

FineCIR: Explicit Parsing of Fine-Grained Modification Semantics for Composed Image Retrieval

Zixu Li¹ Zhiheng Fu¹ Yupeng Hu^{1*} Zhiwei Chen¹ Haokun Wen^{2,3} Liqiang Nie³

¹ School of Software, Shandong University

² School of Data Science, City University of Hong Kong

³ School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)

{Lzx, zivchen, fuzhiheng0215}@mail.sdu.edu.cn

huyupeng@sdu.edu.cn, {whenhaokun, nieliqiang}@gmail.com

Abstract

Composed Image Retrieval (CIR) facilitates image retrieval through a multimodal query consisting of a reference image and modification text. The reference image defines the retrieval context, while the modification text specifies desired alterations. However, existing CIR datasets predominantly employ coarse-grained modification text (CoarseMT), which inadequately captures fine-grained retrieval intents. This limitation introduces two key challenges: (1) ignoring detailed differences leads to imprecise positive samples, and (2) greater ambiguity arises when retrieving visually similar images. These issues degrade retrieval accuracy, necessitating manual result filtering or repeated queries. To address these limitations, we develop a robust fine-grained CIR data annotation pipeline that minimizes imprecise positive samples and enhances CIR systems' ability to discern modification intents accurately. Using this pipeline, we refine the FashionIQ and CIRR datasets to create two fine-grained CIR datasets: Fine-FashionIQ and Fine-CIRR. Furthermore, we introduce FineCIR, the first CIR framework explicitly designed to parse the modification text. FineCIR effectively captures fine-grained modification semantics and aligns them with ambiguous visual entities, enhancing retrieval precision. Extensive experiments demonstrate that FineCIR consistently outperforms state-of-the-art CIR baselines on both fine-grained and traditional CIR benchmark datasets. Our FineCIR code and fine-grained CIR datasets are available at <https://github.com/SDU-L/FineCIR.git>.

1. Introduction

Composed Image Retrieval (CIR) is a crucial paradigm in image retrieval, with widespread applications in fields such as product search [7, 22] and information retrieval [4, 28]. Its queries take a multimodal form, comprising a reference

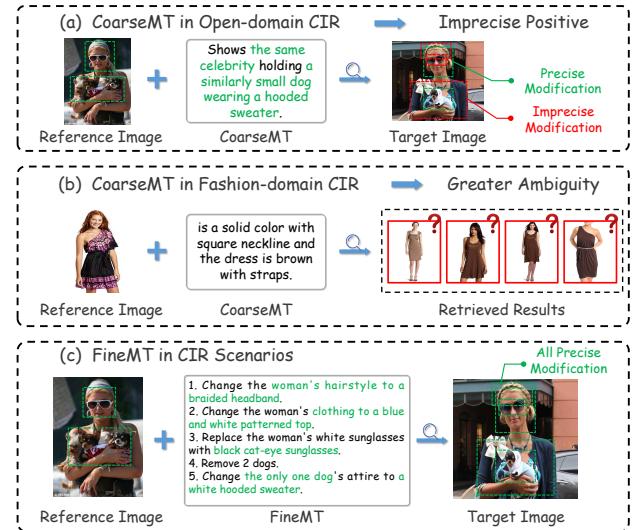


Figure 1. Problems caused by CoarseMT in (a) Open-domain CIR and (b) Fashion-domain CIR. And (c) illustrates our FineMT in CIR Scenarios.

image and a modification text, where the modification text articulates the user's modification intent regarding the reference image. The CIR model retrieves the target image that aligns most closely with the modified reference image by analyzing the multimodal query. Due to the unique ability of CIR to convey user modification intent, it has garnered significant attention in recent years [8, 13, 19, 20, 30, 41].

According to their respective domains, current mainstream CIR datasets can be categorized into two types: open-domain datasets (CIRR [23], LaSCo [16], SynthTriplets18M [10]) and fashion-domain datasets (FashionIQ [39], Shoes [11], Fashion200K [12]). However, these datasets are often annotated using coarse-grained modification text (CoarseMT), which fails to convey users' fine-grained retrieval needs. Due to the following limitations, they may negatively impact training and the CIR process.

L1: Ignoring detailed differences leads to imprecise positive samples.

Open-domain datasets are characterized by rich visual content, often exhibiting significant detail differences between the reference image and the target image. As illustrated in Figure 1(a), taking CIRR dataset as an example, the bounding boxes highlight these differences between the reference image and the target image: green boxes indicate the differences mentioned in the CoarseMT, while red boxes indicate the differences that the CoarseMT fails to capture. This suggests that CoarseMT often overlooks fine details, causing deviations from the intended modifications. We define such target images as imprecise positive samples, which are prevalent in existing datasets. When used for training, these samples may bias CIR models toward accepting candidate images with unintended modifications, even if they do not fully align with the multimodal query. This phenomenon reduces retrieval accuracy and forces users into iterative refinement and searches.

L2: Greater ambiguity arises when retrieving visually similar images.

Fashion-domain datasets are characterized by limited category diversity and a high prevalence of visually similar images. As shown in Figure 1(b), using FashionIQ-Dress as an example, CoarseMT contains only a dozen words or fewer, lacking detail for precise retrieval. This increases ambiguity in retrieval modeling, making it hard for CIR models to differentiate between similar images. Moreover, real-world applications contain even more visually similar images than small-scale datasets, further complicating the retrieval of user-intended results and increasing the burden of manual filtering or repeated queries.

To address these limitations, we introduce fine-grained CIR. As illustrated in Figure 1(c), the key distinction between fine-grained CIR and traditional CIR is the replacement of CoarseMT with fine-grained modification text (FineMT), which comprehensively describes the differences between the reference image and the target image. This approach reduces imprecise positive samples and enhances the accuracy of CIR systems in capturing user modification intent (more examples in Appendix A.1). Specifically, compared to traditional CIR, fine-grained CIR significantly enhances CIR models in multiple aspects, including rate of imprecise positive samples, requirements modeling certainty, and retrieval accuracy, etc (detailed comparison in Appendix A.2). To advance research in this area, we develop a **fine-grained CIR data annotation pipeline** comprising three stages: *Data Selection*, *Construction*, and *Quality Check*. This pipeline enables the enrichment of existing CIR datasets with fine-grained annotations. We apply this pipeline to two widely used CIR datasets, FashionIQ and CIRR, resulting in **Fine-FashionIQ** and **Fine-CIRR**. These datasets support the development of CIR models and enhance their fine-grained retrieval capabilities.

Existing CIR models primarily rely on implicit methods

to interpret modification text, making fine-grained modification recognition challenging. To address this, we propose **FineCIR**, an explicit parsing network for Fine-grained semantics in Composed Image Retrieval. FineCIR is the first CIR framework that explicitly parses modification semantics, facilitating the structured extraction of entities, attributes, and relations from the modification text. By leveraging explicit parsing, FineCIR enables a precise understanding of FineMT’s semantics while aggregating extracted information based on its corresponding entity. We then use these explicitly aggregated entity tokens as guidance to achieve indeterminate relationship alignment between visual entities and FineMT’s semantics, thereby completing multimodal feature fusion. Through this explicit parsing mechanism, FineCIR more accurately captures modification intent, ultimately enhancing the performance of fine-grained CIR.

The main contributions of this work are as follows:

- We construct a fine-grained CIR data annotation pipeline that minimizes imprecise positive samples, improves the handling of visually similar images, and enhances modification intent recognition in CIR systems.
- Using this pipeline, we develop two fine-grained CIR benchmark datasets, Fine-FashionIQ and Fine-CIRR. They’ve both been rigorously validated through automated and manual processes, ensuring their suitability for fine-grained CIR model development. Furthermore, they are fully open-sourced and publicly available.
- We propose FineCIR, the first CIR framework to explicitly parse modification semantics. Extensive experiments on four benchmark datasets demonstrate its superior performance on both traditional CIR and fine-grained CIR.

2. Related Work

Composed Query Datasets. Composed query datasets are structured as triplets, each comprising a multimodal query and a target. The query consists of a reference image/video and a modification text. These datasets are widely used in retrieval tasks, including Composed Image Retrieval (CIR) [10, 35, 37, 40, 46], Zero-shot CIR (ZS-CIR) [9, 15, 43, 44], and Composed Video Retrieval (CoVR) [14, 31, 33, 34]. However, most existing studies rely on CoarseMT, which lacks precision for fine-grained retrieval. Wu *et al.* [45] proposed generating detailed textual descriptions from video content to advance fine-grained CoVR research. CIR datasets often overlook the limitations of CoarseMT, leading to imprecise positive samples and greater ambiguity when retrieving similar images, hindering CIR model training and retrieval. To address this, we introduce fine-grained CIR, improving model training and enabling CIR systems to accurately interpret FineMT.

Composed Image Retrieval. Composed Image Retrieval (CIR) aims to retrieve target images based on a reference

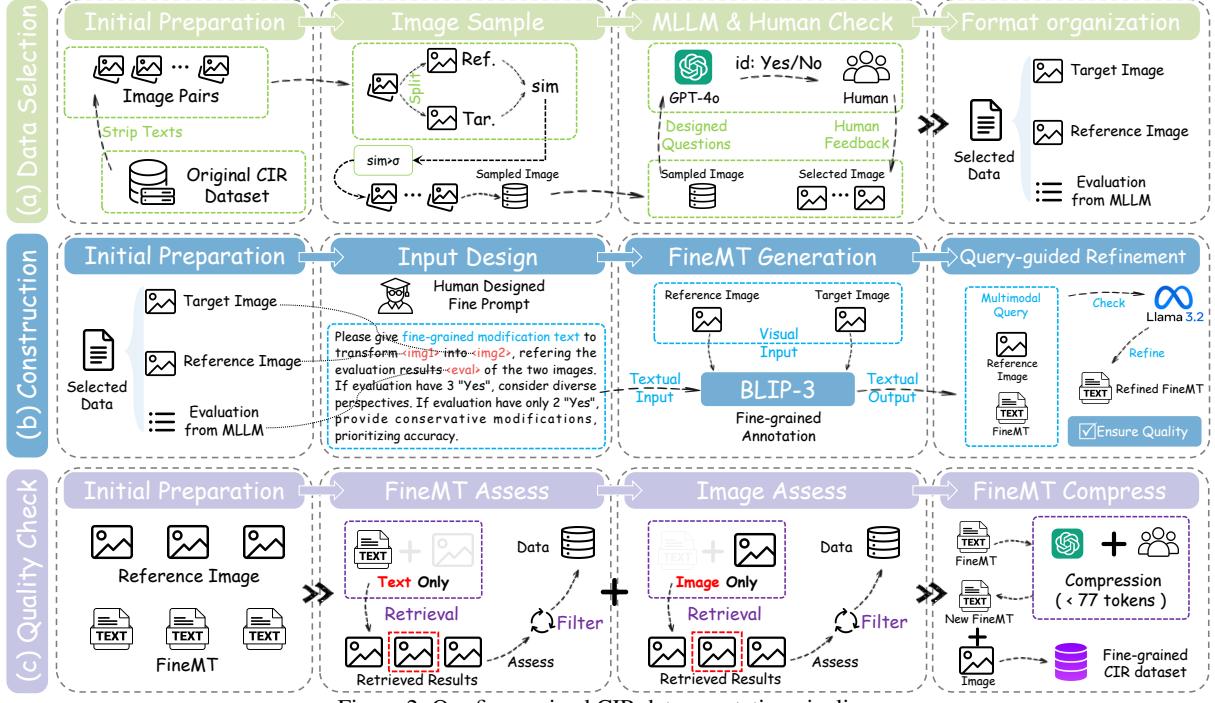


Figure 2. Our fine-grained CIR data annotation pipeline.

image and modification text. According to the utilized backbone, existing CIR methods can be broadly categorized into two streams: traditional models [35, 36] that separately process and compose visual-textual information, and VLP-based models [25, 37, 40] that embed multimodal queries to unified feature spaces. In addition, due to the high cost of labeling training data for CIR tasks, some researchers have turned to the Zero-shot CIR (ZS-CIR) task. Different from the CIR approach, existing ZS-CIR studies [9, 10] tend to leverage image-caption pairs to train text inversion networks to achieve ZS-CIR. However, due to the high cost of triplet annotation, research on CIR datasets remains relatively underexplored. To address this challenge, we construct a robust fine-grained CIR data annotation pipeline, integrating automated and manual processes for quality assurance. This pipeline facilitates the incorporation of fine-grained annotations into existing CIR datasets, enhancing support for fine-grained CIR.

3. Fine-grained CIR data annotation pipeline

As shown in Figure 2, we constructed a **fine-grained CIR data annotation pipeline** to enhance fine-grained CIR research. This pipeline comprises three stages. Except for “Initial Preparation”, each stage goes through 3 steps. First, the *Data Selection stage* (Section 3.1) facilitates image pair selection. Second, the *Dataset Construction and Refinement stage* (Section 3.2) generates FineMT and conducts an initial multimodal query quality check. Finally, the *Quality Check stage* (Section 3.3) verifies triplet data from multiple perspectives to ensure reliability. We applied this pipeline to

two benchmark CIR datasets, FashionIQ and CIRR, resulting in two fine-grained CIR datasets (**Fine-FashionIQ** and **Fine-CIRR**). They were designed to support CIR model development and enhance fine-grained retrieval performance.

3.1. Data Selection

Since existing CIR datasets provide high-quality reference-target image pairs (hereafter referred to as image pairs), we leverage the triplets from the FashionIQ and CIRR datasets as the starting point for dataset construction to minimize overhead. For each triplet, we retain its image pairs and generate fine-grained modification text (FineMT) to describe their differences. As the following fine-grained CIR data annotation pipeline is applicable to various CIR datasets, we do not distinguish between the original dataset names in subsequent sections. To construct a high-quality fine-grained CIR dataset, as shown in Figure 2(a), we first filter the raw image pairs.

Image Sample. First, as shown in the “Image Sample” step of Figure 2(a), we initially filter image pairs based on semantic similarity. Specifically, we use the pre-trained vision-language model BLIP [17] to encode the images and compute the cosine similarity between each pair. A filtering threshold is set to remove image pairs with similarity scores below this threshold. This filtering step accounts for potential mismatches in the original dataset, where some triplets may contain image pairs with excessive scene transitions and deviate from real-world application scenarios, and thus should be excluded. After similarity-based sampling, we obtain the sampled images from the original CIR dataset.

MLLM & Human Check. As shown the “MLLM & Hu-

man Check” step in Figure 2(a), after obtaining sampled images, we refine image pairs using MLLM and human verification. Specifically, we design a set of evaluation questions and employ GPT-4o [27] to assess each image pair:

- Are the contents of the two images related, ensuring a logical and meaningful modification from the reference to the target?
- Can the reference image provide valuable information for retrieving the target image?
- Are there sufficient modifiable aspects between the two images, and are these modifications feasible and meaningful for fine-grained modification?

We instruct GPT-4o (prompts detailed in Appendix B.1) to provide three “Yes/No” responses per image pair, guiding the following actions:

- Three “Yes” answers \Rightarrow Retained without modification.
- Two “Yes” answers \Rightarrow Manually reviewed to determine.
- One or zero “Yes” answers \Rightarrow Discarded.

This process ensures that retained image pairs remain relevant and suitable for fine-grained CIR. Through meticulous verification, we obtain the selected images.

Format organization. As shown in the “Format organization” step of Figure 2(a), to facilitate subsequent use, we reorganized the data structure, retaining the selected images in the form of triplets: <reference image, target image, evaluation from MLLM>. This structured format enhances usability for further processing.

3.2. Dataset Construction and Refinement

As shown in Figure 2(b), based on the dataset obtained from the *Data Selection stage*, we designed a comprehensive dataset construction process. Recognizing the powerful multimodal comprehension capabilities of BLIP-3 [42], we employ it as our primary automatic annotation tool to generate FineMT. Subsequently, we refine the FineMT through a combination of MLLM and human verification.

Input Design. As shown in the “Input Design” step of Figure 2(b), based on the selected data, we manually designed fine prompt to serve as input for BLIP-3. Specifically, in these prompts, we instruct the model to generate an initial FineMT for each image pair, leveraging both the image pair and the evaluation from MLLM. In this context, <img1> denotes the reference image, <img2> denotes the target image, and <eval> represents the evaluation from MLLM.

FineMT Generation. As shown in the “FineMT Generation” step of Figure 2(b), we provide BLIP-3 with the fine prompt (detailed prompt in Appendix B.2) as textual input and image pairs as visual input. BLIP-3 compares the <img1> and <img2> while incorporating <eval> as a reference to perform the following operations: (1) If the <eval> contains three “Yes” responses, the model generates FineMT from multiple perspectives. (2) If the <eval> contains two “Yes” responses, the model conservatively

generates several modifications focusing on visual characteristics while ensuring the text remains error-free. Finally, we obtain the textual output (FineMT) from BLIP-3.

Query-guided Refinement. Since BLIP-3’s direct output may contain hallucinations, it is necessary to verify the obtained FineMT to eliminate such artifacts. As shown in the “Query-guided Refinement” step of Figure 2(b), we design specialized prompts (detailed prompt in Appendix B.1) and utilize Llama3.2 [26] to review all multimodal queries (including the reference image and FineMT). This process will remove the hallucinated content, yielding a refined FineMT.

3.3. Quality Check

Through the aforementioned process, we obtain fine-grained multimodal queries. However, we identify two potential issues in the existing data: (1) FineMT contains more detailed information compared to CoarseMT, which may reduce the relevance of the reference image in the CIR process, making the modification text the dominant factor. (2) Most existing CIR datasets construct image pairs based on semantic similarity. In some cases, the reference image alone may be sufficient to accurately retrieve the target image, rendering the modification text less relevant. This contradicts the fundamental objective of CIR, which relies on both textual and visual information. To address these concerns, as shown in Figure 2(c), we utilize feedback from unimodal retrieval results to filter existing triplets from two perspectives: FineMT and reference image.

FineMT Assess. To avoid issue (1), we exclude the reference image and use only FineMT to retrieve the target image. As shown in the “FineMT Assess” step of Figure 2(c), we employ BLIP [17] to extract features from both FineMT and candidate images, then retrieve candidate images using FineMT features. We filter out triplets where the target image appears as the top-1 retrieval result. Next, we conduct a manual assessment, examining the FineMT in relation to the corresponding image pair: (a) If FineMT is overly detailed and includes critical information originally conveyed by the reference image, we refine it using MLLM to ensure that it does not diminish the reference image’s contribution. (b) If the image pair exhibits excessive differences and FineMT merely describes the target image rather than its transformation from the reference image, we discard the triplet.

Image Assess. To avoid issue (2), we exclude FineMT and use only the reference image to retrieve the target image. As shown in the “Image Assess” step of Figure 2(c), we filter out triplets where the target image appears as the top-1 retrieval result. A manual assessment is then conducted to determine whether to retain these triplets.

FineMT Compress. Finally, we recognize the length constraints imposed by different text encoders, particularly the CLIP encoder, which has a maximum text length of 77 tokens. To prevent excessively long FineMT from overload-

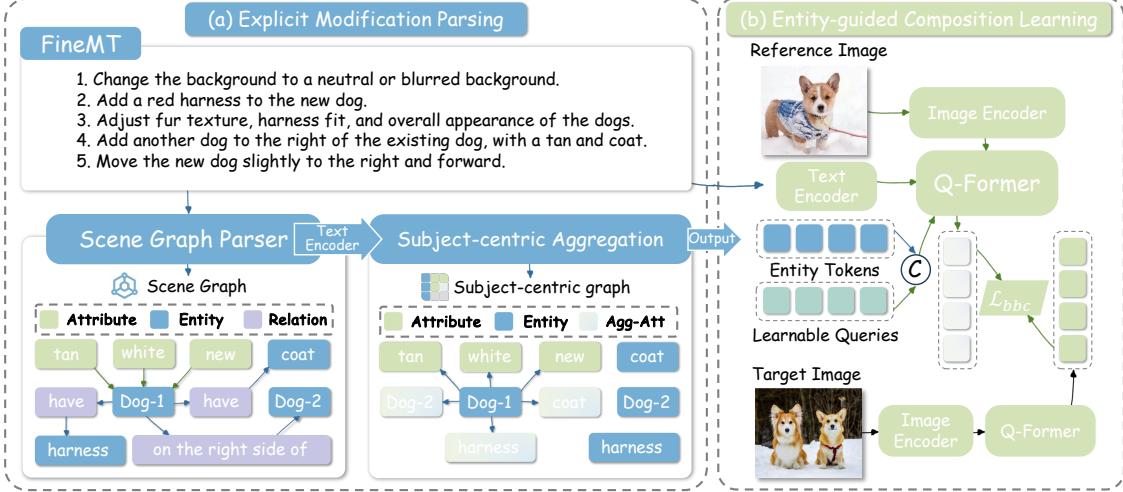


Figure 3. Overall architecture of our proposed FineCIR.

ing the backbone, we use GPT-4o to verify its length. As shown in the ‘‘FineMT Compress’’ step of Figure 2(c), to reduce the workload of manual verification, we first input FineMT exceeding 77 tokens into GPT-4o (detailed prompt in Appendix B.1). It will retain key modification information while omitting redundant expressions, producing a compressed version of FineMT. If the length still exceeds 77 tokens, a manual review is conducted, where human annotators remove irrelevant details while ensuring the completeness of FineMT. Following these steps, we obtain the New FineMT, which, together with the corresponding image pair, forms a triplet. With this, we complete the construction of a fine-grained CIR dataset. In addition, we count the change in query number of Fine-FashionIQ and Fine-CIRR at each step (detail in Appendix B.3).

4. Method

To address the fine-grained CIR, we propose an explicit parsing network of **Fine**-grained semantics for **CIR** (**FineCIR**), which aims to explicitly parse FineMT via the *Scene Graph* (*SG*) to understand fine-grained modification semantics and achieve indeterminate relationship alignment between visual entities and FineMT’s semantics. As illustrated in Figure 3, FineCIR consists of two main components: (a) *Explicit Modification Parsing* (*Eparse*) (detailed in Section 4.2) and (b) *Entity-guided Composition Learning* (*Encompose*) (described in Section 4.3). In the following, we elaborate on each module of our proposed FineCIR.

4.1. Preliminaries

Given $\mathcal{T} = \{(x_r, t_m, x_t)_n\}_{n=1}^N$ as a set of N triplets, where x_r, t_m and x_t refer to the reference image, FineMT and target image, respectively. The designed model aims to learn an embedding space where the multimodal query (x_r, t_m) should be as close as possible to the corresponding target image x_t , which is formulated as $\mathcal{F}(x_r, t_m) \rightarrow \mathcal{F}(x_t)$.

4.2. Explicit Modification Parsing

Since FineMT contains a large amount of fine-grained modification semantics, conventional implicit parsing mechanisms struggle to accurately interpret complex fine-grained semantics. To address this issue, we propose the *Explicit Modification Parsing* (*Eparse*) module. It incorporates a Scene Graph Parser, capable of constructing a scene graph-based structured association framework, which explicitly parses FineMT into an entity-attribute-relation structured semantic representation. Additionally, *Eparse* incorporates Subject-centric Aggregation, which aggregates structured semantics in the scene graph based on subject ownership, serving as guidance for subsequent Entity-guided Composition Learning.

Scene Graph Parser. To explicitly parse the modification semantics in FineMT, we first construct a scene graph-based structured association framework, extracting a multi-dimensional representation of entity-attribute-relation. Specifically, the scene graph parsing of FineMT t_m is formulated as follows:

$$\mathcal{G}(\mathcal{E}, \mathcal{R}) = \Phi_{\mathbb{T}}(\text{SG Parser}(t_m)), \quad (1)$$

where $\Phi_{\mathbb{T}}$ represents the text encoder of BLIP-2, and \mathcal{G} denotes the scene graph corresponding to t_m . $\mathcal{E} = \{\mathbf{E}_1, \dots, \mathbf{E}_u, \dots, \mathbf{E}_U\}$ denotes the entity set contained in FineMT, where U is the entity number. Each entity \mathbf{E}_u has an associated set of attributes, denoted as $\mathcal{A}_{\mathbf{E}_u} = \{\mathbf{A}_1, \dots, \mathbf{A}_k, \dots, \mathbf{A}_K\}$, where K is the number of attributes for entity \mathbf{E}_u . $\mathcal{R} = \{(\mathbf{E}_{sub}, \mathbf{R}_i, \mathbf{E}_{obj})_i\}_{i=1}^r$ denotes the subject-object relation set between entities, where r is the relation number, and \mathbf{E}_{sub} and \mathbf{E}_{obj} represent the subject and object entities, respectively. $\mathbf{E}, \mathbf{A}, \mathbf{R} \in \mathbb{R}^{D_{\mathbb{T}}}$ represent the textual CLS tokens of entities, attributes, and relations, where $D_{\mathbb{T}}$ is the textual embedding dimension.

Subject-centric Aggregation. Given that subject-object relations involve diverse modification semantic operations

(e.g., position adjustment and content addition/removal), they pose challenges to accurately modeling modification semantics. To capture the minimal complete semantics of FineMT, we design *Subject-centric Aggregation*, which aggregates structured semantics in the scene graph based on subject ownership, *i.e.*, it consolidates the object relationships associated with the subject, along with its own attributes, into the subject’s semantic representation.

Specifically, based on the subject-object relation set \mathcal{R} in the scene graph, we reorganize the scene graph \mathcal{G} into a subject-centric graph $\mathcal{G}_{sub} = \{\mathcal{E}_{sub}, \mathcal{U}_{sub}\}$, where \mathcal{E}_{sub} represents the set of all subject entities. \mathcal{U}_{sub} denotes the union of the subject’s attribute set and the associated object set, where the object token integrates the subject-object relation semantics via simple addition, *i.e.*, $\mathbf{E} + \mathbf{R}$.

Subsequently, leveraging Graph Attention Networks [3, 32], we aggregate the object relations associated with the subject, along with its own attributes, into the subject’s semantic representation, formulated as follows:

$$\mathbf{E}_{agg} = \text{GAT}(\mathcal{G}_{sub}), \quad (2)$$

where $\mathbf{E}_{agg} \in \mathbb{R}^{E \times D_{\mathbb{T}}}$ represents the entity tokens after relational aggregation and E denotes the subject number.

4.3. Entity-guided Composition Learning

Since FineMT contains numerous fine-grained modification details, the same visual entity in the reference image may correspond to multiple modification clauses, and a single modification clause may also correspond to multiple visual entities in the reference image. We refer to this phenomenon as indeterminate relationships. To align these indeterminate relationships and prevent incorrect associations between visual entities and fine-grained modification semantics during multimodal query composition, we propose the *Entity-guided Composition Learning (Encompose)* module. This module leverages explicitly aggregated entity tokens as guidance to identify indeterminate relationships between visual entities in the reference image and fine-grained modification semantics in FineMT. Furthermore, it integrates multimodal query semantics while performing composition learning, promoting the composed multimodal query features close to the target image features. In the following, we detail the module design.

Feature Extracting. Specifically, we first extract the features of the reference image and FineMT. Following previous works [34, 40], we utilize the image encoder of BLIP-2 to extract the visual feature $\mathbf{E}_r \in \mathbb{R}^{C \times D_{\mathbb{I}}}$ of the reference image x_r , formulated as,

$$\mathbf{E}_r = \Phi_{\mathbb{I}}(x_r), \quad (3)$$

where C denotes the visual channel and $D_{\mathbb{I}}$ is the visual embedding dimension. $\Phi_{\mathbb{I}}$ represent the BLIP-2 image encoder. Similarly, we extract the visual feature $\hat{\mathbf{E}}_t \in \mathbb{R}^{C \times D_{\mathbb{I}}}$

of the target image and utilize BLIP-2 text encoder to extract the textual feature $\mathbf{E}_m \in \mathbb{R}^{S \times D_{\mathbb{T}}}$ of the FineMT, where S denotes the textual sequence length.

Entity-guided Composition. To explore indeterminate relationships between visual entities in the reference image and fine-grained modification semantics in FineMT, we leverage the entity tokens output from the *Exparse* module as semantic guidance, adaptively aligning the semantic information of the reference image and FineMT.

Specifically, we utilize a set of learnable queries, denoted as $\mathbf{a}_q = \{a_1, \dots, a_k\}$. These queries, along with the entity tokens \mathbf{E}_{agg} (obtained in Eq. (2)) and FineMT textual feature \mathbf{E}_m , are used as the query inputs to the Q-Former. Since entity tokens explicitly aggregate the subject-centric entity-attribute-relation semantics in FineMT, they enable learnable queries to adaptively aggregate multiple fine-grained modification semantics corresponding to the same visual entity and multiple visual entities corresponding to the same modification semantics within the multimodal query. Moreover, previous studies have shown that Q-Former possesses strong multimodal composition capabilities [34, 40]. Thus, we use the reference image’s visual feature \mathbf{E}_r as the cross-modal input to Q-Former, facilitating the adaptive alignment of semantic information between the reference image and FineMT, formulated as,

$$\mathbf{E}_c = \text{Q-Former}([\mathbf{a}_q, \mathbf{E}_{agg}, \mathbf{E}_m], \mathbf{E}_r), \quad (4)$$

where $\mathbf{E}_c \in \mathbb{R}^D$ denotes the multimodal composed token, *i.e.*, the CLS token of the Q-Former’s output, and D denotes the embedding dimension of Q-Former. Similarly, for the target image, we also utilize Q-Former to extract the entity token from its features $\hat{\mathbf{E}}_t$, denoted as $\mathbf{E}_t \in \mathbb{R}^D$.

Token Matching Learning Following previous works [24, 37, 38], to promote multimodal composed tokens to align with their corresponding target entity tokens, we apply the universal batch-based classification loss [5], formulated as,

$$\mathcal{L}_{bbc} = \frac{1}{B} \sum_{i=1}^B -\log \left\{ \frac{\exp \{s(\mathbf{E}_{ci}, \mathbf{E}_{ti}) / \tau\}}{\sum_{j=1}^B \exp \{s(\mathbf{E}_{ci}, \mathbf{E}_{tj}) / \tau\}} \right\}, \quad (5)$$

where as $\mathbf{E}_{ci}, \mathbf{E}_{ti}$ indicate the composed and target token of the i -th triplet, respectively. τ is the temperature coefficient. B is the batch size.

Finally, the loss function of FineCIR is formulated as,

$$\Theta^* = \arg \min_{\Theta} (\mathcal{L}_{bbc}), \quad (6)$$

where Θ^* is the to-be-optimized parameter for FineCIR.

5. Experiments

In this section, we address the detailed experiments on two fine-grained CIR datasets (Fine-FashionIQ and Fine-CIRR) and two CIR datasets (FashionIQ and CIRR).

Method	Year	Fine-FashionIQ										Fine-CIRR									
		Dresses		Shirts		Tops&Tees		Avg		R@k				$R_{subset}@k$			Avg				
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	k=1	k=5	k=10	k=50	k=1	k=2	k=3					
<i>Pre-trained Models</i>																					
Image-Only	-	10.57	25.60	16.94	31.51	14.40	30.09	13.97	29.07	8.53	30.99	44.64	76.58	21.56	41.62	61.58	26.28				
Text-Only	-	14.65	30.74	15.86	32.59	17.11	36.17	15.87	33.17	12.26	31.64	43.54	70.88	56.54	77.39	89.02	44.09				
Image+Text	-	25.38	48.11	29.55	53.11	29.14	54.02	28.02	51.75	27.00	58.17	71.72	93.91	59.77	78.91	90.28	58.97				
<i>Existing Methods</i>																					
TIRG [35]	2019	7.86	16.52	10.21	16.23	9.97	22.31	9.35	18.35	8.71	32.41	46.59	77.35	13.68	40.23	67.88	23.05				
CLVC-Net [36]	2021	13.90	28.63	15.48	31.22	17.08	32.78	15.49	30.88	10.76	35.44	50.72	80.35	31.64	54.94	74.42	33.54				
CLIP4CIR [2]	2022	27.58	51.12	35.62	56.91	34.77	58.43	32.66	55.49	39.48	67.36	80.69	93.87	67.45	83.10	92.34	67.41				
TG-CIR [38]	2023	33.84	59.59	48.25	69.43	46.52	70.11	42.87	66.38	46.06	77.84	87.24	97.40	72.06	87.53	94.91	74.95				
BLIP4CIR [24]	2024	31.49	55.98	39.12	61.57	39.56	63.19	36.72	60.25	41.21	70.43	83.96	95.44	71.52	85.34	94.37	70.98				
CIREVL [15]	2024	24.53	50.78	31.49	51.45	36.58	57.11	30.87	53.11	33.27	62.03	73.11	89.34	66.58	83.02	93.21	64.31				
LinCIR [9]	2024	37.88	61.85	43.01	63.59	47.68	70.33	42.86	65.26	31.48	62.85	72.01	88.59	66.09	81.53	92.98	64.47				
SEIZE [43]	2024	38.45	60.23	41.98	66.47	43.09	69.91	41.17	65.54	35.99	69.01	78.24	91.02	71.58	88.43	93.01	70.30				
Candidate [25]	2024	40.69	61.56	45.63	67.40	45.09	70.37	43.80	66.44	47.96	79.29	86.01	97.28	76.89	88.03	95.62	78.09				
SPRC [40]	2024	43.89	69.18	61.86	81.42	60.35	80.86	55.37	77.15	53.39	81.30	89.31	97.58	80.65	91.60	96.32	80.97				
LIMN [37]	2024	44.04	66.31	53.61	72.90	49.97	72.62	49.20	70.61	47.35	78.44	87.31	97.30	71.69	87.23	94.20	75.07				
CoVR-2 [34]	2024	51.21	76.96	60.59	78.88	60.38	81.27	57.39	79.04	53.48	81.12	89.87	97.66	80.34	91.73	96.55	80.73				
FineCIR (Ours)		55.29	79.61	64.84	84.82	63.42	83.64	61.18	82.69	57.67	85.69	92.02	98.37	83.77	93.86	97.45	84.73				

Table 1. Performance comparison on Fine-FashionIQ and Fine-CIRR relative to R@K(%). The overall best results are in bold, while the best results over baselines are underlined. The Avg metric in Fine-CIRR denotes (R@5 + $R_{subset}@1$) / 2.

5.1. Experimental Settings

Evaluation. We choose the recall at rank K (R@ K) as the evaluation metric, quite similar to the previous CIR task [5, 37, 38]. For Fine-FashionIQ and FashionIQ, we employ R@10, R@50 and their category-wise averages. Similarly, Fine-CIRR and CIRR assessment included R@ k ($k=1, 5, 10$), $R_{subset}@k$ ($k=1, 2$) and the average ($R@5 + R_{subset}@1$) / 2. In addition, we provide a description (detailed in Appendix C.1) and statistical data(detailed in Appendix C.2) of the above datasets.

Implementation Details. We utilize BLIP-2 [18] as the backbone and train FineCIR using the AdamW optimizer with an initial learning rate of $2e-5$. Moreover, we utilize FactualSceneGraph [21] as the scene graph parser in the *Explicit Modification Parsing* module. The batch size is set to 128, and the embedding dimension is set to 256. All experiments are accomplished on a single NVIDIA A40 GPU with 48 GB memory.

5.2. Method Comparison

We conduct a comprehensive evaluation of the proposed FineCIR model against several significant baselines on the above four datasets. Since the source codes for some methods are not accessible, we select a Vision-Language Pre-training (VLP) model, BLIP [17] (for Image-Only, Text-Only, and Image+Text), and some open-source baselines (Candidate [25], SPRC [40], LIMN [37], etc.) and the baselines of CIR-related tasks, including ZS-CIR and CoVR. We retrain and test all of them according to their original configurations on Fine-FashionIQ and Fine-CIRR. Furthermore, we also evaluate our FineCIR on traditional CIR datasets, *i.e.*, FashionIQ and CIRR, to verify its generalizability. We now detail the analysis as follows.

Comparison on Fine-grained CIR Datasets. From the results presented in Table 1, we can obtain the following conclusions: **1)** Our proposed FineCIR achieves superior

performance on both Fine-FashionIQ and Fine-CIRR, indicating its excellent ability to comprehend the fine-grained modification semantics. **2)** FineCIR demonstrates a significant performance advantage over the baselines, particularly gaining 6.6% improvement on Avg.R@10 for Fine-FashionIQ. This may be attributed to the enhancements provided by our *Explicit Modification Parsing* module, which improves the model’s ability to grasp the complex semantics of FineMT. **3)** Previous baselines exhibit suboptimal performance on certain metrics, whereas our FineCIR achieves optimal results on all metrics. This indicates that FineCIR keeps superior generalizability in comprehending fine-grained modifications on fashion and open domains. **4)** Text-Only and Image-Only trained with BLIP exhibit significantly lower performance than Image+Text and other VLP-based CIR frameworks, confirming that both the reference image and FineMT in our fine-grained triplets are highly relevant to the multimodal query. This prevents a single modality from dominating the compositional process, aligning with the fundamental objective of the CIR task.

Method	Year	FashionIQ		CIRR
		Avg. R@10	Avg. R@50	Avg
TIRG [35]	2019	17.40	37.39	35.52
CLVC-Net [36]	2021	30.70	58.41	-
CLIP4CIR [2]	2022	38.40	61.74	69.09
TG-CIR [38]	2023	51.32	73.09	75.57
BLIP4CIR [24]	2024	42.63	66.79	72.08
CoVR-2 [34]	2024	49.96	71.17	78.92
Candidate [25]	2024	51.17	73.13	80.90
SPRC [40]	2024	54.72	74.97	81.39
LIMN [37]	2024	<u>55.91</u>	<u>77.82</u>	72.19
FineCIR		59.86	80.43	81.63

Table 2. Performance comparison on original FashionIQ and CIRR relative to R@K(%). The overall best results are in bold, while the best results over baselines are underlined. The Avg metric in CIRR denotes (R@5 + $R_{subset}@1$) / 2.

Comparison on Traditional CIR Datasets. Due to space constraints, we present the averaged results on the Fash-

Method	Fine-FashionIQ			Fine-CIRR		
	R@10	Δ	R@50	Δ	Avg	Δ
<i>Explicit Modification Parsing (Exparse)</i>						
w/o SG	57.33	-3.85	77.77	-4.92	81.12	-3.61
w/o SC-Agg	59.01	-2.17	79.29	-3.40	82.81	-1.92
w/ SG_AMRBART	60.76	-0.42	82.25	-0.44	84.24	-0.49
w/ SG.Stanford	60.11	-1.07	81.96	-0.73	84.01	-0.72
<i>Entity-guided Composition Learning (Encompose)</i>						
w/o Entity_Guide	59.87	-1.31	79.87	-2.82	83.56	-1.17
w/o Q-Former	59.63	-1.55	79.98	-2.71	82.65	-2.08
<i>Loss Function</i>						
w/o BBC	58.45	-2.73	79.80	-2.89	81.14	-3.59
FineCIR (Ours)	61.18	-0.00	82.69	-0.00	84.73	-0.00

Table 3. Ablation study on Fine-FashionIQ and Fine-CIRR datasets. We compute Avg-R@10, R@50 for Fine-FashionIQ, and Avg (mean of R@5 and R_{subset} @ 1) for Fine-CIRR, respectively.

ionIQ and CIRR datasets in Table 2, while the full results are provided in Appendix D. The results indicate that our proposed FineCIR still achieves the best performance on traditional CIR datasets, demonstrating its ability to effectively comprehend coarse-grained modification semantics. This confirms its strong generalization capability across both coarse and fine granularities.

5.3. Ablation Study

In this section, we introduce the ablation study of our proposed FineCIR with different variants, as shown in Table 3. The compared variants are as follows.

- **w/o SG.** We replace the scene graph with the original FineMT for subject interaction to verify the scene graph’s ability to comprehend fine-grained modifications.
- **w/o SC-Agg.** We remove the *Subject-centric Aggregation* module and use mean pooling on entity features to verify the effectiveness of learning subject-object relations.
- **w/ SG_AMRBART & w/ SG_Stanford.** To verify the robustness of FineCIR, we replace the FactualSceneGraph with AMRBART [1] and StanfordSceneGraph [6].
- **w/o Entity_Guide.** To verify the efficacy of entity tokens in learning indeterminate relations, we replace them with the mean pooling feature of all tokens in the scene graph.
- **w/o Q-Former.** To verify the role of Q-Former, we replace it with an MLP for composition.
- **w/o BBC.** To verify the point of BBC loss, we replace it with the commonly used KL divergence loss for this task [36, 38], which is also used to optimize the distribution of multi-modal queries and target images.

From the ablation results of FineCIR in Table 3, we can obtain the following observations. 1) The performance drop of **w/o SG** is most significant, indicating the superiority of the scene graph in comprehending fine-grained modification semantics. With the aid of scene graph, FineCIR can more precisely comprehend the modification details in FineMT. 2) **w/o SC-Agg** outperforms **w/o SG** but still shows a noticeable performance drop, indicating that subject-centric aggregation benefits in learning fine-grained subject modification relationships. 3) **w/ SG_AMRBART**

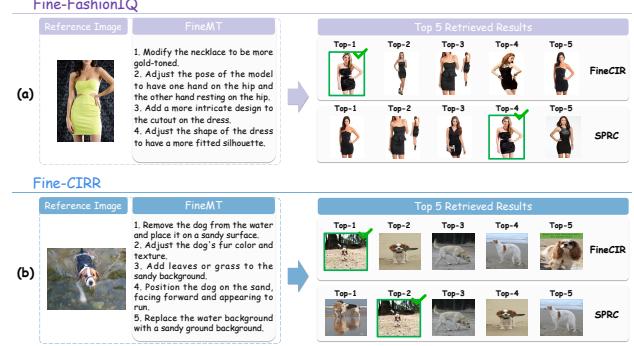


Figure 4. Qualitative examples of our proposed FineCIR compared to the sub-optimal CIR model SPRC.

and **w/ SG_Stanford** both outperform **w/o SC-Agg**, with only a slight performance drop. This indicates the necessity of scene graphs for fine-grained modification learning and demonstrates that FineCIR’s subsequent architecture is robust in learning relationships from various scene graph representations. 4) **w/o Entity_Guide** underperforms compared to FineCIR, implying that entity tokens play a crucial role in guiding the multimodal composition process to effectively learn indeterminate relationships. 5) **w/o Q-Former** performs worse than FineCIR but better than **w/o SG**, indicating that Q-Former possesses superior multimodal composition capabilities, while also highlighting the indispensability of scene graphs in comprehending fine-grained modifications. 6) **w/o BBC** underperforms compared to FineCIR, implying that BBC loss is more effective in optimizing metric learning.

5.4. Case Study

To qualitatively evaluate the performance of our proposed FineCIR in fine-grained CIR scenarios, we present comparative examples against the sub-optimal CIR model SPRC [40] in Figure 4. On the Fine-FashionIQ dataset, as shown in Figure 4(a), FineCIR retrieves the target image at Top-1, while SPRC ranks it at Top-4, highlighting FineCIR’s superior understanding of fine-grained modification texts (FineMT). Similarly, on the Fine-CIRR dataset, as shown in Figure 4(b), FineCIR captures modifications like repositioning subjects and background changes, ranking the target image at Top-1, compared to SPRC’s Top-2, indicating that FineCIR has a stronger capacity to prioritize fine-grained scenarios. These qualitative results demonstrate FineCIR’s effectiveness in modeling fine-grained semantics across both fashion and open-domain datasets.

5.5. More Experiments

Due to space constraints, Appendix E compares the training and inference efficiency of FineCIR and SOTA, further demonstrating the practicality of FineCIR. Furthermore, Appendix F provides additional case studies.

6. Conclusion

In this work, we developed a fine-grained CIR data annotation pipeline and constructed two corresponding datasets that reduced imprecise positive samples and enhanced CIR systems' ability to accurately capture user modification intent. Additionally, we proposed FineCIR, the first CIR framework to explicitly parse modification semantics. Extensive experiments exhibited that FineCIR outperformed existing models in both fine-grained and traditional CIR, highlighting its superiority and generalization capability.

References

- [1] Xuefeng Bai, Yulong Chen, and Yue Zhang. Graph pre-training for amr parsing and generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 6001–6015, 2022. 8
- [2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4959–4968, 2022. 7, 8
- [3] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022. 6
- [4] Dapeng Chen, Min Wang, Haobin Chen, Lin Wu, Jing Qin, and Wei Peng. Cross-modal retrieval with heterogeneous graph embedding. In *Proceedings of the International ACM Conference on Multimedia*, pages 3291–3300. ACM, 2022. 1
- [5] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2998–3008. IEEE, 2020. 6, 7
- [6] Woo Suk Choi, Yu-Jung Heo, Dharani Punithan, and Byoung-Tak Zhang. Scene graph parsing via abstract meaning representation in pre-trained language models. In *Proceedings of the 2nd workshop on deep learning on graphs for natural language processing*, pages 30–35, 2022. 8
- [7] Yali Du, Yinwei Wei, Wei Ji, Fan Liu, Xin Luo, and Liqiang Nie. Multi-queue momentum contrast for microvideo-product retrieval. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 1003–1011. ACM, 2023. 1
- [8] Hongfei Ge, Yuanchun Jiang, Jianshan Sun, Kun Yuan, and Yezheng Liu. Llm-enhanced composed image retrieval: An intent uncertainty-aware linguistic-visual dual channel matching model. *ACM Transactions on Information Systems*, 43(2):1–30, 2025. 1, 8
- [9] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3, 7
- [10] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *Transactions on Machine Learning Research*, 2024. 1, 2, 3
- [11] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogério Schmidt Feris. Dialog-based interactive image retrieval. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 676–686. MIT Press, 2018. 1
- [12] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471, 2017. 1
- [13] Fuxiang Huang, Lei Zhang, Xiaowei Fu, and Suqi Song. Dynamic weighted combiner for mixed-modal image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2303–2311, 2024. 1
- [14] Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. Egocvr: An egocentric benchmark for fine-grained composed video retrieval. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024. 2
- [15] S Karthik, K Roth, M Mancini, Z Akata, et al. Vision-by-language for training-free compositional image retrieval. In *The International Conference on Learning Representations*, 2024. 2, 7
- [16] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2991–2999, 2024. 1
- [17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3, 4, 7
- [18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 7
- [19] Shenshen Li, Xing Xu, Xun Jiang, Fumin Shen, Xin Liu, and Heng Tao Shen. Multi-grained attention network with mutual exclusion for composed query-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1
- [20] Shenshen Li, Xing Xu, Xun Jiang, Fumin Shen, Zhe Sun, and Andrzej Cichocki. Cross-modal attention preservation with self-contrastive learning for composed query-based image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 1
- [21] Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Ghollamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. FACTUAL: A benchmark for faithful and consistent textual scene graph parsing. In *Findings of the Association for Computational Linguistics*, pages 6377–6390. Association for Computational Linguistics, 2023. 7
- [22] Han Liu, Yinwei Wei, Jianhua Yin, and Liqiang Nie. HS-GCN: hamming spatial graph convolutional networks for

- recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):5977–5990, 2023. 1
- [23] Zheyuan Liu, Cristian Rodriguez Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2105–2114. IEEE, 2021. 1, 5, 7
- [24] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5753–5762, 2024. 6, 7, 1, 8
- [25] Zheyuan Liu, Weixuan Sun, Damien Teney, and Stephen Gould. Candidate set re-ranking for composed image retrieval with dual multi-modal encoder, 2024. 3, 7, 8
- [26] Meta. The llama 3 herd of models. 2024. 4, 2
- [27] OpenAI. Gpt-4o system card, 2024. 4
- [28] Leigang Qu, Meng Liu, Wenjie Wang, Zhedong Zheng, Liqiang Nie, and Tat-Seng Chua. Learnable pillar-based re-ranking for image-text retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1252–1261. ACM, 2023. 1
- [29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 8
- [30] Chull Hwan Song, Taebaek Hwang, Jooyoung Yoon, Shunghyun Choi, and Yeong Hyeon Gu. Syncmask: Synchronized attentional masking for fashion-centric vision-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13948–13957, 2024. 1
- [31] Omkar Thawakar, Muzammal Naseer, Rao Muhammad Anwer, Salman Khan, Michael Felsberg, Mubarak Shah, and Fahad Shahbaz Khan. Composed video retrieval via enriched context and discriminative embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26896–26906, 2024. 2
- [32] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. 6
- [33] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Güл Varol. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5270–5279, 2024. 2
- [34] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Güл Varol. Covr-2: Automatic data construction for composed video retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 6, 7, 8
- [35] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval - an empirical odyssey. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6439–6448. IEEE, 2019. 2, 3, 7, 8
- [36] Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie. Comprehensive linguistic-visual composition network for image retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1369–1378. ACM, 2021. 3, 7, 8
- [37] Haokun Wen, Xuemeng Song, Jianhua Yin, Jianlong Wu, Weili Guan, and Liqiang Nie. Self-training boosted multi-factor matching network for composed image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 3, 6, 7, 8
- [38] Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. Target-guided composed image retrieval. In *Proceedings of the ACM International Conference on Multimedia*, pages 915–923, 2023. 6, 7, 8
- [39] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021. 1, 5
- [40] Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, Chun-Mei Feng, et al. Sentence-level prompts benefit composed image retrieval. In *International Conference on Learning Representations*, 2024. 2, 3, 6, 7, 8
- [41] Yahui Xu, Yi Bin, Jiwei Wei, Yang Yang, Guoqing Wang, and Heng Tao Shen. Align and retrieve: Composition and decomposition learning in image retrieval with text feedback. *IEEE Transactions on Multimedia*, 2024. 1
- [42] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 4
- [43] Zhenyu Yang, Shengsheng Qian, Dizhan Xue, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Semantic editing increment benefits zero-shot composed image retrieval. In *Proceedings of the ACM International Conference on Multimedia*, pages 1245–1254, 2024. 2, 7
- [44] Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, and Changsheng Xu. Ldre: Llm-based divergent reasoning and ensemble for zero-shot composed image retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 80–90, 2024. 2
- [45] WU Yue, Zhaobo Qi, Yiling Wu, Junshu Sun, Yaowei Wang, and Shuhui Wang. Learning fine-grained representations through textual token disentanglement in composed video retrieval. In *The International Conference on Learning Representations*, 2025. 2
- [46] Gangjian Zhang, Shikun Li, Shikui Wei, Shiming Ge, Na Cai, and Yao Zhao. Multimodal composition example mining for composed query image retrieval. *IEEE Transactions on Image Processing*, pages 1149–1161, 2024. 2

FineCIR: Explicit Parsing of Fine-Grained Modification Semantics for Composed Image Retrieval

Supplementary Material

This is the supplementary material of “FineCIR: Explicit Parsing of Fine-Grained Modification Semantics for Composed Image Retrieval”. Specifically, Section A illustrates detailed information about the difference between fine-grained CIR and the traditional CIR. Section B shows the details of dataset construction, including various prompts for the LLM we utilized in several steps. Section C presents statistics of the proposed two datasets and compares them with the original CIR datasets. Section D shows a comprehensive evaluation of the proposed FineCIR model against several significant baselines on the traditional CIR datasets FashionIQ and CIRR. In Section E we determine the computational costs of our proposed FineCIR, and make comparison with the sub-optimal model. Section F presents a case study on our proposed fine-grained CIR datasets, Fine-FashionIQ and Fine-CIRR.

A. Fine-grained CIR VS. Traditional CIR

A.1. FineMT in Solving CoarseMT’s Problems

Reducing the Imprecise Positive Samples. Ground-truth labeling in CIR datasets is frequently inadequate due to the presence of numerous visually similar images and the limited expressive capability of modification text. For a given multimodal query, there may exist candidate images that, despite subtle differences from the labeled ground truth, actually align better with the query requirements compared to the target image. In such cases, these candidate images should ideally be regarded as ground-truth. However, target images that align less effectively with query requirements are often still labeled as positive samples, leading to what we term false-positive samples.

To demonstrate the effectiveness of our proposed fine-grained datasets in reducing the imprecise positive samples, we conducted a comparative experiment between the original coarse modification text (CoarseMT) and our fine-grained modification text (FineMT), as depicted in Figure 5. In the figure, correctly attended modifications are indicated with green boxes, while incorrectly attended modifications are marked in red. For example, in Case 1, FineMT covers a more comprehensive set of subjects from multiple perspectives, explicitly including the “crab”, “plate”, “bottle”, and even the “butter”. Conversely, CoarseMT narrowly focuses only on the “crab” and “plate”, neglecting several other key objects. In situations like these, when a candidate image meets the query criteria more precisely and contains more distinguishing features than the designated

target image, the target image can inadvertently be an imprecise positive sample. By specifying more detailed and comprehensive modification requirements, FineMT significantly reduces the likelihood that suitable candidate images are incorrectly labeled as positive samples. Consequently, this ensures that the target image is accurately classified as a true-positive, effectively mitigating the risk of false-positive classifications.

Recognizing Visually Similar Images. Ground-truth labeling in CIR datasets is often insufficient due to numerous visually similar images and the limited descriptive capability of modification text. For a given multimodal query, candidate images that subtly differ from the ground-truth yet fulfill the query are often visually similar to the target image, complicating the retrieval process by making it challenging to distinguish between visually similar images.

To validate the advantages of our proposed Fine-FashionIQ and Fine-CIRR datasets in differentiating visually similar samples, we selected BLIP4CIR [24], a straightforward baseline incorporating multimodal query features for retrieving target images. Experiments were conducted comparing FineMT and the original CoarseMT scenarios. Figures 6 and 7 illustrate the top-5 retrieval results in both scenarios, highlighting target images in green.

In the fashion-domain dataset Fine-FashionIQ, Figure 6(a) (bottom) shows that although the top-5 retrieved results satisfy the multimodal query, they are incorrectly classified as negative samples. This occurs because the CoarseMT failed to mention the presence of “a necklace” in the target image. After applying FineMT relabeling (top), this detail is captured, correctly identifying the target image as the only positive sample, significantly improving discrimination among visually similar samples. A similar improvement is observed in Figure 6(b), where the target image’s ranking improves from second to first place after FineMT relabeling.

In the open-domain dataset Fine-CIRR, Figure 7(a) demonstrates that the CoarseMT includes basic descriptions (“side angle view on buffalo”, “in pond”, and “sharp horns”), whereas FineMT introduces additional crucial details (“standing in the water”, “dirt path”, and “trees in the background”). The enhanced FineMT query accurately retrieves the target image, effectively reducing confusion from visually similar images. Likewise, Figure 7(b) shows the target image’s retrieval rank improving from fifth to first after FineMT relabeling. These results confirm that FineMT’s detailed multimodal queries successfully differ-

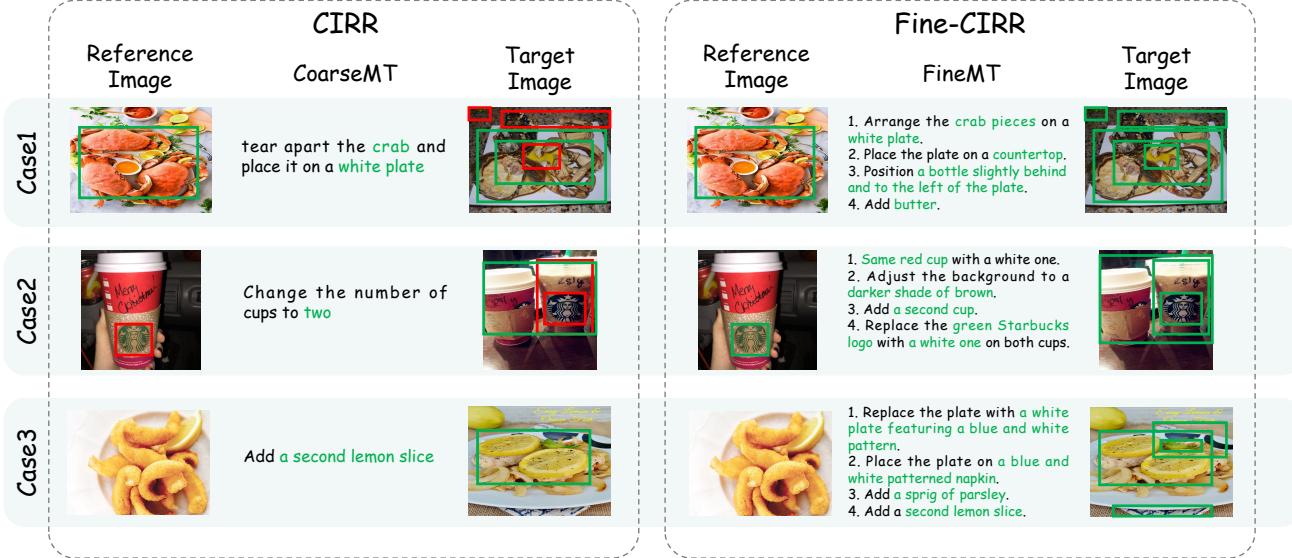


Figure 5. Capability of reducing the imprecise positive samples by using the fine-grained modification text (FineMT).

entiate visually similar samples, converting previously ambiguous cases into true negatives.

A.2. Comparison of Features

As shown in Figure 8, we present the comparison between coarse-grained CIR and fine-grained CIR. The table highlights key differences in their ability to handle modification requirements and user intent. Coarse-grained CIR typically relies on brief, one-to-two-sentence modification texts, limiting its ability to express complex or nuanced requirements. This results in weaker support for multi-dimensional modifications and lower certainty in modeling precise user needs, making it harder to distinguish visually similar images. Additionally, the retrieval accuracy is relatively low, which reduces overall user satisfaction.

In contrast, fine-grained CIR employs detailed modification descriptions, often comprising multiple sentences, to express multi-level and multi-dimensional modifications more effectively. This approach reduces the rate of imprecise positive samples, enabling the model to better adhere to specific modification requirements. Fine-grained CIR demonstrates stronger certainty in capturing user intent, significantly improving retrieval accuracy by providing results that better align with user expectations. Consequently, it enhances the user experience by quickly finding images that meet the desired specifications, reducing the need for iterative adjustments and increasing overall satisfaction.

B. Details of Dataset Construction

In subsection B.1, we provided a more detailed analysis of prompts, including the prompts for LLM in the *Data Selection stage*, the *Dataset Construction stage*, and the *Quality Check stage*. In subsection B.2, we analyzed the detailed prompts used for FineMT generation during the process of BLIP-3, which is the key step in the construction process. And in subsection B.3, we reported the changes in the query number during the construction of our proposed Fine-FashionIQ and Fine-CIRR.

B.1. Analysis of Prompts for Check

Figure 9 illustrates our designed prompt used in the “MLLM & Human Check” step. Here, the LLM analyzes reference-target image pairs based on three provided questions, returning “Yes” or “No” answers. These answers serve as critical assessments for subsequent evaluation. For instance, as shown, the LLM examines the reference and target images and answers <Yes, Yes, Yes>. That is because the target can be obtained by modifying details such as the background and subject of the reference image, and the two images are highly related, being dogs of similar breeds, so this pair is well suited for constructing fine-grained datasets.

Figure 10 displays the designed prompt for the “Query-guided Refinement” step. Since inputting FineMT alone may omit context regarding the modified objects, we input the entire multimodal query, including the reference image and FineMT, into Llama3.2 [26]. This comprehensive input enables Llama3.2 to accurately interpret modifications in relation to the reference image, thereby ensuring FineMT’s accuracy and eliminating any irrelevant or hallucinated content. In the provided example, the sentence “Show a brown dog in a similar stance and posture” solely describes the tar-

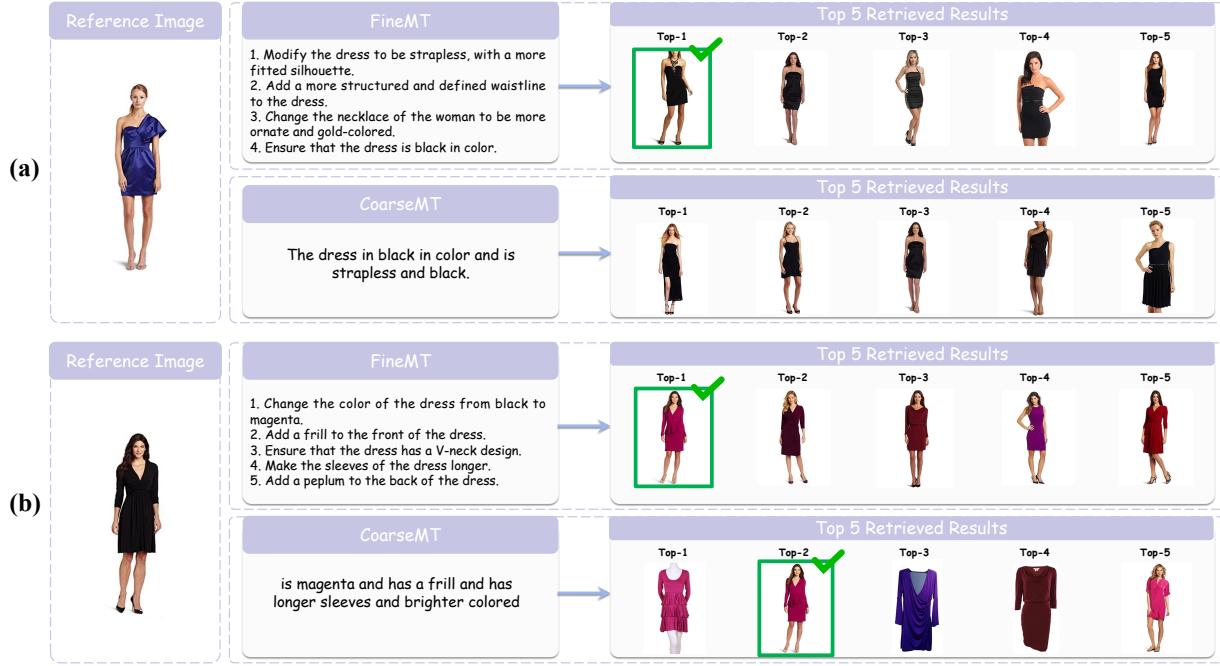


Figure 6. The ability to distinguish between visually similar samples using FineMT. We showed the top 5 retrieved results on both Fine-FashionIQ and FashionIQ. The target images are framed in green.



Figure 7. The ability to distinguish between visually similar samples using FineMT. We showed the top 5 retrieved results on both Fine-CIRR and CIRR. The target images are framed in green.

Feature	Conventional Coarse-grained CIR	vs	Fine-grained CIR
Length of modification text	Short, typically 1-2 sentences		Detailed, with multiple sentences describing modification details
Ability to express modification requirements	Weak, limited to simple modifications		Strong, capable of expressing multi-level and multi-dimensional modifications
Rate of imprecise positive samples	High, with minimal adverse impact on model training		Low, enhancing the model's ability to strictly follow modification requirements
Requirements modelling certainty	Weak, hard to distinguish visually similar images		Strong, capable of precisely capturing users' modification intents
Retrieval accuracy	Low, requiring users to repeatedly filter results		High, better satisfying users' specific modification requirements
User experience	Requires multiple adjustments and modifications		Quickly finds images that meet requirements, improving satisfaction

Figure 8. Comparison of the characteristics of coarse-grained CIR and fine-grained CIR.

get image, making it irrelevant to the reference image and thus a candidate for removal.

Figure 11 demonstrates our prompt used in the ‘‘FineMT Compress’’ step, which aims to refine textual descriptions by minimizing redundancy while retaining key information. The prompt guides GPT-4o [27] in summarizing lengthy FineMT descriptions exceeding 77 tokens. This ensures essential modification details are maintained, eliminating excessive prepositions and unnecessary elaboration. The provided example highlights the contrast between verbose input and the refined output, simplifying overly detailed expressions such as ‘‘to enhance its design and make it more visually appealing’’ to concise phrases like ‘‘to the top half of the dress’’. This compression effectively balances information retention and conciseness, thus enhancing the quality of FineMT.

B.2. Analysis of Prompts for FineMT Generation

In Figure 12, we present the prompts employed to generate foundational texts for FineMT using BLIP-3 [42]. Our requirement was that BLIP-3 outputs include additional details present in the image pairs but omitted in the original CoarseMT. For instance, while the original CoarseMT might only reference a single object, the target image typically reflects modifications across multiple objects relative to the reference image. Thus, FineMT provides a comprehensive and detailed description of the modifications, addressing the shortcomings of the original CoarseMT.

or a thorough analysis, we compare generated texts obtained using various designed prompts for BLIP-3, as illustrated in Figure 12. In this comparison, prompt (a), which lacks evaluation by an LLM, generates texts limited in detail, failing to capture comprehensive modifications. Con-

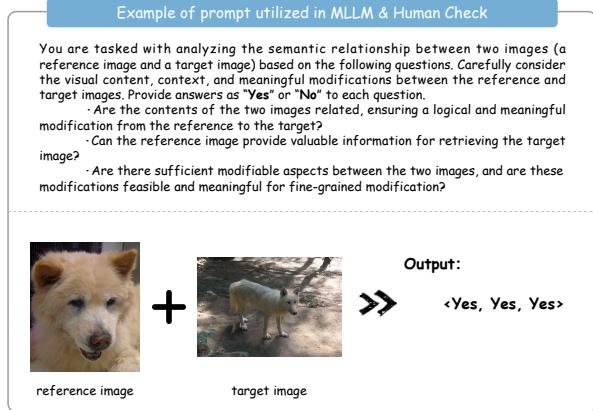


Figure 9. Example of prompt utilized in MLLM&Human Check.

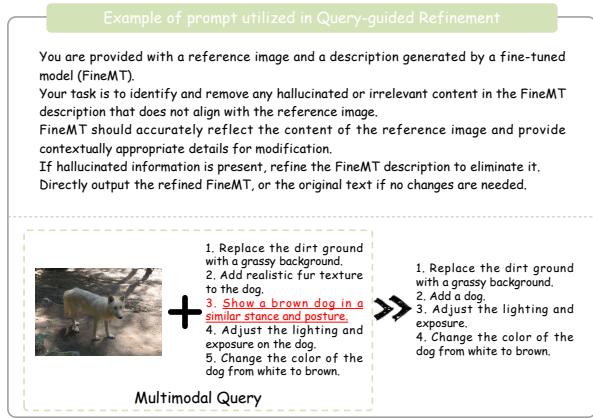


Figure 10. Example of prompt utilized in Query-guided Refinement.

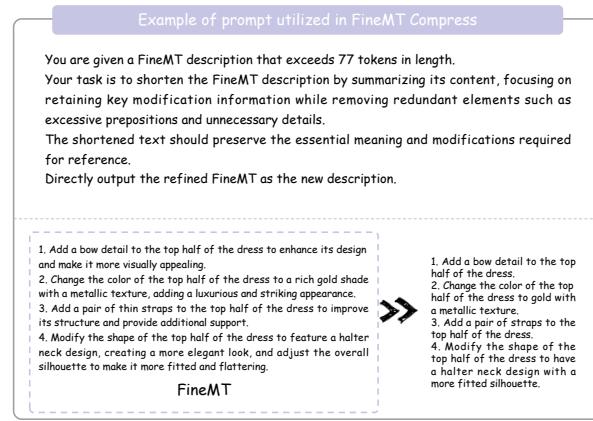


Figure 11. Example of prompt utilized in FineMT Compress.

versely, prompts (b) and (c), aided by LLM evaluation, successfully capture and elaborate on fine-grained modifications from diverse perspectives. This demonstrates the effectiveness of incorporating LLM evaluation to enhance the

accuracy and completeness of captured fine-grained details.

Furthermore, texts generated using prompt (c) were notably more precise and comprehensive, addressing various perspectives and details effectively. Therefore, we selected prompt (c) as the optimal approach for generating fine-grained texts.

Recognizing differences between the fashion-domain data and real-life scenarios, we employed specifically designed additional prompts for Fine-FashionIQ. Given the original FashionIQ dataset's focus on garment details, such as the presence or absence of straps and sleeve lengths—these tailored prompts guide BLIP-3's output towards meticulous attention to garment components, holistic interpretation of clothing items, and comparative analysis between reference and target images. This meticulous approach enables the generation of highly detailed modification descriptions.

In contrast, Fine-CIRR presents a more complex challenge due to the open-domain nature of the original CIRR dataset, typically involving multiple objects and background elements. This complexity underscores the essential role of FineMT, whose detailed descriptions precisely capture the genuine modifications present in the reference and target images, further validating the need for enhanced granularity.

B.3. Query Number Changes throughout the Construction

Furthermore, since the fine-grained datasets we constructed have a large number of checking and filtering steps, we statistically recorded the changes in the query number throughout the dataset construction process and presented it as line graphs, as shown in Figure 13 and Figure 14.

C. Traditional and Fine-grained CIR Datasets

In subsection C.1, we present our proposed datasets and their underlying datasets, FashionIQ [39] and CIRR [23]. In subsection C.2, we compared and analyzed the statistics of the datasets.

C.1. Dataset Details

To evaluate the validity of the fine-grained modification text, we constructed two fine-grained datasets for CIR. We now describe each dataset in detail as follows,

- **Fine-FashionIQ** is based on the classic CIR dataset, **FashionIQ** [39], whose content belongs entirely to the fashion domain. It consists of 77,684 images which are divided into three categories: *Dresses*, *Shirts*, and *Tops&Tees*. Following the FashionIQ, we treat it as three independent datasets. And there are 18K triples for training and ~6K triples for testing. Based on it, after our detailed filtering and constructing process, we finally uti-

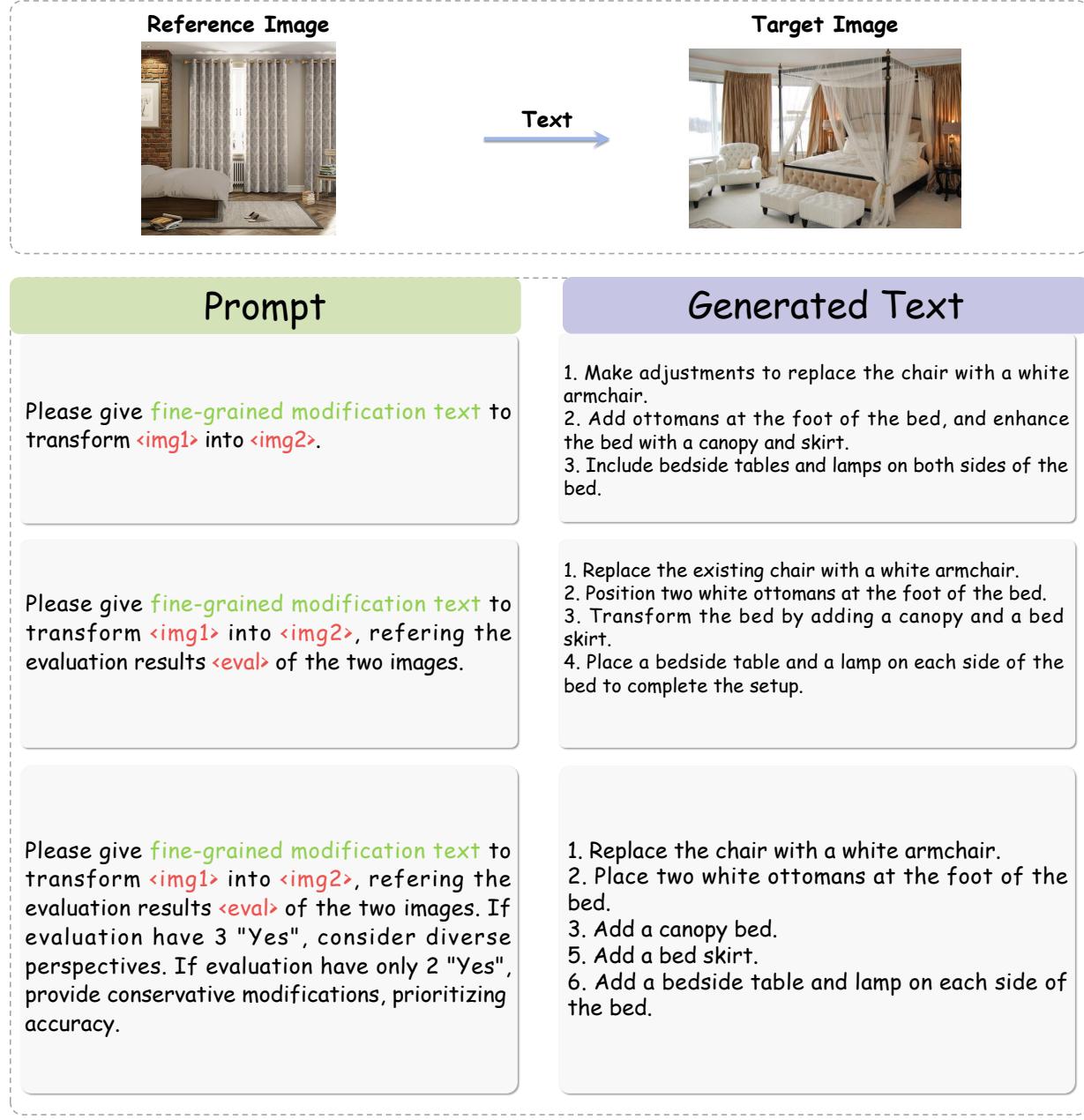


Figure 12. Generated texts using various prompts for BLIP-3.

	Domain	Granularity	Train triplets	Test triplets	MT length*	Maximal tokens
FashionIQ	fashion	coarse-grained	18000	6016	24.7	37.0
FineFashionIQ	fashion	fine-grained	14812	4285	59.2	77.0
CIRR	open	coarse-grained	28225	4181	12.8	50.0
FineCIRR	open	fine-grained	15344	3808	62.9	77.0

Table 4. Statistics of fine-grained datasets for CIR compared to two classic CIR datasets, FashionIQ and CIRR. MT length* denotes the average length of modification texts in the dataset.

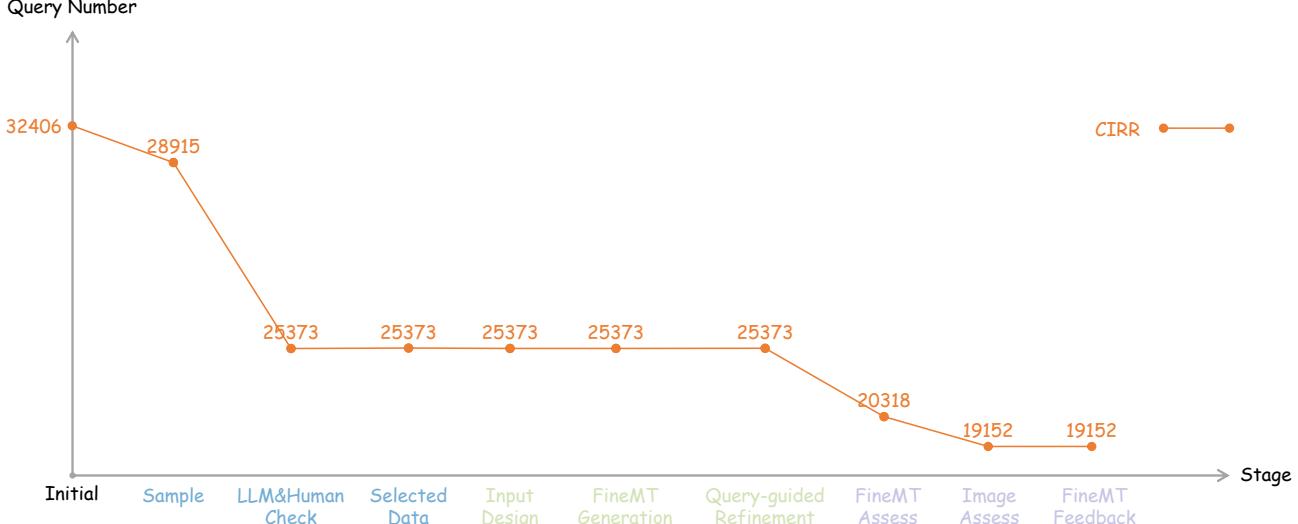


Figure 13. Changes in the query number throughout the construction of Fine-CIRR.

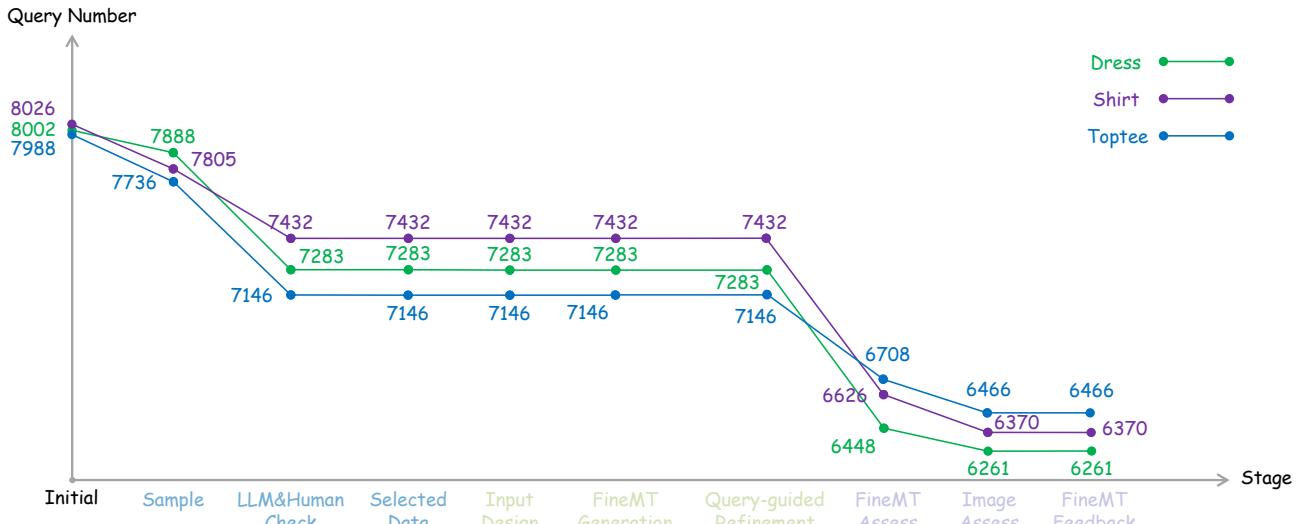


Figure 14. Changes in the query number throughout the construction of Fine-FashionIQ.

lized 14812 and 4285 triplets for training and testing on Fine-FashionIQ.

- **Fine-CIRR** is based on the classic open-domain CIR dataset, **CIRR** [23]. It contains 21,552 real images taken from the renowned language reasoning dataset NLVR², which is well-known for its natural language reasoning applications. CIRR contains a total of 36.5 K triples, of which 80% are used for training, 10% for validation and 10% for testing. Specifically, similar to Fine-FashionIQ, we utilizing a detailed process and finally obtained 15344 and 3808 triplets for training and testing on Fine-CIRR. In addition, following the CIRR, the fine-grained dataset Fine-CIRR includes a specialized subset designed for fine discrimination. This subset focuses on negative images

that exhibit a high degree of visual similarity and is utilized to assess the model’s performance in distinguishing false-negative images.

C.2. Statistics

We provide detailed statistics for Fine-FashionIQ and Fine-CIRR, comparing them with their original counterparts, FashionIQ and CIRR. Since previous annotations led to portions of the data being unsuitable for constructing fine-grained datasets, we performed iterative checks and filtration, resulting in a streamlined dataset with fewer but higher-quality queries. Regarding the length of modification texts, FineMT inherently contains finer-grained descriptions, thus resulting in increased length compared to

Method	Year	FashionIQ										CIRR									
		Dresses		Shirts		Tops&Tees		Avg		R@k				Rsubset@k			Avg				
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	k=1	k=5	k=10	k=50	k=1	k=2	k=3					
TIRG [35]	2019	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39	14.61	48.37	64.08	90.03	22.67	44.97	65.14	35.52				
CLVC-Net [36]	2021	29.85	56.47	28.75	54.76	33.50	64.00	30.70	58.41	-	-	-	-	-	-	-	-				
CLIP4CIR [2]	2022	33.81	59.40	39.99	60.45	41.41	65.37	38.40	61.74	38.53	69.98	81.86	95.93	68.19	85.64	94.17	69.09				
TG-CIR [38]	2023	45.22	69.66	52.60	72.52	56.14	77.10	51.32	73.09	45.25	78.29	87.16	97.30	72.84	89.25	95.13	75.57				
BLIP4CIR [24]	2024	40.65	66.34	40.38	64.13	46.86	69.91	42.63	66.79	40.17	71.81	83.18	95.69	72.34	88.70	95.23	72.08				
CoVR-2 [34]	2024	46.53	69.60	51.23	70.64	52.14	73.27	49.96	71.17	50.43	81.08	88.89	98.05	76.75	90.34	95.78	78.92				
Candidate [25]	2024	48.14	71.34	50.15	71.25	55.23	76.80	51.17	73.13	50.55	81.75	89.78	97.18	80.04	91.90	96.58	80.90				
SPRC [40]	2024	49.18	72.43	55.64	73.89	59.35	78.58	54.92	74.97	51.96	82.12	89.74	97.69	80.65	92.31	96.60	81.39				
LIMN [37]	2024	50.72	74.52	56.08	77.09	60.94	81.85	55.91	77.82	43.64	75.37	85.42	97.04	69.01	86.22	94.19	72.19				
IUDC [8]	2025	35.22	61.90	41.86	63.52	42.19	69.23	39.76	64.88	-	-	-	-	-	-	-	-				
FineCIR (Ours)		53.15	76.95	63.98	81.35	62.45	82.98	59.86	80.43	52.75	82.96	90.84	98.27	80.29	92.39	96.70	81.63				

Table 5. Performance comparison on original FashionIQ and CIRR relative to R@K(%). The overall best results are in bold, while the best results over baselines are underlined. The Avg metric in CIRR denotes $(R@5 + R_{subset}@1) / 2$.

the original modification texts, as demonstrated in Table 4. Nevertheless, we controlled the length during dataset construction to avoid text overpowering the visual information and weakening image relevance. Additionally, controlling text length ensures compatibility with the CLIP [29] encoder, which has a maximum input token length limitation of 77 tokens.

D. Detailed Comparison on Traditional CIR Datasets

As illustrated in Table 5, we conduct a comprehensive evaluation of the proposed FineCIR model against several significant baselines on the traditional CIR datasets, *i.e.*, FashionIQ and CIRR. From the results presented in the table, we can obtain the following observations. **1)** Our proposed FineCIR achieves outstanding performance on both FashionIQ and CIRR, demonstrating its ability to capture essential modification semantics even in coarse-grained modification text. **2)** Compared to the baselines, FineCIR exhibits significant performance advantages on both fashion-domain and open-domain datasets. This may be attributed to the superior domain generalization ability of our designed modules. **3)** Notably, FineCIR achieves better results on fine-grained CIR datasets compared to traditional CIR datasets, whereas previous baselines (*e.g.*, LIMN, BLIP4CIR) exhibit significant performance drops on fine-grained CIR. This suggests that prior methods struggle to capture fine-grained modification semantics, leading to the lack of critical modification details and performance degradation. In contrast, FineCIR demonstrates superior generalization capability in capturing modification semantics across both coarse and fine granularities.

E. Computation Cost Analysis

We determine the computational costs of our proposed FineCIR compared to the representative CIR models, Candidate [25] and SPRC [40], on the Fine-CIRR dataset. Specifically, we choose the train time, test time, and GPU memory for evaluation, as shown in Table 6. All experi-

ments are performed on a single NVIDIA A40 GPU. The train time describes the time it takes for the model to optimize to the optimum, while the test time is the time it takes for the inference on one sample. From the results presented in Table 6, we observe that FineCIR’s **w/o SG** setting obtains optimal training and testing efficiency, and it has a higher recall than Candidate and SPRC. Importantly, this study prioritizes enhanced retrieval accuracy through improved model architecture and more effective input information handling, rather than targeting scenarios where speed is prioritized over precision. Consequently, we accept a modest increase in computational overhead to gain up to an 11.4% absolute recall improvement compared to SPRC on fine-grained CIR benchmarks.

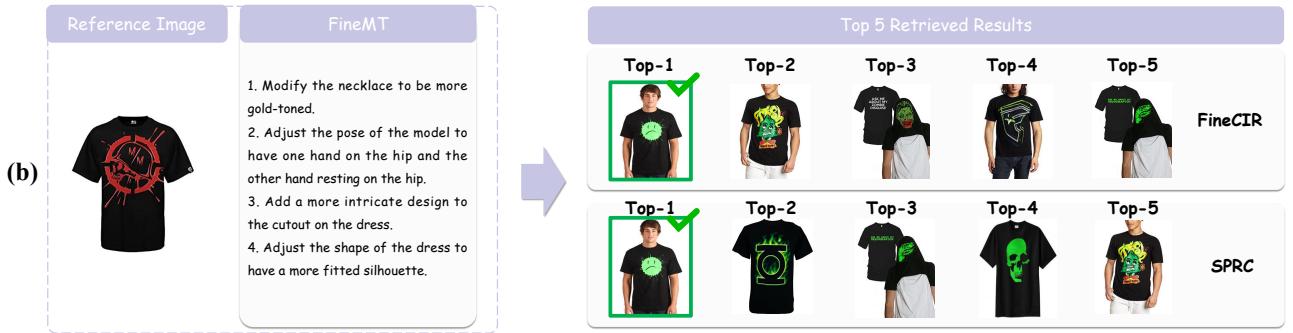
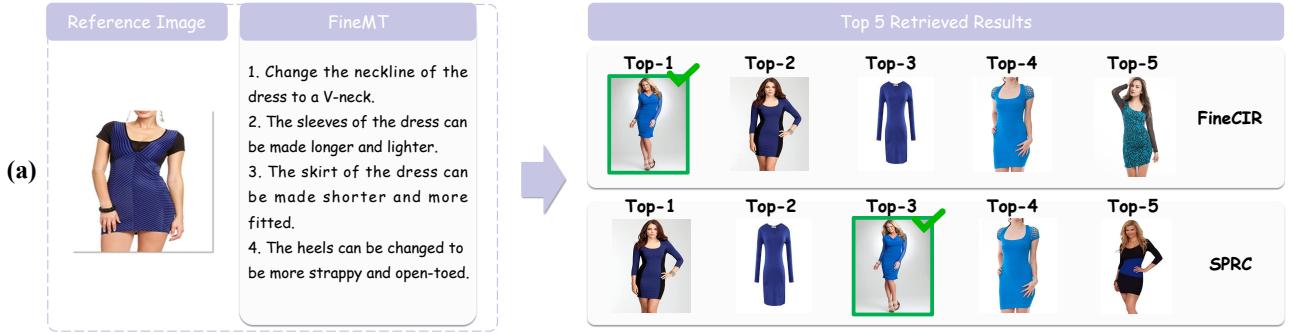
Method	Backbone	Train ↓	Test ↓	Memory ↓	Avg.R ↑
Candidate	BLIP	16h	16.7ms/sample	47.3G	78.09
SPRC	BLIP-2	4.5h	17.7ms/sample	25.2G	80.97
FineCIR w/o SG	BLIP-2	2.7h	15.6 ms/sample	25.4G	81.12
FineCIR (Ours)	BLIP-2	2.8h	23.1 ms/sample	28.9G	84.73

Table 6. Comparison on computation cost. The better results are in bold.

F. Case Study

To intuitively validate the performance of our proposed FineCIR on fine-grained scenarios of CIR, we present several examples demonstrating FineCIR’s retrieval results to this scenario, along with the comparison to sub-optimal CIR model SPRC [40], as shown in Figure 15. In the Fine-FashionIQ dataset, Figure 15(a) shows that FineCIR accurately retrieves the target image at Top-1 by capturing fine-grained modifications to the dress, including changes to the neckline, sleeves, skirt, and heels, while SPRC ranks the target at Top-3, reflecting a less precise understanding of the modifications. Similarly, in Figure 15(b), FineCIR retrieves the target image at Top-1 for modifications involving the necklace, pose, and design details of a shirt, whereas SPRC ranks it at Top-1 too, demonstrating both models’ ability to prioritize subtle changes. On the Fine-CIRR dataset, Figure 15(c) illustrates FineCIR’s success in retrieving the tar-

Fine-FashionIQ



Fine-CIRR

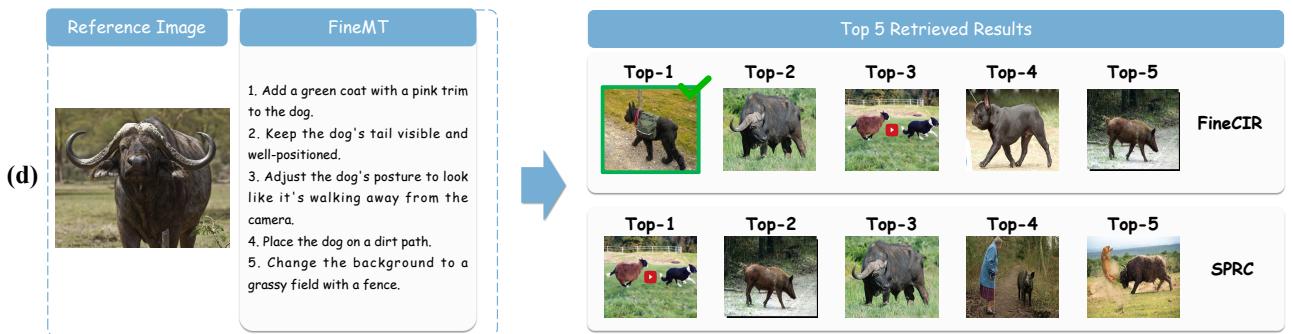
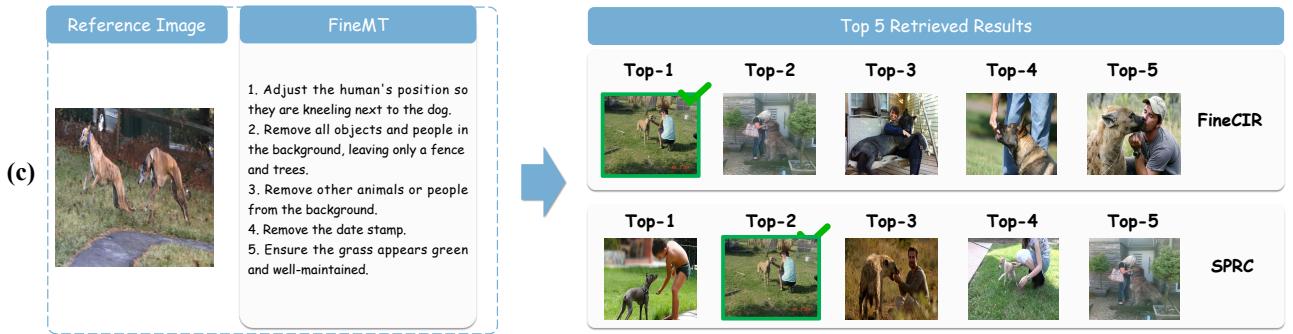


Figure 15. Qualitative examples of our proposed FineCIR compared to the sub-optimal CIR model SPRC.

get image at Top-1 based on modifications like repositioning a person and refining background elements, while SPRC ranks the target at Top-2. In Figure 15(d), FineCIR also ranks the target at Top-1 by accurately addressing modifications to the dog’s position, posture, and background, in con-

trast to SPRC’s out of Top 5 ranking. These results demonstrate that FineCIR consistently outperforms SPRC by effectively capturing and prioritizing fine-grained semantics of modifications across both fashion and open-domain CIR tasks.