively preserve only what remains unchanged while applying the instructed modification(relative text). Motivated by this insight, we introduce the intersection paradigm for semantic fusion, which explicitly models the overlap $I_r \cap T$ between the reference image and the relative text. By filtering out the residual set $I_r \setminus T$, our approach focuses retrieval on target-relevant regions of the reference image and eliminates interference. Unlike prior models that im-

*Corresponding author.

plicitly rely on LLM-generated pseudo-descriptions (e.g., LDRE (Yang et al. 2024b)), context-dependent pseudo-words (Context-I2W (Tang et al. 2024b)), or object-centric tokenization (MOA (Li et al. 2025)), these approaches, despite their surface differences, remain anchored in the union paradigm: their objective is to integrate as much semantic content as possible from the reference and text. This inclusiveness inevitably carries over irrelevant cues, which dilute the intended modification and hinder retrieval precision. By contrast, our intersection-driven strategy pioneers a new paradigm for information fusion—it selectively retains only the overlapping, text-consistent semantics. While this may sacrifice some potentially useful information, it substantially reduces noise and enhances the authenticity of the retained cues. This trade-off yields a more faithful alignment between the composed query and the target image, and, by narrowing the candidate pool, facilitates more effective downstream re-ranking.

Yet, even with cleaner candidates, CLIP's global cosine similarity struggles to discriminate subtle variations (e.g., distinguishing "a red striped shirt" from "a red plain shirt"). To address this limitation, our SETR (Semantic-enhanced Two-Stage Retrieval) introduces a complementary second stage: an MLLM-based fine-grained scorer. Adapted via Low-Rank Adaptation (LoRA), the MLLM transforms compositional queries and candidate images into binary semantic relevance judgments ("Yes/No"), which we aggregate into structured scores for re-ranking. Unlike CLIP's coarse global matching, this scorer explicitly verifies fine-grained consistency, yielding more reliable fine-grained alignment. Our design offers both conceptual clarity and portability as a general enhancement to retrieval pipelines.

Together, these two stages form a collaborative pipeline: intersection-driven coarse retrieval narrows the search space with high-precision candidate sets, and MLLM-based re-ranking refines them with nuanced semantic reasoning. This synergy delivers consistent improvements across CIRR, CIRCO, and FashionIQ (Table 1), establishing intersection-driven fusion as a pioneering paradigm for information alignment in ZS-CIR.

Our main contributions are as follows.

- We introduce a new paradigm for compositional query construction in ZS-CIR. Unlike union-based methods that indiscriminately aggregate all cues, our intersection strategy selectively preserves only overlapping, text-consistent semantics. This reduces interference and enhances authenticity, achieving a precision-oriented trade-off.
- We adapt a pretrained multimodal LLM with LoRA to perform binary semantic relevance scorer, enabling explicit verification of relational and attribute-level consistency. This provides finer discrimination than CLIP's global cosine similarity, especially for subtle modifications (e.g., striped vs. plain).
- We integrate intersection-driven coarse retrieval with MLLM-based re-ranking into the SETR framework. The two stages are complementary: coarse retrieval narrows the candidate pool with high recall, and re-ranking refines candidates with high precision. Experiments on CIRR, CIRCO, and FashionIQ demonstrate consistent state-of-the-art performance, highlighting SETR as a general and portable paradigm for ZS-CIR.

2 Related Work

2.1 Zero-Shot CIR (ZS-CIR)

Recent advances in ZS-CIR (Saito et al. 2023; Baldrati et al. 2023; Gu et al. 2023) have eliminated the need for annotated triplets by leveraging pretrained vision–language models such as CLIP. Most approaches follow a **union-based composition strategy**, combining reference and text features to maximize semantic coverage. However, this inclusiveness inevitably imports irrelevant cues from the reference image, which dilute the intended modification and inject semantic noise into retrieval.

Despite employing different mechanisms, recent methods all share this limitation. SEIZE (Yang et al. 2024b) use LLM-generated multiple pseudo-descriptions, Context-I2W (Tang et al. 2024b) design context-dependent pseudo-words, MOA (Li et al. 2025) adopt object-centric tokenization, and CoTMR (Sun, Jing, and Lu 2025a) use CoT to reason about all possible information, each of these strategies ultimately aggregates information indiscriminately under the union paradigm, leaving semantic interference unresolved.

In contrast, our intersection-driven strategy introduces a new paradigm for information fusion: rather than integrating all cues, it selectively preserves only the overlapping, text-consistent semantics. This reduces interference and enhances the authenticity of the retained cues, establishing a cleaner foundation for fine-grained re-ranking.

2.2 Vision–Language Models and MLLMs

Vision–Language Models (VLMs) such as UNITER (Chen et al. 2020) and CLIP (Radford et al. 2021) have become central to multimodal representation learning by aligning image–text pairs in a shared embedding space. CLIP, in particular, achieves strong performance in zero-shot tasks and is widely adopted for CIR.

Beyond VLMs, the rise of Multimodal Large Language Models (MLLMs) such as GPT-4o (OpenAI et al. 2024) and Qwen2.5-VL (Bai et al. 2025) enables fine-grained cross-modal reasoning. Unlike CLIP's global feature matching, MLLMs incorporate visual inputs into a language-centric reasoning, allowing them to interpret nuanced relational semantics (Liu et al. 2024). Our method leverages this capability in a two-stage pipeline: coarse filtering stage using CLIP, followed by semantic re-ranking powered by MLLMs.

2.3 Cosine Similarity

Cosine similarity is a popular metric in CIR due to its simplicity and compatibility with contrastively trained models like CLIP. However, it assumes isotropic embeddings and cannot adequately model fine-grained semantic changes. This limitation is especially pronounced in ZS-CIR, where visually similar yet semantically incorrect images are often retrieved (Yang et al. 2024b).

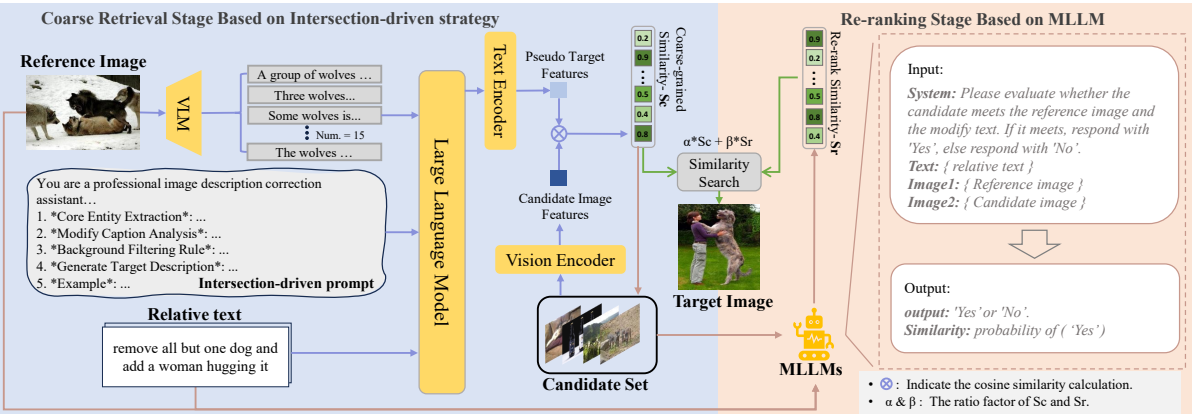

Figure 2: Overall framework of the proposed SETR for ZS-CIR.

Recent work (Sun, Jing, and Lu 2025b; Borgeaud et al. 2023) shows that replacing static similarity metrics with reasoning-capable models can alleviate this issue. MLLMs offer context-aware interpretation of image–text pairs and support dynamic understanding of compositional instructions. Motivated by this, our method integrates classical similarity for coarse candidate generation with MLLM-based semantic reasoning for refined ranking, enabling more accurate retrieval under complex transformations.

3 Method

3.1 Overview

We propose SETR, a two-stage framework. As illustrated in Figure 2, SETR consists of: (1) coarse retrieval stage employing an intersection-driven strategy to prune the candidate set, and (2) fine-grained re-ranking stage leveraging a fine-tuned MLLM scorer for semantic consistency scoring.

3.2 Intersection-driven Coarse Retrieval stage

Intersection-driven strategy. To improve the semantic alignment of the pseudo-target description, we introduce a intersection-driven strategy that identifies the core semantic intersection between the reference image I_r and the relative text T, denoted as $I_r \cap T$.[1] This retains only the visual elements from I_r that are intended to persist in the target image, minimizing semantic drift and irrelevant carry-over.

We additionally identify and discard visual elements in I_r that are not clear supported by T, represented by the set difference $I_r \setminus T$.[2] Removing these mitigates interference information and ambiguity from overdescriptive captions, thereby sharpening the retrieval focus.

To implement this semantic filtering, we design a prompt-based framework that guides to reason jointly over visual description from the reference image and modification request from the relative text. Inspired by prior work on prompt design (Zhao et al. 2021), we propose a modular engineering to ensure semantic extraction across diverse CIR scenarios. The resulting pseudo-target descriptions are used as text queries during coarse retrieval, improving matching precision.

The prompt content of the Perception enhancement step is as follows. Note that we provide the reference image and relative text in the context and refer to it accordingly in the prompt description.

Perception Enhancement Prompt Framework

```
You are a professional image
description correction assistant.
Please process the input information
according to the following rules:
```

1. `Core Entity Extraction: Identify main entity types (e.g., person, object) on the Image Description from user input:[Image Content]......`
2. `Relative Text Analysis: The modification Caption from user input is:[Instruction]......`
3. `Background Filtering: Delete the content that......`
4. `Target Description Generation: The description will be kept as simple and concise as possible. It will focus solely on the primary subject or action outlined in the Instruction.......`
5. `Example: Image Description:"The image shows a woman standing on top of a lush green field,holding a large Irish Wolfhound in her arms." Caption:"shows two dogs eating." Target Description:"Two dogs eating."`

[1] $I_r \cap T$ represents the shared visual semantic between the reference image and the relative text.

[2] $I_r \setminus T$ refers to visual content in the I_r that contradicts or is irrelevant to the relative text.

Coarse Retrieval with CLIP. Given a gallery $B = \{I_i\}_{i=1}^{M}$ and a query in the form of a pseudo-target text, we compute normalized CLIP embeddings:

$$\mathbf{v}_i = \frac{\psi_I(I_i)}{\|\psi_I(I_i)\|_2}, \quad \mathbf{t} = \frac{\psi_T(T)}{\|\psi_T(T)\|_2}, \tag{1}$$

where ψ_I and ψ_T are the CLIP image and text encoders, respectively. The cosine similarity is computed as Equation (2):

$$S_c(I_i, T) = \mathbf{v}_i^\top \mathbf{t}. \tag{2}$$

Sorting $\{S_c(I_i, T)\}$ in descending order yields a permutation σ, we select the top $K = 50$ images as Equation (3):

$$C_{\text{coarse}} = \{I_{\sigma(1)}, \ldots, I_{\sigma(K)}\}. \tag{3}$$

This step efficiently reduces the retrieval space while retaining semantically plausible candidates for re-ranking.

3.3 MLLM-based Fine-grained Re-ranking stage

To refine the coarse candidate set, we introduce an MLLM-based scorer that converts multimodal reasoning into structured retrieval scores. Instead of relying on CLIP's global cosine similarity, the scorer evaluates each candidate through binary semantic relevance judgments ("Yes/No") with respect to the composed query. This scorer explicitly verifies fine-grained semantic consistency, providing a principled scoring mechanism for fine-grained re-ranking.

To adapt the scorer to this binary judgment task, we apply lightweight LoRA adapters, fine-tuning only 1–2% of parameters while keeping the backbone frozen. Training pairs are constructed from the Visual Genome corpus(Krishna et al. 2017) by combining annotated relation triplets (subject, predicate, object) into complex region-level captions. Importantly, we do not construct retrieval triplets, and all images used in fine-tuning are strictly disjoint from our test sets. The purpose of this adjustment is not to learn retrieval supervision, but to stabilize the scoring mechanism, ensure reliable yes/no consistency judgments, and improve the output speed of scoring results through task adaptation. Thus, the retrieval stage remains strictly zero-shot, with re-ranking implemented as a portable scorer module.

At inference time, the scorer assigns a semantic-consistency score to each candidate independently. For efficiency, we default to re-ranking the top $K = 10$ candidates from the coarse retrieval stage.

Semantic Representation Construction. For each candidate image $c_i \in C_{\text{top10}}$, we construct an input P_i by encoding the visual features from c_i and the combined context of the relative text and reference image, as shown in Equation (4):

$$P_i = \Phi_{\text{vis}}(c_i) \oplus \Phi_{\text{text}}(T, I_{\text{r}}), \tag{4}$$

where Φ_{vis} and Φ_{text} are the image and text encoders of the MLLM, and $\oplus$ denotes multimodal fusion.

Consistency Scoring. The scorer is prompted to assess the relevance of each P_i and outputs probability for 'yes' options, as shown in Equation (5):

$$S_r(c_i) = \text{Probability}(\text{MLLM}_{\text{YES}}). \tag{5}$$

This score reflects the model's confidence that candidate c_i semantically matches the intended transformation.

Dynamic Re-ranking. We combine the original CLIP cosine similarity S_c and the MLLM consistency score S_r to produce a final relevance score as Equation (6):

$$L_{\text{final}} = \text{SORT}\Big(\{\alpha \cdot S_c(j) + \beta \cdot S_r(j)\}_{j=1}^{10} \cup \{S_c(j)\}_{j=11}^{50}\Big), \tag{6}$$

where α and β control the weighting between coarse similarity and re-ranking score, and $\cup$ denotes set concatenation. This late fusion strategy balances efficiency and accuracy by applying MLLM scoring only to the top-ranked candidates.

4 Experiment

4.1 Experimental Setup

Datasets: Researchers widely use CIRR(Liu et al. 2021), FashionIQ(Wu et al. 2021), and CIRCO(Baldrati et al. 2022) for compositional image retrieval, each targeting different aspects of the task. CIRR emphasizes complex real-world semantics, including attribute variations, spatial relations, and compositionality. CIRCO introduces a one-to-many retrieval challenge that requires the model to reason about semantics while filtering out irrelevant noise. FashionIQ focuses on modifying specific attributes of images in the fashion field. CIRR and CIRCO datasets focus on semantically rich natural world images, which are more complex and realistic, and also include the retrieval of fashion information to a certain extent. Therefore, we focus on evaluating the performance changes on CIRCO and CIRR datasets.

Evaluation Metrics: We evaluate retrieval performance using Recall@K (R@K) for one-to-one settings on FashionIQ and CIRR. For CIRR, we additionally report $Recall_{\text{subset}}@K$ (Rs@K) on a subset where each query is matched against six semantically similar candidates. For the one-to-many scenario in CIRCO, we use mean average precision (mAP), which captures both relevance and ranking quality.

Implementation Details: Following the findings of Yang et al., who have analyzed various configurations of VLMs and LLMs, we set the number of reference captions to 15 for optimal performance(Yang et al. 2024b). We use CLIP with ViT-L/14 and ViT-G/14 backbones as the vision–language model, and GPT-4o (OpenAI et al. 2024) as the language model, consistent with baseline settings to ensure fair comparison, and Qwen2.5-VL (Bai et al. 2025) as the MLLM for task-oriente fine-tuning. During re-ranking, we train for one epoch on a single A100 GPU and re-ranking for the top-10 coarse retrieval candidates due to computational cost.

4.2 Comparison with SOTA Methods

We compared SETR with state-of-the-art (SOTA) methods on ZS-CIR benchmark. These methods include **Pic2Word**(Saito et al. 2023), **LinCIR**(Gu et al. 2024): maps the visual features of a reference image into a pseudo-word token, **Context-I2W**(Tang et al. 2024b): selectively maps text description-relevant visual information from the reference image, **MOA**(Li et al. 2025): through object recognition and text-based object screening, the objects to be modified in the reference image are adaptively converted into

Backbone	Method	LLM-based	CIRCO				CIRR					FashionIQ
			mAP@k				Recall@k			Rs@k		R@k(avg)
			k=5	k=10	k=25	k=50	k=1	k=5	k=10	k=1	k=2	k=10
ViT-L/14	Pic2Word(CVPR'23) †	✗	8.55	9.51	10.64	11.29	23.90	51.70	65.30	53.76	74.46	24.70
	LinCIR(CVPR'24) †	✗	12.59	13.58	15.00	15.85	25.04	53.25	66.68	57.11	77.37	26.28
	Context-I2W(AAAI'24) †	✗	13.04	14.62	16.14	17.16	25.60	55.10	68.50	-	-	27.80
	MOA(SIGIR'25) †	✗	15.30	17.10	18.50	19.30	27.10	56.50	69.20	-	-	30.10
	CIReVL(ICLR'24) †	✓	18.57	19.01	20.89	21.80	24.55	52.31	64.92	59.54	79.88	28.55
	LDRE(SIGIR'24) †	✓	23.35	24.03	26.44	27.50	26.53	55.57	67.54	60.43	80.31	24.81
	OSrCIR(CVPR'25) †	✓	23.87	25.33	27.84	28.97	29.45	57.68	69.86	62.12	81.92	33.26
	SETR (Ours)	✓	**30.87**	**29.65**	**31.70**	**32.73**	**41.68**	**66.53**	**71.76**	**72.17**	**87.52**	**35.32**
ViT-G/14	Pic2Word(CVPR'23) †	✗	5.54	5.59	6.68	7.12	30.41	58.12	69.23	68.92	85.45	31.28
	LinCIR(CVPR'24) †	✗	19.71	21.01	23.12	24.18	34.80	64.07	75.11	68.72	84.70	**45.11**
	CIReVL(ICLR'24) †	✓	26.77	27.59	29.96	31.03	34.65	64.29	75.06	67.95	84.87	32.19
	LDRE(SIGIR'24) †	✓	31.12	32.24	34.95	36.03	36.15	66.39	77.25	68.82	85.66	32.49
	OSrCIR(CVPR'25) †	✓	30.47	31.14	35.03	36.59	37.26	67.25	77.33	69.22	85.28	37.57
	SETR (Ours)	✓	**37.98**	**37.72**	**40.34**	**41.37**	**44.36**	**73.13**	**79.71**	**75.66**	**89.45**	37.87

Table 1: Results on ZS-CIR benchmarks CIRCO, CIRR and FashionIQ. The top two scores are emphasized, with the highest in bold and the runner-up underlined. (†indicates the results are cited from (Tang et al. 2024a; Li et al. 2025; Yang et al. 2024a).)

pseudo-words, **OSrCIR**(Tang et al. 2024a): use the reflective thought chaining framework to infer pseudo-target descriptions by combining relative text with contextual clues in the reference image, as well as **CIReVL**(Karthik et al. 2024) and the baseline model **LDRE**(Yang et al. 2024b): the LLM-based method on the union-based strategy obtains the pseudo-target image description.

Table 1 reports evaluation results on ZS-CIR benchmarks. Our method consistently outperforms all recall metrics, achieving state-of-the-art results. On CIRR with a ViT-L/14 backbone, we improve Recall@1 by 15.15 percentage points(pp) over LDRE and 12.23 pp over the prior SOTA (OSrCIR), even surpassing the gains of ViT-G/14 on previous SOTA model. On CIRCO, we boost mAP@5 by 7.52 pp versus the baseline and 7.00 pp over the previous SOTA. These one-to-one and one-to-many improvements demonstrate the efficacy of our approach in ZS-CIR.

Our method slightly surpasses prior SOTA on FashionIQ thanks to fine-grained semantic alignment, though gains are limited by the dataset's simple attribute edits. With ViT-L/14, we match or exceed top methods; with ViT-G/14, we place second behind LinCIR, which benefits from explicit CLIP alignment training. This gap likely reflects LinCIR's specialized alignment procedure, which our framework lacks—an important limitation in the fashion domain where base CLIP's pretraining may not capture terms like "sequin corset." In contrast, on natural-image benchmarks like CIRCO, our method outperform all text-inversion techniques. Future work will explore tighter coupling between reasoning and retrieval to further improve domain-specialized performance.

4.3 Ablation Studies

To assess the contribution of each component in SETR, we conduct a series of ablation experiments on the CIRR test set, and report Recall@5 results in Table 2. All models are based on CLIP with a ViT-G/14 backbone. Minor differences between our reproduced baselines and previously reported values may stem from variations in hyperparameter tuning or training settings.

Method Variant	R@5
1. *Reported Baseline from LDRE (Yang et al. 2024b)*	66.39
2. Reproduced Baseline	62.46
Significance of key modules of SETR	
3. Baseline + Intersection-driven strategy	68.87
4. Baseline + MLLM-based Re-ranking	68.31
Impact of different MLLM backbone size	
5. w/ Qwen2.5-VL 3B backbone	69.88
6. w/ Qwen2.5-VL 7B backbone	**73.13**
7. SETR (ours)	**73.13**

Table 2: Ablation results on the CIRR test set. See section 4.3 for more details of method variants settings.

Significance of key modules. Model "3.", which omits the MLLM-based re-ranking module, suffers a 4.26 pp performance drop relative to the full model. Model "4.", which replaces our intersection-driven strategy with a union-based baseline strategy, incurs a 4.82 pp decrease—underscoring the essential contributions of both components.

Impact of MLLM Backbone Size. Replacing the Qwen2.5-VL 7B model with its smaller 3B variant (model "5.") causes a 3.25 pp decrease. Nevertheless, the 3B variant still outperforms the baseline union-based fusion model (62.46 R@5) by 10.67 pp, demonstrating that even compact MLLMs contribute meaningfully to semantic refinement, with larger models offering further gains.

This trend confirms that our method effectively suppresses irrelevant background while preserving target-relevant semantics, especially in challenging scenarios. These findings also provide more concrete and quantitative explanation for semantic benefits of intersection-driven filtering strategy.

4.4 Prompt Performance Comparison

To evaluate the Prompt Framework driven by intersection, we compare different prompt configurations in CIRR using Recall@5 (Table 3). The unstructured baseline prompt

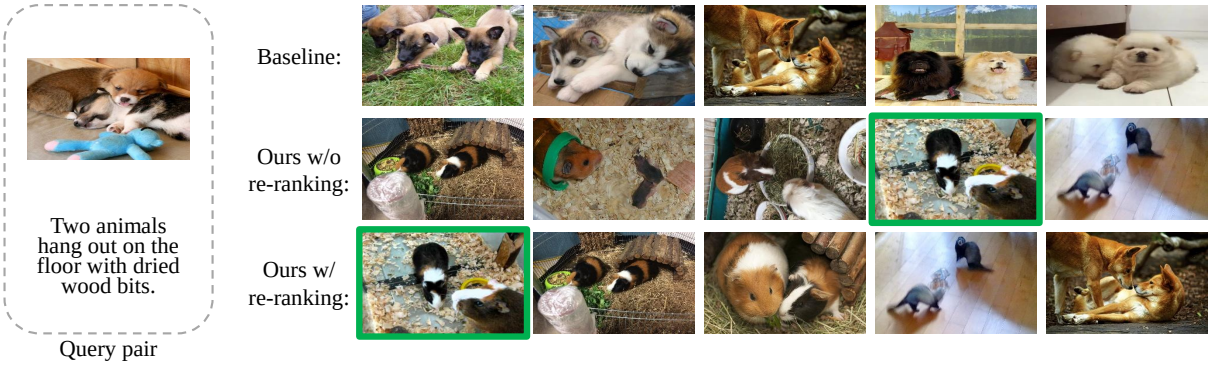

Figure 3: Examples from CIRR showing SETR's improvement over union-based baselines. Our intersection-driven retrieval eliminates background noise, and re-ranking resolves compositional phrasing to recover the correct target.

Prompt Mode	Recall@K			Rs@K	
	k=1	k=5	k=10	k=1	k=2
Union-based strategy	33.66	62.46	73.95	63.37	82.02
LLM-Generated	33.88	61.66	74.00	64.72	82.10
Instructed-Filtered	32.24	61.49	73.54	63.61	82.07
Ours	**39.23**	**68.87**	**79.71**	**72.80**	**88.08**

Table 3: Comparison of prompt modes. "LLM-Generated" denotes the template produced by the LLM when conditioned only on the reference image and its relative text to generate a pseudo-target description. "Instructed-Filtered" additionally guides the LLM to filter out irrelevant background details from the reference image.

("Union-based strategy") yields 62.46. the LLM-Generated Prompt scores 61.66 ("LLM-Generated Prompt"), and adding instructions to LLM to filter out irrelevant background details ("Instructed-Filtered Prompt") results in 61.49. By contrast, our prompts achieve 68.87, a 6.41 pp gain over baseline and 7.21–7.38 pp over LLM-Generated variants. These results confirm that our intersection-driven strategy is a important improvment in filter out irrelevant background details, and systematic prompt engineering—not merely incidental wording—substantially enhances the fidelity of pseudo-target descriptions and drives downstream retrieval performance in ZS-CIR.

4.5 Qualitative Analysis

Figure 3 illustrates SETR's two-stage collaboration on a CIRR query where the reference image contains multiple animals and the text specifies "two animals." The union-based baseline mistake retrieves dog-centric results due to interference in reference image and fails to isolate the intended pair. In contrast, SETR's intersection-driven coarse retrieval filters out inteference information, bringing the ground truth into the top-5 candidates, and its MLLM-based re-ranking then interprets the relation "on the floor with dried wood bits," correctly promoting the ground truth to top-1. This example demonstrates the effectiveness of semantic filtering, the MLLM's fine-grained relational reasoning, and the synergy between SETR's coarse retrieval and re-ranking stages.

5 Conclusion and Limitations

Despite clear gains of SETR, it still faces practical hurdles. First, the MLLM-based re-ranking module adds nontrivial latency, hindering real-time or edge deployment. Second, our hard intersection-driven filter can over-prune: subtle, occluded, or context-dependent attributes may fall outside prompt, causing occasional recall misses. Third, SETR's performance depends on prompt design—small wording or template shifts can induce subtle retrieval variance—and its zero-shot transfer to sharply different domains (e.g., medical imagery, remote sensing) remains unverified without prompt or lightweight fine-tuning.

We have presented SETR, a two-stage ZS-CIR framework that first reduces the search space via intersection-driven filtering and then refines results through MLLM-based reranking, achieving state-of-the-art performance on CIRR, FashionIQ, and CIRCO. Going forward, we will pursue lightweight reranking alternatives, develop context-aware filtering to recover nuanced semantics—all while striving to reduce inference cost and improve practical applicability.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Baldrati, A.; Agnolucci, L.; Bertini, M.; and Del Bimbo, A. 2023. Zero-Shot Composed Image Retrieval with Textual Inversion. *arXiv preprint arXiv:2303.15247*.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Bimbo, A. D. 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.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; et al. 2023. Improving language models by retrieving from trillions of tokens. *Nature*, 613(7947): 734–739.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, 104–120. Springer.

Gu, G.; Chun, S.; Jun, H.; Kang, Y.; Kim, W.; and Yun, S. 2023. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*.

Gu, G.; Chun, S.; Kim, W.; Kang, Y.; and Yun, S. 2024. Language-only efficient training of zero-shot composed image retrieval. In *CVPR*.

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

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1): 32–73.

Li, Z.; Zhang, L.; Zhang, K.; Chen, W.; Zhang, Y.; and Mao, Z. 2025. Rethinking Pseudo Word Learning in Zero-Shot Composed Image Retrieval: From an Object-Aware Perspective. In *Proceedings of the 2025 ACM SIGIR Conference*. Pittsburgh, PA, USA.

Liu, Y.; Chen, P.; Cai, J.; Jiang, X.; Hu, Y.; Yao, J.; Wang, Y.; and Xie, W. 2024. LamRA: Large Multimodal Model as Your Advanced Retrieval Assistant. arXiv:2412.01720.

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 (ICCV)*, 2125–2134.

OpenAI; Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; and others. 2024. GPT-4o System Card. arXiv:2410.21276.

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 *ICML*, 8748–8763. PMLR.

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.

Sun, Z.; Jing, D.; and Lu, Z. 2025a. CoTMR: Chain-of-Thought Multi-Scale Reasoning for Training-Free Zero-Shot Composed Image Retrieval. arXiv:2502.20826.

Sun, Z.; Jing, D.; and Lu, Z. 2025b. CoTMR: Chain-of-Thought Multi-Scale Reasoning for Training-Free Zero-Shot Composed Image Retrieval. *arXiv preprint arXiv:2502.20826*.

Tang, Y.; Qin, X.; Zhang, J.; Yu, J.; Gou, G.; Xiong, G.; Ling, Q.; Rajmohan, S.; Zhang, D.; and Wu, Q. 2024a. Reason-before-Retrieve: One-Stage Reflective Chain-of-Thoughts for Training-Free Zero-Shot Composed Image Retrieval. arXiv:2412.11077.

Tang, Y.; Yu, J.; Gai, K.; Zhuang, J.; Xiong, G.; Hu, Y.; and Wu, Q. 2024b. 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.

Wen, H.; Zhang, X.; Song, X.; Wei, Y.; and Nie, L. 2023. Target-guided composed image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, 915–923.

Wu, H.; Gao, Y.; Guo, X.; Al-Halah, Z.; Rennie, S.; Grauman, K.; and Feris, R. 2021. The Fashion IQ Dataset: Retrieving Images by Combining Side Information and Relative Natural Language Feedback. *CVPR*.

Yang, Z.; Qian, S.; Xue, D.; Wu, J.; Yang, F.; Dong, W.; and Xu, C. 2024a. Semantic Editing Increment Benefits Zero-Shot Composed Image Retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, 1245–1254. New York, NY, USA: Association for Computing Machinery.

Yang, Z.; Xue, D.; Qian, S.; Dong, W.; and Xu, C. 2024b. 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.

Zhang, G.; Wei, S.; Pang, H.; and Zhao, Y. 2021. Heterogeneous feature fusion and cross-modal alignment for composed image retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, 5353–5362.

Zhao, T. Z.; Wallace, E.; Feng, S.; Klein, D.; and Singh, S. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. arXiv:2102.09690.