Rethinking Composed Image Retrieval Evaluation: A Fine-Grained Benchmark from Image Editing

Tingyu Song [1,2,3*] **Yanzhao Zhang** [2] **Mingxin Li** [2] **Zhuoning Guo** [4,2]
Dingkun Long [2] **Pengjun Xie** [2] **Siyue Zhang** [5] **Yilun Zhao** [6] **Shu Wu** [1,3]

[1]CASIA [2]Tongyi Lab, Alibaba Group [3]UCAS
[4]HKUST(GZ) [5]NTU [6]Yale

Abstract

Composed Image Retrieval (CIR) is a pivotal and complex task in multimodal understanding. Current CIR benchmarks typically feature limited query categories and fail to capture the diverse requirements of real-world scenarios. To bridge this evaluation gap, we leverage image editing to achieve precise control over modification types and content, enabling a pipeline for synthesizing queries across a broad spectrum of categories. Using this pipeline, we construct **EDIR**, a novel fine-grained CIR benchmark. EDIR encompasses 5,000 high-quality queries structured across five main categories and fifteen subcategories. Our comprehensive evaluation of 13 multimodal embedding models reveals a significant capability gap; even state-of-the-art models (e.g., RzenEmbed and GME) struggle to perform consistently across all subcategories, highlighting the rigorous nature of our benchmark. Through comparative analysis, we further uncover inherent limitations in existing benchmarks, such as modality biases and insufficient categorical coverage. Furthermore, an in-domain training experiment demonstrates the feasibility of our benchmark. This experiment clarifies the task challenges by distinguishing between categories that are solvable with targeted data and those that expose intrinsic limitations of current model architectures.

SighingSnow/EDIR

1 Introduction

Composed Image Retrieval (CIR) aims to retrieve a target image given a query composed of a reference image and a natural language description that specifies a desired modification (Du et al., 2025; Song et al., 2025b; Wan et al., 2025). This task has attracted increasing research interest due to its broad applicability in domains such as web search and e-commerce. Consequently, a number of benchmarks have been proposed to evaluate CIR models, including CIRR (Liu et al., 2021), FashionIQ (Wu et al., 2021), and CIRCO (Baldrati et al., 2023).

Despite these contributions, current CIR benchmarks suffer from two primary drawbacks. (1) **Coarse-grained Evaluation**: Existing benchmarks (Liu et al., 2021; Wu et al., 2021) provide a coarse-grained evaluation by focusing on a narrow range of modification categories. As a result, they neglect the broader spectrum of real-world requirements, as shown in Figure 1(a). (2) **Limited Query Scale**: While some benchmarks (Baldrati et al., 2023; Wu et al., 2021) introduce query categories, they often suffer from insufficient scale and ambiguous category definitions. For instance, many queries in CIRCO are labeled with the *"direct addressing"* tag, which often overlaps with more specific categories (i.e., *"color"*), thereby diluting the granularity of the evaluation. These limitations largely stem from the methodology used to construct these datasets. Specifically, the standard approach retrieves a target image for a source image first, and then annotates a query post-hoc to describe the difference. This dependence on the retriever's output leads to the absence of certain modification categories and an insufficient number of queries for others.

To address these limitations, we first propose a comprehensive taxonomy for *fine-grained* CIR evaluation, organizing real-world requirements into five main categories and fifteen subcategories, as illustrated in Figure 1(b). We then introduce a novel data synthesis pipeline that leverages image editing (Brooks et al., 2023; Wu et al., 2025; Liu et al., 2025) to populate this taxonomy. By initiating the process with a textual modification to synthesize the target image, our pipeline provides precise control over query types and content. Using this pipeline, we construct **EDIR**, an Image **E**diting **D**erived Benchmark for Composed **I**mage

Correspondence to: Shu Wu (shu.wu@nlpr.ia.ac.cn)

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 46224–46242
July 2-7, 2026

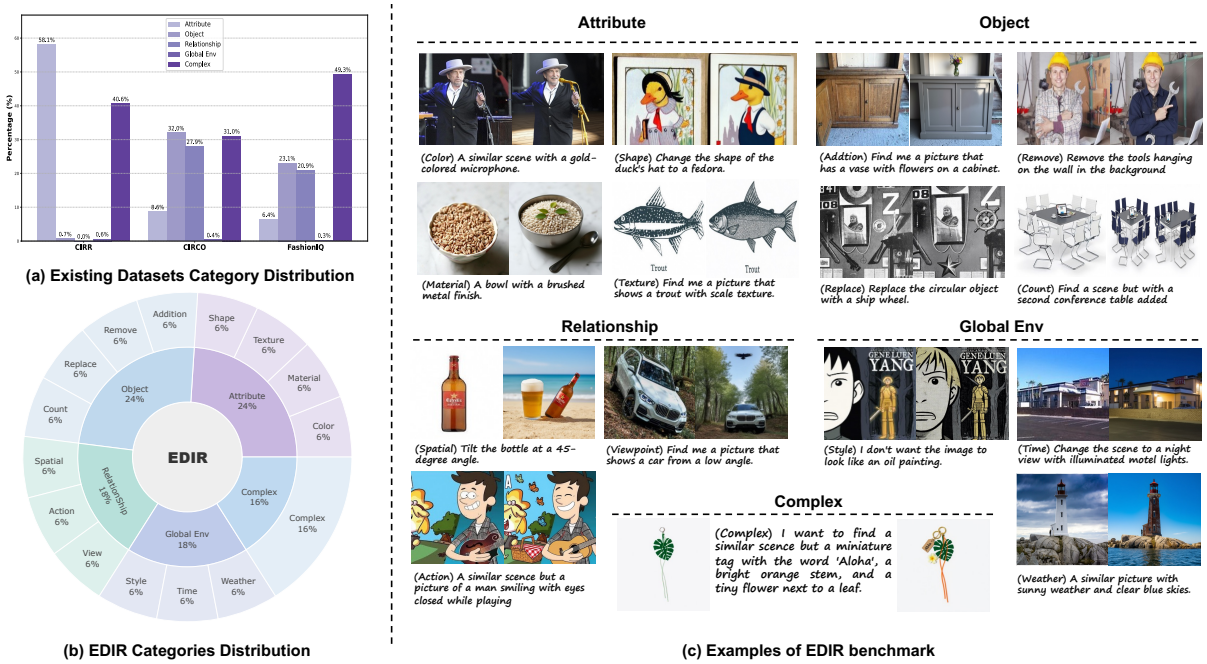

Figure 1: (a) Query category distribution in existing benchmarks, re-categorized using our taxonomy. (b) The balanced distribution of EDIR across five main categories and fifteen subcategories. (c) Example queries illustrating the fine-grained nature of each subcategory. Left is the source image, right is the target image, and the below text is the CIR query text.

Retrieval, a comprehensive CIR benchmark comprising 5,000 high-quality queries and an image gallery of 178,645 images.

We conduct an extensive evaluation of current multimodal embedding models on EDIR. Our assessment includes models trained on Multimodal Large Language Models (MLLM) (Wang et al., 2024; Bai et al., 2025) and CLIP (Hafner et al., 2021). We observe that even the strongest models cannot consistently perform well across all subcategories. We attribute these shortcomings to both the inherent limitations of the models and a scarcity of suitable training data. Additionally, we compare EDIR with existing CIR benchmarks and conduct a detailed analysis of their characteristics. We conclude that existing benchmarks suffer from significant evaluation gaps, including insufficient coverage of fine-grained categories and modality bias, which allows models to achieve high scores by over-relying on text.

Furthermore, to analyze the unique challenges presented by EDIR, we conduct an in-domain training experiment. We train a model, **EDIR-MLLM**, on our synthesized data. The resulting performance on EDIR facilitates a critical analysis, allowing us to differentiate between challenges that can be overcome with sufficient in-domain data and those that expose fundamental, intrinsic limitations of current model architectures.

Benchmark	# Qry	# Corpus	# Category
CIRR (Liu et al., 2021)	4,148	2,315	-
FashionIQ (Wu et al., 2021)	6,016	15,536	-
CIRCO (Baldrati et al., 2023)	1,000	123,403	9
I-CIR (Psomas et al., 2025)	1,813	NA[1]	7
EDIR (ours)	5,000	178,645	15

[1] I-CIR uses an instance-level corpus and does not report exact count.

Table 1: Comparing EDIR with previous benchmarks.

We summarize our contributions as follows:

- We propose a comprehensive taxonomy for CIR and a controllable data synthesis pipeline that leverages image editing to populate it.
- We introduce EDIR, a new fine-grained benchmark designed to facilitate comprehensive evaluation in the CIR domain.
- We analyze current models and existing CIR benchmarks using EDIR, revealing significant gaps in model capabilities and inherent limitations in prior benchmarks.
- We conduct an in-domain training experiment that provides insights for future model development by distinguishing between data-solvable challenges and intrinsic model weaknesses.

2 Related Works

2.1 Composed Image Retrieval Benchmarks

Recently lots of retrieval benchmarks (Meng et al., 2025; Zhang et al., 2025; Song et al., 2025a) has been proposed in multimodal and text domains. However, the evaluation of CIR models is currently constrained by a limited number of available benchmarks. Early CIR benchmarks are either domain-specific or coarse-grained in query categories. For instance, FashionIQ (Wu et al., 2021) is confined to the fashion domain, while CIRR (Liu et al., 2021) provides a broader domain but lacks fine-grained CIR queries to diagnose model failures. To address these limitations, CIRCO (Baldrati et al., 2023) introduces more detailed categories and careful annotations. However, it suffers from an imbalanced distribution of queries across categories and ambiguous category definitions. More recent benchmarks further extend CIR evaluation toward more complex reasoning settings. GeneCIS (Vaze et al., 2023) is proposed to evaluate a model's capacity to dynamically adapt its understanding of "similarity" based on a textual condition. I-CIR (Psomas et al., 2025) introduces instance-level retrieval in the CIR task. However, it only provides seven categories, which dilutes the evaluation's granularity. Furthermore, many existing benchmarks exhibit a significant modality bias (Huynh et al., 2025; Psomas et al., 2025), where models can achieve high scores by over-relying on text-only signals, failing to test genuine multimodal compositionality.

2.2 Composed Image Retrieval Methods

Methods for CIR have evolved from specialized attribute classifiers (Ak et al., 2018; Yang et al., 2020; Hou et al., 2021) to approaches built upon VLMs (Hafner et al., 2021; Li et al., 2022, 2023). Current literature generally categorizes CIR approaches into three main streams: (1) *Text-Inversion* (Saito et al., 2023; Baldrati et al., 2023; Gu et al., 2024a), which maps the reference image to a textual token for text-image fusion; (2) *Data Synthesis* (Ventura et al., 2024; Zhang et al.; Zhou et al., 2024, 2025), which leverages generative models to create large-scale CIR triplets for training (though recent works (Zhou et al., 2024; Gu et al., 2024b) use this for training data, they do not address the benchmark evaluation gaps we identify); and (3) *Training-Free Methods* (Karthik et al., 2024; Wu et al., 2024; Yang et al., 2024), which leverage modular pipelines combining off-the-shelf vision and language models for tasks such as captioning and zero-shot reasoning.

Recently, MLLMs (Google, 2024; Bai et al., 2025; Zhu et al., 2025) have shown strong performance on a wide range of tasks. Consequently, universal multimodal embedding models (Lin et al.; Jiang et al., 2024b) have been developed based on MLLM architectures. These MLLM-based models achieve superior performance on existing CIR benchmarks. However, given the restricted benchmark design and limited evaluation coverage mentioned above, it remains unclear whether the observed performance gains reflect genuine compositional reasoning or merely the exploitation of benchmark biases.

3 EDIR Benchmark Construction

We propose EDIR, a comprehensive benchmark designed to evaluate the capabilities of current multimodal embedding models in CIR in a fine-grained manner. Formally, each instance in our benchmark is a triplet $\{I_r, T_m, I_t\}$, comprising a reference image I_r, a target image I_t, and a text query T_m. Constructing such a benchmark at scale presents two primary technical challenges that are not adequately addressed by existing CIR benchmarks. First, we must ensure *fine-grained, category-level diversity* to guarantee that the benchmark reflects the broad range of real-world CIR needs rather than a narrow set of edits. Second, we require a *systematic and scalable* method to construct the text query T_m for each category, ensuring that the modification is unambiguous, controllable, and aligned with the intended evaluation signal.

To address these challenges, we first propose a comprehensive and hierarchical taxonomy that covers a wide spectrum of real-world modifications, as illustrated in Figure 1. Then, to systematically construct queries for these categories, we build an automated pipeline based on state-of-the-art image editing technology. This pipeline leverages our taxonomy to guide the generation process, ensuring that each resulting triplet is accurately aligned with a specific, fine-grained category. Specifically, we first generate an initial triplet $\{I_r, T_{edit}, I_t\}$, where a source image I_r is modified to produce a target image I_t based on a raw edit instruction T_{edit} (§3.2). Next, we refine the edit instruction T_{edit} into a natural CIR query T_m(§3.3). Finally, we apply a two-stage filtering process and human validation to ensure overall dataset quality (§3.4, §3.5).

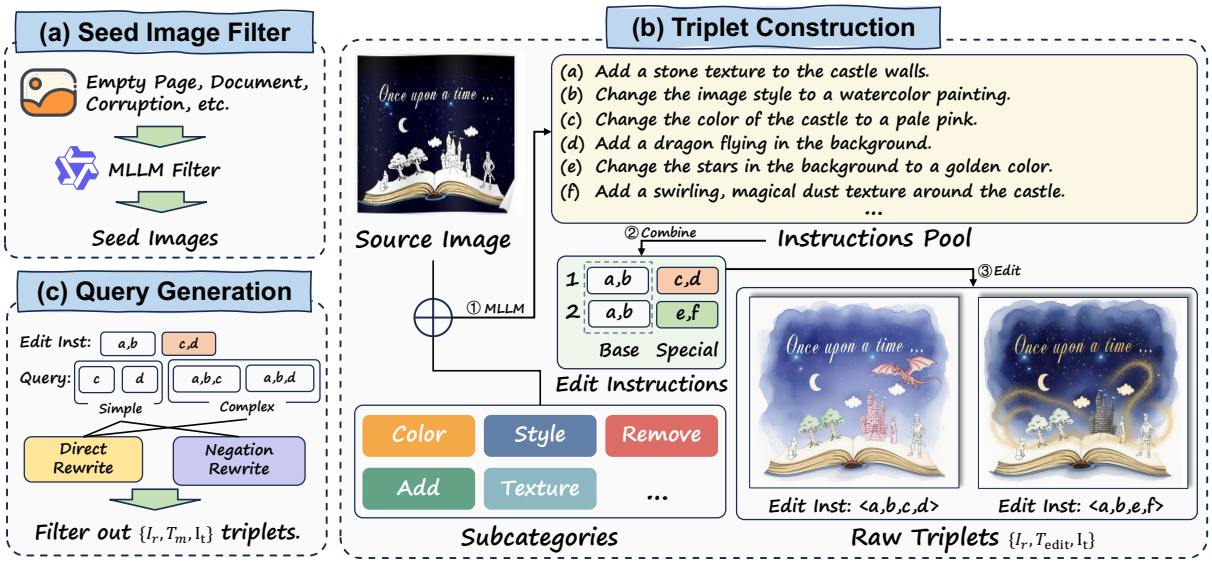

Figure 2: Overview of our data synthesis pipeline: (1) Seed Image Selection: Unsuitable images are filtered from a large pool to select high-quality source images. (2) Triplet Generation: For each source image, multiple edit instructions are generated and applied to create ⟨source image, edit instruction, target image⟩ triplets. (3) Query Formulation: The edit instructions are automatically rewritten into natural language CIR queries.

3.1 Preliminary Setup

Taxonomy Definition. Our taxonomy consists of five major categories: *Attribute*, *Object*, *Relationship*, *Global Environment*, and *Complex*. (1) The *Attribute* category focuses on object properties, mirroring e-commerce scenarios where a user might ask to see a product in a different color or material; (2) *Object* involves operations such as adding or removing objects, which is fundamental for practical applications like photo editing and content creation; (3) *Relationship* pertains to the spatial or semantic connections between objects, reflecting sophisticated needs like rearranging a scene for interior design or changing a viewpoint; (4) *Global Environment* addresses holistic changes to the scene, such as style or weather, supporting creative searches for different moods or artistic effects; (5) *Complex* queries combine multiple modifications from the other categories, representing the most realistic and challenging user requests that involve several simultaneous constraints. As detailed in Figure 1 and Table 2, these five categories are further broken down into fifteen subcategories.

Seed Image Selection. We select seed images from the LAION-400M (Schuhmann et al., 2021) dataset due to its vast coverage of real-world scenes. However, the dataset contains numerous corrupted, blank, or document-style images that are unsuitable for both CIR and image editing. To address this, we employ an MLLM (i.e., Qwen2.5VL-32B) to automatically filter out these low-quality images. The specific prompt used for this filtering step is detailed in Appendix C.

Subcategory	Definition
Category 1: Attribute	
(1) Color	Changes the color of an object or an entire area.
(2) Material	Modifies the surface material of an object.
(3) Shape	Alters the geometric form or outline of an object.
(4) Texture	Adds or alters fine surface details, patterns on an object.
Category 2: Object	
(5) Addition	Introduces a new, distinct object into the scene.
(6) Remove	Completely eliminates an existing object or element.
(7) Replace	Swaps an existing object with another object.
(8) Count	Changes the number of instances of a specific object.
Category 3: Relationship	
(9) Spatial	Modifies the position, orientation, or background of elements,spatial relationships between items.
(10) Action	Makes a subject in the image perform a new action.
(11) Viewpoint	Alters the camera's position, angle, or in-door and out-door transform.
Category 4: Global Environment	
(12) Style	Changes the artistic style of the image.
(13) Time	Changes the time of day depicted in the scene.
(14) Weather	Modifies the weather conditions shown in the image.
Category 5: Complex	
(15) Complex	A query that combines two or more modifications from the simple categories above.

Table 2: Detailed categories definitions.

3.2 Raw Triplet Construction

For each source image I_r, we first use an MLLM (i.e., Qwen2.5-VL-32B) to identify 5-6 suitable subcategories from our taxonomy. For each suitable subcategory, the MLLM generates three distinct edit instructions, creating an instruction pool.

Our core strategy leverages this pool to synthesize a set of related, complex images $\{I_1, I_2, ...I_n\}$ using an image editing model (i.e., Qwen-Image-Edit (Wu et al., 2025)). Within this set, one image is randomly selected to serve as the *target image* I_t for a given query, while the others serve as *hard negatives*. To achieve this, each image I_i is generated by applying a composite of instructions $\{a, b, c, d\}$. One part consists of base modifications $\{a, b\}$ sampled from different categories; these establish a shared visual context crucial for our hard negative mining strategy. The second part consists of distinctive modifications $\{c, d\}$, which are randomly sampled to prevent the retrieval task from becoming trivial. If a target image were generated with only one unique change, the retrieval task would be overly simple for current CIR models. The full process is illustrated in Figure 2(b).

3.3 Query Rewrite

Since the raw edit instructions are not directly suitable for use as CIR queries, we need to further refine the edit instruction into a natural CIR query. As established in §3.2, each editing process is guided by a composite instruction $\{a, b, c, d\}$. The instructions a and b are basic operations that create a shared visual context, while c and d are the distinctive modifications. For simple queries, we use one of the distinctive modifications, c or d, as the basis. For complex queries, we combine one distinctive modification with the two basic operations, resulting in a query based on $\{a, b, c\}$ or $\{a, b, d\}$, as illustrated in Figure 2(c). We avoid using the full set of instructions $\{a, b, c, d\}$ for a single query, as this would make the query overly specific and could negatively impact retrieval performance. We utilize an LLM (i.e., Qwen3-32B) to rewrite the edit instruction. Following previous work (Zhou et al., 2025; Huynh et al., 2025), we employ several prompt templates to rephrase the edit instructions into natural language queries. In addition to direct rewrites, we recognize the importance of negation queries. For example, a positive query might be "I want the same dress but in red," whereas a corresponding negation query could be "Show me this dress in a different color." These types of queries are common in daily life, especially for categories such as *Color* and *Shape*. Therefore, we intentionally construct negation-based queries for these specific categories.

3.4 Data Quality Control

To improve data quality, we implement a two-stage filtering pipeline using an MLLM (i.e., QwenVL-32B). The first stage occurs after the raw triplet construction (Figure 2(b)). The MLLM assesses whether the generated image matches the full composite edit instruction, filtering out 312,009 of the 368,437 initial images. However, the complexity of these instructions occasionally allows partially correct images to pass. Therefore, a second filtering stage is applied after query rewriting (Figure 2(c)). In this step, the MLLM re-evaluates the $\{I_r, I_t\}$ pair against the more concise CIR query T_m. This second pass filtered out 889,013 from 1,087,710 triplets, improving the final dataset's alignment with the queries. We provide details of the construction process in Appendix A.1 and prompts in Appendix C.

3.5 Dataset Analysis

Statistics. We initially sample 70,000 images and generate 368,437 edited images. After filtering, 889,013 high-quality $\{I_r, T_m, I_t\}$ triplets are obtained. From these, we construct our benchmark by randomly sampling 300 queries for each of the 14 simple categories and 800 queries for the *Complex* category, resulting in a total of 5,000 queries. For each query, we include its target image along with three hard negatives generated from the same source image. To ensure corpus diversity, this set is augmented with 150,000 additional edited images which are also derived from the 70,000 source images. The final benchmark comprises 5,000 queries and a corpus of 178,645 images.

Human Validation. To assess dataset quality, we conduct a human validation study on a randomly selected 12% sample. In this study, annotators evaluate three primary error types. The *False Positive Rate* measures instances where the target image I_t does not match the query $\{I_r, T_m\}$. The *False Negative Rate* identifies cases where a provided hard negative image also satisfies the query. Finally, to measure the *Global False Negative Rate*, we use a state-of-the-art CIR model (Zhou et al., 2025), MMRet-MLLM, to retrieve the top-5 images from the corpus for each query $\{I_r, T_m\}$. Annotators then check if any of these retrieved images, other than the target image I_t, are also positive. Our manual annotation reveals a False Positive Rate of 8.0%, a False Hard Negative Rate of 7.3%, and a Global False Negative Rate of 11.7%.

Metric	Total	Attribute				Object				Relationship			Style			Complex
		Color	Material	Shape	Texture	Add	Remove	Replace	Count	Spatial	Action	View	Style	Weather	Time	Complex
Non MLLM-based Models																
PIC2WORD	**21.2**	**22.0**	**15.3**	18.0	16.7	**28.7**	11.7	24.3	**23.3**	**19.0**	27.7	12.0	21.3	**26.7**	**29.3**	**21.8**
SEARLE	17.1	20.7	15.0	15.7	12.0	25.0	**8.3**	21.3	20.0	10.7	**28.7**	**8.0**	10.0	22.0	20.0	18.1
MAGICLENS	16.8	19.3	10.3	**9.7**	**6.7**	22.0	12.0	**29.3**	18.3	18.0	24.3	10.0	**23.3**	11.0	18.7	17.4
Avg.	18.4	20.7	13.6	14.4	11.8	25.2	10.7	25.0	20.6	15.9	26.9	10.0	18.2	19.9	22.7	19.1
MLLM-based Models																
RzenEmbed-7B	**47.2**	44.7	37.3	35.7	36.7	74.0	**28.0**	**71.0**	49.0	**45.7**	**60.7**	24.0	44.3	**50.7**	**58.0**	47.5
Ops-embedding	**47.2**	**45.7**	**38.3**	**38.7**	**40.0**	**75.0**	23.3	66.3	**50.3**	44.0	59.7	**24.7**	**46.3**	**50.7**	52.3	**49.0**
GME-2B	42.4	39.0	35.3	34.3	37.3	65.0	21.7	63.3	47.3	34.0	56.3	22.0	42.0	44.7	50.3	42.8
GME-7B	40.1	36.7	34.3	30.7	29.7	63.3	23.3	62.7	45.3	36.7	52.0	24.3	35.7	42.0	48.0	39.0
MMRet-MLLM	36.8	36.3	26.3	26.7	27.3	57.7	18.7	54.3	40.3	36.7	40.7	20.7	23.0	42.7	45.0	43.6
E5-V	34.0	26.3	27.3	26.0	27.3	55.0	14.7	51.0	39.3	37.0	49.0	11.7	35.7	34.3	38.7	34.9
VLM2Vec-2B	32.4	32.7	25.0	26.7	32.3	55.7	18.7	33.7	36.0	32.7	35.3	18.0	23.7	34.7	38.7	36.0
UniME-7B	31.8	23.7	16.0	23.3	20.3	48.7	17.0	46.0	41.0	35.3	46.3	17.7	29.0	27.7	35.7	38.5
UniME-2B	28.6	28.0	23.0	21.0	21.0	44.3	17.3	49.7	33.7	22.7	40.7	14.7	21.7	23.3	30.3	31.8
mmE5	28.1	20.7	21.0	23.0	24.7	41.0	19.3	39.7	26.3	28.7	36.7	20.3	30.3	31.3	31.0	28.1
Avg.	36.9	33.4	28.4	28.6	29.7	58.0	20.2	53.8	40.9	35.3	47.7	19.8	33.2	38.2	42.8	39.1
EDIR-MLLM	59.9	57.7	59.0	44.0	56.3	86.0	37.7	74.3	58.3	48.0	71.0	33.0	66.7	76.3	72.0	59.1

Table 3: Recall@1 performance of models on EDIR. *Avg.* is computed as the average performance across categories for each type of models, excluding EDIR-MLLM.

4 Experiments and Results

4.1 Experiment Setup

We evaluate a wide range of multimodal models using Recall@1, including both MLLM-based and Non-MLLM-based types, as follows:

Non-MLLM-based Models. We evaluate the following Non-MLLM-based methods and models: (1) **PIC2WORD** (Saito et al., 2023), which implements the *text-inversion method* for CIR. (2) **SEARLE** (Baldrati et al., 2023), which is also based on the *text-inversion method*. (3) **MAGICLENS** (Zhang et al.), which is trained on a large scale of CIR triplets.

MLLM-based Models. We evaluate the following frontier MLLM-based models: (1) **GME-Qwen2-VL** (Zhang et al., 2024), for which we include both the 2B and 7B versions. (2) **BGE-VL** (Zhou et al., 2025), where we use the BGE-VL-MLLM-S1 version, which is not further finetuned on MMEB. (3) **VLM2Vec** (Jiang et al., 2024b), where we use VLM2Vec-V2.0, which is based on Qwen2-VL-2B. (4) **Ops-embedding** (OpenSearch-AI, 2025), where we use Ops-MM-embedding-v1-7B for evaluation, which shows competitive performance on relevant tasks. (5) **E5-V** (Jiang et al., 2024a), where we use the 7B version of E5-V. (6) **UniME** (Gu et al., 2025), where we use UniME-Qwen2-VL and include both the 2B and 7B versions. (7) **mmE5** (Chen et al., 2025), where we use the model trained based on Llama-3.2-11B-Vision. (8) **RzenEmbed-7B** (Jian et al., 2025), where we use the RzenEmbed-V2-7B based on Qwen2-VL.

4.2 Results

Non-MLLM-based Models. Non-MLLM-based models achieve an average total score of only 18.4%. We attribute this underperformance primarily to the limitations of the CLIP architecture upon which these models are built. Since many candidate images in EDIR are visually similar edits of a single source, these models can identify the correct group of images but cannot accurately distinguish the target based on the fine-grained text query. This fundamental limitation explains their low scores in nuanced categories like *remove* and *texture*. This confirms that EDIR is also a challenging benchmark for Non-MLLM-based models.

MLLM-based Models. From Table 3, we observe that MLLM-based models consistently outperform Non-MLLM baselines. They achieve relatively strong performance on the *addition*, *replace*, and *action* categories. However, they perform notably worse on others, especially *texture*, *remove*, and *shape*. We therefore conduct a detailed error analysis of these models (§4.3). In addition, to verify that EDIR is a meaningful and complementary benchmark, we compare model performance on EDIR against existing CIR benchmarks and provide a thorough analysis (§4.4).

4.3 Error Analysis

To better understand the current weaknesses of multimodal embedding models, we examine cases with low Recall@1 scores and develop the following taxonomy of error types. (1) **Failure in Handling Negation**: Models consistently struggle with queries involving negation, both in removal commands (e.g., "remove the hat") and with explicit negative terms (e.g., "not red"). (2) **Deficiencies in Compositional Reasoning**: Models exhibit poor performance on categories like *count*, *spatial*, *style*, and *viewpoint*. These tasks demand a form of compositional reasoning. The model must correctly interpret relationships between objects (*spatial*, *count*) or apply global transformations that affect the entire scene (*style*, *viewpoint*). For instance, executing a *viewpoint* query to change an indoor scene to an outdoor one requires the model to reason about the global scene context and its constituent elements. This is a capability that current models appear to lack. (3) **Struggles with Multiple Constraints**: In the *complex* category, queries provide multiple conditions. Models often retrieve images that only partially satisfy all constraints. This indicates a weakness in composing and verifying multiple distinct instructions from a single query. (4) **Insensitivity to Fine-Grained Details**: For categories such as *texture*, *material*, and *shape*, the distinctions between the source and target images can be subtle. Current models tend to overlook these fine-grained visual changes, leading to errors. We provide a detailed error case study in the Appendix A.2.

This underperformance stems from two interconnected issues: intrinsic model weaknesses and inadequate training data. Weaknesses in the foundational MLLMs (Fu et al., 2025a,b) explain the observed **Failure in Handling Negation** and **Deficiencies in Compositional Reasoning**, as these base models inherently struggle with logical and spatial operations. Simultaneously, the embedding models' inability to handle **Multiple Constraints** and **Fine-Grained Details** is exacerbated by training on data that lacks such complexity. This highlights the critical need for more carefully curated datasets to address these specific shortcomings and enhance model capabilities.

4.4 Benchmark Analysis

To better understand the limitations of existing CIR benchmarks, we analyze the performance correlation of MLLM-based models across EDIR and four prominent CIR benchmarks: CIRCO, CIRR, FashionIQ, and GeneCIS. We compute the Spearman correlation coefficients between model performances. Performance on each benchmark is measured using its respective standard metric, including Recall@1 for EDIR. For CIRR and CIRCO, we use their validation sets to measure performance. As shown in Figure 3, EDIR has a positive value between all categories and the target models. This verifies that EDIR is qualified to evaluate the CIR abilities of current models. However, the results also reveal varying correlations. This confirms the two critical limitations of existing benchmarks mentioned in §2.1: a fine-grained evaluation bias and a significant modality bias.

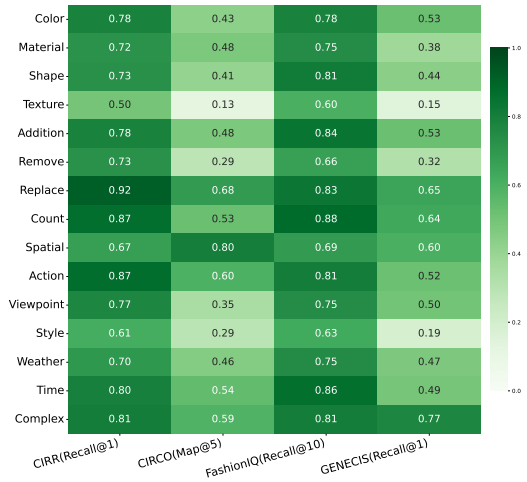

Figure 3: Performance correlation of MLLM-based models between EDIR and prior CIR benchmarks.

Fine-grained Evaluation Bias. Existing benchmarks lack balanced, fine-grained evaluation. Using an LLM (i.e., Qwen-32B) to classify their queries, we find a heavy skew towards *complex* modifications, as shown in Figure 1. Meanwhile, they lack sufficient coverage of specific categories like *remove*, *spatial*, and *texture*. For example, CIRCO only has 10 *remove* queries, and CIRR has no *spatial* queries in its validation set. This overall categorical imbalance helps explain why, in our correlation analysis illustrated in Figure 3, the performance correlation for these specific abilities is consistently lower relative to other categories within the same benchmark's results. This indicates that EDIR addresses a critical evaluation gap by providing comprehensive coverage of these overlooked compositional skills.

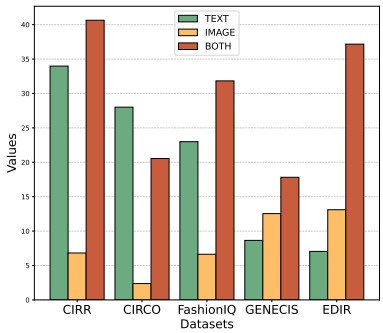

Figure 4: Average performance of MLLM-based models across CIR benchmarks.

Modality Bias. Existing benchmarks can also exhibit a strong modality bias. We test this by evaluating MLLM-based models in *text-only*, *image-only*, and *text-image* modes. As illustrated in Figure 4, on CIRCO, models perform even better with only text, indicating the reference image is almost redundant. This text-centric shortcut also partially explains CIRCO's low correlation with EDIR, which requires a genuine synthesis of both modalities and thus offers a more robust test of the CIR task. In conclusion, EDIR provides a more fine-grained evaluation that simultaneously demands a compositional understanding of both image and text.

5 In-domain Training and Analysis

To further investigate the unique challenges posed by EDIR and its relationship with existing benchmarks, we conduct an in-domain training experiment. This experiment is designed to assess the solvability of EDIR's fine-grained categories when a model is trained on specialized data. Leveraging our data synthesis pipeline Figure 2, we generated an additional pool of approximately 1.1 million high-quality edit triplets. From this pool, we curate a specialized training set by sampling 15,000 triplets for each of our 15 categories, totaling 225,000 training instances. We train a model based on Qwen2.5-VL (Bai et al., 2025), which we refer to as EDIR-MLLM, on this dataset for 2,500 steps with a batch size of 128. We provide training details in Appendix B.2.

To determine if the challenges in EDIR are solvable and to identify which categories remain difficult, we define a category as *solvable* if its Recall@1 exceeds 60% or shows an improvement of over 20 percentage points after in-domain training. The in-domain performance of EDIR-MLLM demonstrates that our benchmark is indeed solvable. As shown in Table 3, EDIR-MLLM achieves a new state-of-the-art Recall@1 of 59.9% on EDIR. This is a substantial improvement over the average of other MLLM-based methods, which is 36.9%. To gain a more granular understanding of these results, we analyze the performance on a per-category basis. These results directly corroborate our model analysis in §4.3. As mentioned, categories requiring sensitivity to fine-grained details, such as *color*, *material*, *texture*, and *action*, see dramatic improvements. This confirms our hypothesis that such challenges, often stemming from inadequate training data, can be largely overcome with corresponding examples. Conversely, categories demanding complex compositional reasoning, including *count*, *spatial*, and *viewpoint*, exhibit modest gains. These issues represent intrinsic model weaknesses in operations involving reasoning, which are not easily resolved even with in-domain data. This illustrates that EDIR can effectively distinguish between data-solvable challenges and the more fundamental architectural limitations of current models.

6 Conclusion

We introduce EDIR, a large-scale diagnostic benchmark specifically designed for the granular evaluation of Composed Image Retrieval (CIR) tasks. Constructed through an innovative automated data synthesis pipeline that leverages image editing, EDIR comprises 5,000 queries across fifteen detailed subcategories. The benchmark is crafted to address two key limitations of existing CIR datasets: coarse-grained assessment and limited, ambiguously defined query categories. Our comprehensive evaluation of 13 multimodal embedding models reveals their significant shortcomings on EDIR, highlighting a clear gap in current model capabilities regarding compositional generalization. Furthermore, a thorough comparison against existing CIR benchmarks confirms that EDIR effectively uncovers model weaknesses that other evaluations overlooked. Finally, to validate the unique challenges posed by our benchmark, we conduct an in-domain training experiment. This not only demonstrates the solvability of EDIR but also reveals its ability to distinguish between data-solvable issues and intrinsic model limitations. In conclusion, EDIR provides the community with a robust tool to drive the development of more genuinely compositional and less biased CIR models.

Limitations

While our work introduces a fine-grained benchmark for Composed Image Retrieval (CIR), we acknowledge several limitations that open avenues for future research. First, a key limitation is the cost and scalability of our data synthesis pipeline. Although leveraging programmatic image editing provides precise control over modifications, the process remains computationally expensive, making large-scale data generation a challenge. Second, the complexity of our *Complex* queries is bounded. The queries in our EDIR benchmark are typically composed of three distinct conditions. While more challenging than single-edit queries, they do not yet represent highly complex scenarios with four or more interdependent instructions. This presents an opportunity to develop even more challenging benchmarks. Finally, our work is intentionally focused on evaluation. We designed EDIR primarily as a benchmark to diagnose model weaknesses, rather than as a universal training solution. The development of scalable training methods tailored to address these weaknesses remains an open research direction. In conclusion, while our benchmark serves as an important diagnostic tool, addressing these limitations in scalability, complexity, and training will be crucial for advancing the next generation of CIR models.

References

Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham. 2018. Learning attribute representations with localization for flexible fashion search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7708–7717.

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

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

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402.

Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. 2025. mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. *arXiv preprint arXiv:2502.08468*.

Longye Du, Shuaiyu Deng, Ying Li, Jun Li, and Qi Tian. 2025. A survey on composed image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2025a. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025b. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Google. 2024. Gemini-2.0.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, , Yoohoon Kang, and Sangdoo Yun. 2024a. Language-only training of zero-shot composed image retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Tiancheng Gu, Kaicheng Yang, Kaichen Zhang, Xiang An, Ziyong Feng, Yueyi Zhang, Weidong Cai, Jiankang Deng, and Lidong Bing. 2025. Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning. *arXiv preprint arXiv:2510.13515*.

Markus Hafner, Maria Katsantoni, Tino Köster, James Marks, Joyita Mukherjee, Dorothee Staiger, Jernej Ule, and Mihaela Zavolan. 2021. Clip and complementary methods. *Nature Reviews Methods Primers*, 1(1):20.

Yuxin Hou, Eleonora Vig, Michael Donoser, and Loris Bazzani. 2021. Learning attribute-driven disentangled representations for interactive fashion retrieval. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 12147–12157.

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

Weijian Jian, Yajun Zhang, Dawei Liang, Chunyu Xie, Yixiao He, Dawei Leng, and Yuhui Yin. 2025. Rzenembed: Towards comprehensive multimodal retrieval. *arXiv preprint arXiv:2510.27350*.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024a. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2024b. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*.

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

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. In *The Thirteenth International Conference on Learning Representations*.

Ruiqi Liu, Yi Han, Zhengbo Zhang, Li Yao, Zhiyuan Yan, Jialiang Shen, Zhijin Chen, Bo Sun, Lubin Weng, Jing Dong, Yan Wang, and Shu Wu. 2025. Beyond artifacts: Real-centric envelope modeling for reliable ai-generated image detection. *ArXiv*, abs/2512.20937.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. 2025. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*.

OpenSearch-AI. 2025. Opensearch-ai/ops-mm-embedding-v1-7b.

Bill Psomas, George Retsinas, Nikos Efthymiadis, Panagiotis Filntisis, Yannis Avrithis, Petros Maragos, Ondrej Chum, and Giorgos Tolias. 2025. Instance-level composed image retrieval. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Tingyu Song, Guo Gan, Mingsheng Shang, and Yilun Zhao. 2025a. Ifir: A comprehensive benchmark for evaluating instruction-following in expert-domain information retrieval. In *North American Chapter of the Association for Computational Linguistics*.

Xuemeng Song, Haoqiang Lin, Haokun Wen, Bohan Hou, Mingzhu Xu, and Liqiang Nie. 2025b. A comprehensive survey on composed image retrieval. *ACM Transactions on Information Systems*, 44(1):1–54.

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

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. 2024. CoVR-2: Automatic data construction for composed video retrieval. *IEEE TPAMI*.

Yongquan Wan, Guobing Zou, and Bofeng Zhang. 2025. Composed image retrieval: a survey on recent research and development. *Applied Intelligence*, 55(6):482.

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

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. 2025. Qwen-image technical report.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317.

Ren-Di Wu, Yu-Yen Lin, and Huei-Fang Yang. 2024. Training-free zero-shot composed image retrieval via weighted modality fusion and similarity. In *International Conference on Technologies and Applications of Artificial Intelligence*, pages 77–90. Springer.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xin Yang, Xuemeng Song, Xianjing Han, Haokun Wen, Jie Nie, and Liqiang Nie. 2020. Generative attribute manipulation scheme for flexible fashion search. In *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval*, pages 941–950.

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

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. In *Forty-first International Conference on Machine Learning*.

Siyue Zhang, Yuan Gao, Xiao Zhou, Yilun Zhao, Tingyu Song, Arman Cohan, Anh Tuan Luu, and Chen Zhao. 2025. Mrmr: A realistic and expert-level multidisciplinary benchmark for reasoning-intensive multimodal retrieval. *ArXiv*, abs/2510.09510.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024. Vista: Visualized text embedding for universal multi-modal retrieval. *arXiv preprint arXiv:2406.04292*.

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A EDIR

A.1 Details of Construction

As shown in Figure 2, we use Qwen25-VL-32B-Instruct (Bai et al., 2025) for both seed image selection and edit-instruction generation. For image editing, we use Qwen-Image-Edit (Wu et al., 2025) (version Qwen-Image-Edit-2509) to generate the target images. For query rewriting, we use Qwen3-32B (Yang et al., 2025) to rewrite each edit instruction into a CIR query according to a predefined template. We adopt two rewriting strategies: (i) directly rewriting the instruction into a CIR query, and (ii) rewriting it into a negation-form query. Since not all categories are suitable for negation, we only apply negation rewriting to the *color*, *shape*, *material*, *texture*, *style*, *weather*, and *time* categories. Our study is reviewed and approved by the Institutional Review Board (IRB) of the first author's affiliated institution.

A.2 Error Analysis

For error analysis, we examine representative examples where the state-of-the-art model, RzenEmbed-7B, achieved a low Recall@1 score.

Failure in Handling Negation. As shown in Figure 5, we observe two types of negation-related queries. The first type is explicit negation, where the user requests *not* to keep an attribute of the reference image (e.g., not keeping the T-shirt in its original color), as shown in Figure 5(a). The second type corresponds to *remove* edits, where the user requests an object or region to be removed (e.g., an empty wall above the bed), as shown in Figure 5(b). In both cases, the retrieved results tend to preserve the negated attribute or fail to realize the removal, indicating difficulty in mapping negation to the intended target state.

Deficiencies in Compositional Reasoning. Models exhibit poor performance on categories such as *count*, *spatial*, *style*, and *viewpoint*, which require compositional reasoning. As shown in Figure 6, the query asks for a similar object *with a classroom background*. However, the retrieved images often match the object appearance while failing to align the global scene context, suggesting limited capability in jointly reasoning about foreground content and background context.

Struggles with Multiple Constraints. In the *complex* category, queries specify multiple constraints, yet models frequently retrieve images that only partially satisfy them. As shown in Figure 7, the top retrieved result matches the presence of "a jug" and "a sponge" and roughly matches "the garage-like background", but fails to satisfy the fine attribute constraint that "the jug handle is black". This indicates a weakness in composing and verifying multiple distinct requirements from a single query.

Insensitivity to Fine-Grained Details. For categories such as *texture*, *material*, and *shape*, the distinctions between the source and target images can be subtle, and current models tend to overlook such fine-grained visual cues. As shown in Figure 8, the model ignores the fine-grained details of the jar in the reference image and retrieves results that merely contain a jar, without preserving the intended subtle characteristics.

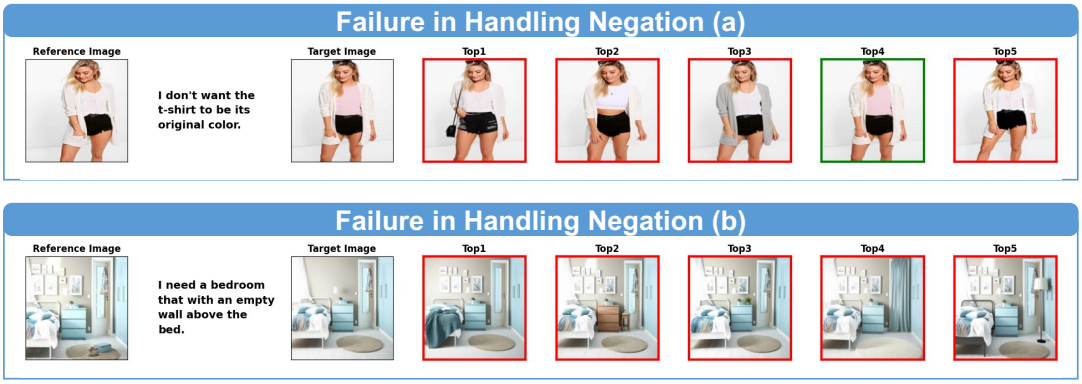

Figure 5: Example of Error Type: *Failure in Handling Negation*

Figure 6: Example of Error Type: *Deficiencies in Compositional Reasoning*

Figure 7: Example of Error Type: *Struggles with Multiple Constraints*

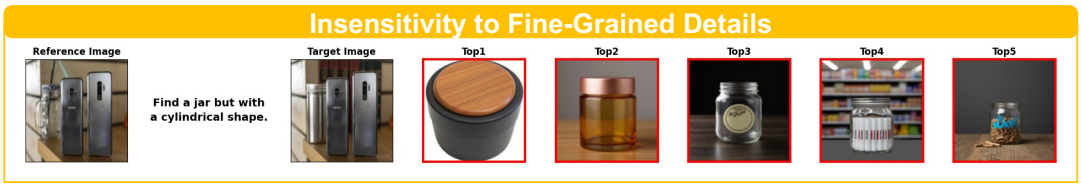

Figure 8: Example of Error Type: *Insensitivity to Fine-Grained Details*

B Experiment Settings

B.1 Evaluation Details

Model Settings We provide the details of the evaluated models in Table 4. For Non-MLLM-based models, we use their CLIP-L/14 variants to ensure a fair comparison, including PIC2WORD, SEARLE and MAGICLENS. For MLLM-based models, we set the maximum sequence length to 2048 and the maximum number of pixels to 1280. And the instruction we used is "Given an image, find a similar image satisfying the query. ".

Benchmark Settings We evaluate models on EDIR using Recall@1. For the other benchmarks, we follow their standard evaluation metrics: CIRR (Recall@1), CIRCO (mAP@5), FASHIONIQ (Recall@10), and GENECIS (Recall@1). For CIRR, we follow the evaluation protocol in (Jiang et al., 2024a), excluding the reference image from the retrieval corpus. For both CIRR and CIRCO, we report the results on the validation set.

B.2 Training Settings

Using our data synthesis pipeline, we edit 500,000 images from LAION-400M, producing 1,087,710 training instances. Each instance consists of a reference image, a query, a target image, and three hard negatives sampled from the same source. From this pool, we sample 15,000 triplets per category across 15 categories, yielding a final training set of 225,000 instances. We train Qwen2.5-VL-7B-Instruct (Bai et al., 2025) with a batch size of 128. The maximum number of image tokens is set to 1,280, and the maximum sequence length is 1,500. The learning rate is 3e-5 with a weight decay of 0.01. We apply LoRA only to the q_proj, k_proj, v_proj, up_proj, down_proj, and gate_proj layers. Training uses an InfoNCE-style loss with a temperature of 0.03.

B.3 Results

We provide further details on model performance on EDIR using additional metrics (i.e., Recall@3), as shown in Table 5. All models exhibit a significant performance increase when evaluated with Recall@3. However, the average performance of MLLM-based models remains close to 60, indicating a substantial gap that still needs to be addressed. In addition, the zero-shot models (e.g., MMRet-MLLM, E5-V, and MAGICLENS) still fail to perform well. Moreover, our benchmark aims to evaluate the fine-grained capabilities of these models. As shown in Table 5, EDIR can still reveal model weaknesses in specific categories, such as *remove* and *view*.

Model	Release	Version	Pretrained	Backbone
		MLLM-based Models		
RzenEmbed-7B	2025-11	`RzenEmbed-7B-V2`	✗	Qwen2-VL-7B
Ops-embedding	2025-07	`Ops-MM-Embedding-v1-7B`	-	Qwen2-VL-7B
GME-2B	2024-12	`gme-Qwen2-VL-2B-Instruct`	✓	Qwen2-VL-2B
GME-7B	2024-12	`gme-Qwen2-VL-7B-Instruct`	✓	Qwen2-VL-7B
MMRet-MLLM	2025-04	`BGE-VL-MLLM-S1`	✓	Llava-Mistral-7B
E5-V	2024-07	`e5-v`	✗	Llava-llama3-8B
VLM2Vec-2B	2025-05	`VLM2Vec-V2.0`	✓	Qwen2-VL-2B
UniME-2B	2025-10	`UniME-V2-Qwen2VL-2B`	✓	Qwen2-VL-2B
UniME-7B	2025-10	`UniME-V2-Qwen2VL-7B`	✓	Qwen2-VL-7B
mmE5	2025-02	`mmE5-mllama-11b-instruct`	✗	Llama-3.2-Vision
		Non-MLLM-based Models		
PIC2WORD	2023-02	`PIC2WORD(CLIP-L/14)`	✗	CLIP-L/14
SEARLE	2023-03	`SEARLE(CLIP-L/14)`	✗	CLIP-L/14
MAGICLENS	2024-03	`MAGICLENS(CLIP-L/14)`	✗	CLIP-L/14

Table 4: Details of the evaluated multimodal embedding models in EDIR.

Metric	Total	Attribute				Object				Relationship			Style			Complex
		Color	Material	Shape	Texture	Add	Remove	Replace	Count	Spatial	Action	View	Style	Weather	Time	Complex
						Non MLLM-based Models										
PIC2WORD	**42.2**	**42.7**	**34.3**	**37.3**	**38.0**	**50.0**	**31.0**	40.0	**47.7**	**35.7**	**50.0**	**32.0**	**42.3**	**49.3**	**51.0**	**45.9**
SEARLE	33.0	34.0	29.0	28.0	26.3	39.0	19.7	36.0	40.3	24.7	48.0	22.0	21.7	42.3	37.0	38.5
MAGICLENS	29.9	29.7	17.7	17.7	14.7	37.0	28.7	**41.3**	31.7	33.0	38.3	23.7	**42.3**	21.0	34.3	32.5
Avg.	35.0	35.4	27.0	27.7	26.3	42.0	26.4	39.1	39.9	31.1	45.4	25.9	35.4	37.6	40.8	39.0
						MLLM-based Models										
Ops-embedding	**71.3**	**70.3**	**65.7**	**64.3**	**70.0**	86.7	49.7	81.7	**73.7**	**69.3**	**81.7**	56.0	**70.0**	**70.0**	**76.3**	**76.4**
RzenEmbed-7B	69.6	64.3	57.7	63.3	60.7	**87.0**	**57.7**	**84.0**	71.0	**69.3**	80.3	**57.0**	67.7	66.3	74.0	74.8
gme-2B	66.1	61.0	59.7	61.7	62.7	79.7	48.7	75.3	68.3	64.3	75.7	48.3	66.7	68.7	71.7	71.0
gme-7B	62.9	59.0	57.0	58.3	53.0	77.7	52.3	75.3	65.0	60.3	68.3	50.3	58.7	63.7	68.7	67.9
VLM2Vec-2B	61.7	59.7	52.3	61.7	67.0	76.3	49.3	60.7	68.7	61.0	68.7	45.7	51.3	58.3	66.0	68.0
MMRet-MLLM	58.0	53.3	44.7	49.3	49.3	75.0	48.3	68.3	57.7	63.7	62.7	46.0	41.7	57.7	63.0	69.8
E5-V	56.2	47.0	47.0	48.0	54.0	70.7	37.3	69.3	61.7	60.7	70.0	36.7	56.7	50.0	59.7	63.1
mmE5	53.8	53.3	44.3	48.0	49.3	64.3	48.3	64.0	56.0	56.0	62.3	37.7	53.7	52.0	54.7	57.1
UniME-7B	50.2	36.3	29.3	37.7	40.0	64.3	38.7	56.3	60.7	56.0	66.0	37.0	48.7	46.3	51.0	63.4
UniME-2B	49.4	45.7	39.7	40.7	46.3	59.7	40.7	67.0	51.0	48.3	61.7	38.7	41.3	41.0	51.3	56.1
Avg.	59.9	55.0	49.7	53.3	55.2	74.1	47.1	70.2	63.4	60.9	69.7	45.3	55.6	57.4	63.6	66.8
EDIR-MLLM	80.8	76.3	79.0	73.3	81.3	92.7	66.7	87.3	82.0	73.0	87.7	63.7	86.0	89.7	87.7	82.4

Table 5: Models Recall@3 performances on EDIR.

Seed Image Selection

You are an AI assistant that judges if an image is suitable for common image editing tasks like adding/removing objects, replacing elements, or changing the background.
Analyze the provided image and determine its suitability.

An image is considered **NOT suitable** if it is:

1. **Primarily Text:** A screenshot of a document, a presentation slide, or code with no significant visual elements.
2. **Too Simple:** A solid color, a simple gradient, or a basic pattern with no distinct objects to manipulate.
3. **Poor Quality:** The image is low-resolution, blurry, or heavily pixelated, especially when the composition is complex. This combination makes it impossible to identify or edit objects cleanly.
4. **Too Abstract or Cluttered:** An abstract pattern, a dense texture, or a chaotic collage where there is no clear subject or distinction between foreground and background.
5. **Functional:** A QR code, barcode, or captcha, where editing would destroy its purpose.

Based on your analysis, provide your output ONLY in the following JSON format:

```
{
    "useful": <true_or_false>,
    "reason": "<A brief explanation for your decision.>"
}
```

Figure 9: Prompt used to judge whether an image is suitable for image editing.

C Prompts.

As shown in Figure 2, we first prompt Qwen25-VL-32B-Instruct to filter out the images that are not suitable for editing. The prompt is shown in Figure 9. After obtaining the seed images, we prompt Qwen25-VL-32B-Instruct to generate edit instructions for these seed images. For each image, the MLLM is required to generate edit instructions for 5-6 categories and 3 edit instructions for each category. The prompt is shown in Figure 10. As we have two methods for prompt rewriting, we provide the prompt for direct rewriting in Figure 11, and we provide the query negation rewrite prompt as shown in Figure 12. For the two stage filtering, we utilize the same prompt template, as shown in Figure 13.

Edit Instruction Generation

The model must output a single valid JSON object with the following structure:

```
{
   "image_description":  "<one-sentence description of the image>",
   "categories":  {
      "<category_1>":  {
         "instructions":  [
            "<instruction_1>",
            "<instruction_2>",
            "<instruction_3>"
         ]
      },
      "<category_2>":  { "instructions":  [ ...  ]  },
      ...
   }
}
```

— RULES —

1. **Top-level keys:** The JSON root *must* contain exactly two keys: `"image_description"` and `"categories"`.
2. **Categories count:** The `"categories"` object must contain 5–6 keys, each selected from the allowed category list.
3. **Allowed category keys:**
   `color`, `material`, `shape`, `texture`, `addition`, `remove`, `replace`, `cardinality`, `spatial`, `action`, `viewpoint`, `style`, `time`, `weather`.
4. **Instructions list:** Each chosen category must contain an `"instructions"` list with 2–3 atomic editing instructions.
5. **Instruction independence:** Instructions across different categories must be combinable without logical conflicts. Do not create edits that negate each other (e.g., removing an object and also recoloring it).
6. **Atomicity:** Each instruction must describe a single concrete change applied to one object or one cohesive group.
7. **Real-world plausibility:** All edits must be realistic and physically plausible; avoid fantasy-like transformations.
8. **Concreteness:** Avoid vague terms like "enhance" or "improve"; instead, specify explicit changes (e.g., "Change the sky to a clear, bright blue.").
9. **Category balance:** Pay particular attention to `remove`, `replace`, `cardinality`, `viewpoint`, `shape`, `time`, and `texture`, ensuring these are used and not neglected.

— EXAMPLE —

```
{
   "image_description":  "A woman wearing a dress standing in a living room."     ,
   "categories":  {
      "remove":  {
         "instructions":  [
            "Remove the coffee table from the scene.",
            "Remove the rug from under the furniture,
            exposing the floor.",
            "Remove the woman, leaving an empty living
            room."
         ]
      },
   }
}
```

Figure 10: Prompt used to generate image editing instructions.

CIR Query Generation

Given an image edit query, rewrite it into an image search query. The goal is to create a search that finds an image matching the final, desired scene.

— GUIDELINES —

1. **Describe the Final State:** Convert action commands (like "add", "make", "move") into descriptive phrases ("a picture of...", "a scene where...").

2. **Omit Comparative Words:** Always remove words that compare to the original image, like "larger", "more", "brighter".

3. **Handle Relational Details Intelligently (CRITICAL):**

   (a) If adding a NEW object: You can often omit its location relative to existing objects to get a better search. Focus on the new object itself. (e.g., "Add a bird on the fence" -> "A picture with a bird").

   (b) If changing the relationship between EXISTING objects: The new relationship is the most important detail and MUST be included in the search. (e.g., "Move the cat onto the sofa" -> "A picture of a cat on a sofa").

— EXAMPLES —

CASE 1: Adding a new object (Omit relation)
Edit Query: Add a flock of seagulls flying near the kitesurfer.
Rewritten Search: I want to see a picture with seagulls flying.
(Reason: The core request is to add seagulls. Their exact position 'near the kitesurfer' is secondary and omitted.)

CASE 2: Changing an object's relationship (Keep relation)
Edit Query: Move the dog so it is sitting at the man's feet.
Rewritten Search: A picture of a dog sitting at a man's feet.
(Reason: The entire point of the edit is the new relationship between the dog and the man. This detail is essential and must be kept.)

CASE 3: Changing a scene attribute (Omit comparative)
Edit Query: Make it a windy day with larger waves.
Rewritten Search: I want to see a windy weather of this place.
(Reason: The comparative "larger" is omitted. The core idea is the windy weather.)

— TASK —
Query: [FILL_THE_QUERY]
You only need to output the rewritten query without any other words or characters. The output should begin with the following prefix. Here is the prefix: [FILL_THE_PREFIX]
If the prefix is "empty", you should simply return a description of the final scene.

Figure 11: Prompt used to convert edit instruction to CIR query. This prompt corresponds to the direct rewriting strategy.

CIR Negation Query

Given an image edit query, rewrite it into an image search query. The goal is to create a search that finds an image matching the final, desired scene.

— GUIDELINES —

1. Convert positive statements about an attribute into a negative or relative query.
2. Focus on the attribute being changed, not the final state.
3. Avoid describing the final appearance; instead, state what should be different.

— EXAMPLES —

CASE 1: Changing an attribute (Color)
Edit Query: Change the dress color to red.
Rewritten Search: Find this dress but in a different color.
(Reason: The query asks for any color other than the original, not specifically red.)

CASE 2: Changing an attribute (Style)
Edit Query: Change the style to a watercolor painting.
Rewritten Search: Show me this picture but not as a photograph.
(Reason: The query negates the current style to find alternatives.)

— TASK —

Query: [FILL_THE_QUERY]
You only need to output the rewritten query without any other words or characters. The output should begin with the following prefix. Here is the prefix: [FILL_THE_PREFIX]
If the prefix is "empty", you should simply return a description of the final scene.

Figure 12: Prompt used to convert an edit instruction to a negation CIR query.

Image Pair Matching

Task:
Your task is to act as a quality control checkpoint. You will be given a `source image`, a `text description`, and a `target image`. Your task is to determine if the `text description` accurately describes the transition from the `source image` to the `target image`.

Failure Criteria:
The text description is considered a 'fail' if it meets **ANY** of the following conditions:

1. **Description Mismatch:** The text does not accurately reflect the actual changes between the source and target images. (e.g., The text describes "making the sky blue," but the actual change from source to target shows the sky turning red).
2. **Subject Inconsistency:** The core subject or scene in the target image is fundamentally different from the source image, and this difference is **not** mentioned in the text description. (e.g., The source shows a dog, the target shows a cat, but the text only mentions "removing the background"). However, if the subject basically belongs to the same category, it is acceptable.
3. **Transition Gap:** The text fails to describe significant visible changes between the source and target images, leaving important transitions unexplained.
4. **Over-Description:** The text describes changes that are not actually present in the transition from source to target image.

Here is the text description: [FILL_THE_QUERY]

Required Output Format:
Please respond strictly in json format, without any additional comments. The json should contain two keys: 'verdict' and 'reason'.

```
{
verdict: [pass / fail]
reason: [If 'fail', provide a brief, specific reason based on the Failure
Criteria.]
}
```

Examples of "fail" Reasons:

- "Reason: The text describes changing the car's color to red, but no color change is visible between source and target images."
- "Reason: The text fails to mention the significant change in background scenery from urban to rural."
- "Reason: The text describes adding a dog, but the target image shows a cat was added instead."

Figure 13: Prompt used to assess the match among a source image, a text description, and a target image, serving as a quality-control checkpoint in our data-filtering pipeline.