Mixture of Attribute-Aware Attention Experts for Fine-grained E-Commerce Composed Image Retrieval

Yufei Ma[*], **Zihan Liang**[*], **Zhipeng Qian**[*], **Huangyu Dai**
Lingtao Mao, **Ben Chen**[†], **Chenyi Lei**[†], **Wenwu Ou**

Kuaishou Technology
✉benchen4395@gmail.com

Abstract

Composed Image Retrieval (CIR) enables users to express modification-based search intent by combining a reference image with manipulation text. Despite its promise for e-commerce, existing methods ignore readily available item textual descriptions, suffer from multi-attribute confounding where samples exhibit uncontrolled variations in unrelated attributes, and lack explicit fine-grained attribute modeling. We propose **MoA^{3}CIR** , featuring an asymmetric five-tower architecture that incorporates item descriptions on both query and target sides, a self-reflective data augmentation pipeline generating controlled attribute modifications through LVLM-driven refinement, and a Mixture of Attribute-Aware Attention (MoA^3) mechanism with grouped multi-head attention and dynamic routing coupled with a two-phase training strategy. Besides, we contribute ECCIR, the first dual-modal e-commerce CIR benchmark with 110K samples featuring authentic item descriptions. Extensive experiments demonstrate that **MoA^{3}CIR** achieves SoTA performance on ECCIR, FashionIQ, and Shoes datasets, with substantial improvements over existing methods.

1 Introduction

Composed Image Retrieval (CIR) addresses a practical challenge, enabling users to express specific search intent by combining a reference image with manipulation text. Instead of relying solely on text descriptions or image queries, CIR allows users to specify modifications such as "same chair but in blue" or "this dress with floral patterns", which is particularly valuable for e-commerce search where precise intents are difficult to express through a single modality.

Despite its promise, existing CIR methods [2, 4, 8, 35, 28] face three critical limitations that hinder their real-world deployment. **First**, current approaches typically process only the reference image and manipulation text on the query side while treating target items as purely visual inputs. However, in real e-commerce scenarios, items are invariably accompanied by textual descriptions containing essential non-visual attributes, such as material ("100% cotton"), functionality ("machine washable"), or size specifications ("XS-XXL"), that cannot be inferred from images alone or generated by LVLMs. In CIR, these non-visual attributes become critical as users may explicitly or implicitly require retrieved items to preserve certain ones. Ignoring this readily available textual information creates a fundamental gap between CIR research and real-world applications. Moreover, existing datasets like FashionIQ [43], while pioneering, suffer from quality issues [12] including ambiguous manipulation texts and noisy annotations (e.g., "is not clothing" for a dress description). **Second**,

[*]Equal Contribution.
[†]Corresponding Author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

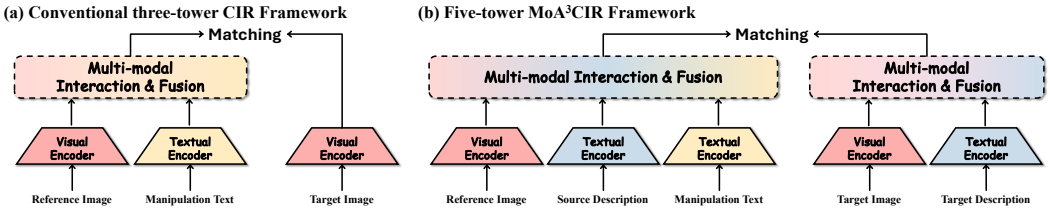

Figure 1: Comparison between conventional three-tower CIR framework (a) and our five-tower **MoA^{3}CIR** framework (b).

conventional contrastive learning struggles with fine-grained attribute learning due to uncontrolled attribute variations in training samples. Negative samples differ from queries in multiple attributes simultaneously, while positive samples also exhibit uncontrolled variations beyond the manipulation text. For instance, when the manipulation specifies "change to red color," negative samples may also vary in pattern or material, and positive samples may still differ in sleeve length or style, making it difficult to learn precise attribute correspondences. **Third**, most methods adopt coarse representations that encode the reference image and manipulation text into a single embedding without explicitly modeling fine-grained attributes such as color, material, texture, or pattern. This prevents models from capturing precise attribute-level changes specified in the manipulation text, limiting their performance.

To address the above limitations, we propose **MoA^{3}CIR** (**M**ixture **o**f **A**ttribute-**A**ware **A**ttention Experts for Fine-grained E-Commerce **C**omposed **I**mage **R**etrieval), featuring three key innovations. **First**, we incorporate item textual descriptions to capture comprehensive semantic information. We introduce a three-tower query-side encoder for the reference image, source description, and manipulation text, along with a two-tower target-side encoder that processes both target images and their descriptions. The overall five-tower architecture is illustrated in Figure 1(b), contrasted with the conventional three-tower design in Figure 1(a). To validate this design and bridge the research-practice gap, we construct the **E-C**ommerce **CIR** (ECCIR) Dataset, the first dual-modal e-commerce CIR benchmark where both query and target sides contain realistic textual information, better aligning with practical scenarios. **Second**, we propose a self-reflective fine-grained data augmentation pipeline leveraging large vision-language models (LVLMs) to construct attribute-specific training pairs. Unlike conventional augmentation, our approach employs LVLM-driven self-reflective loops to generate images with only targeted attribute modifications while preserving all other characteristics. This controlled generation strategy naturally produces hard negatives that differ solely in the manipulated attributes, providing clean supervision for fine-grained attribute learning. **Third**, to enable fine-grained attribute understanding, we design the MoA^3 (**M**ixture **o**f **A**ttribute-**A**ware **A**ttention) mechanism. MoA^3 employs grouped multi-head attention where each group acts as an attribute expert. Learnable attribute query tokens compute dynamic routing weights based on the manipulation text, selectively activating relevant experts to achieve precise attribute-level modeling. We further adopt a two-phase training strategy: first train on augmented data with AWPSE loss to specialize attribute experts, then optimize on real data with InfoNCE loss to achieve holistic retrieval alignment, thereby decoupling attribute specialization from global semantic optimization. Our main contributions are summarized as follows:

- We introduce **ECCIR**, the first dual-modal e-commerce CIR dataset with paired image-text data on both query and target sides, offering high annotation quality and better alignment with real-world scenarios.
- We propose a self-reflective data augmentation pipeline that generates controlled attribute modifications through LVLM-driven loops, contributing to effective fine-grained attribute learning.
- We design **MoA^{3}CIR**, featuring the MoA^3 mechanism that enables explicit attribute-level modeling through grouped multi-head attention with learnable attribute tokens and dynamic routing.

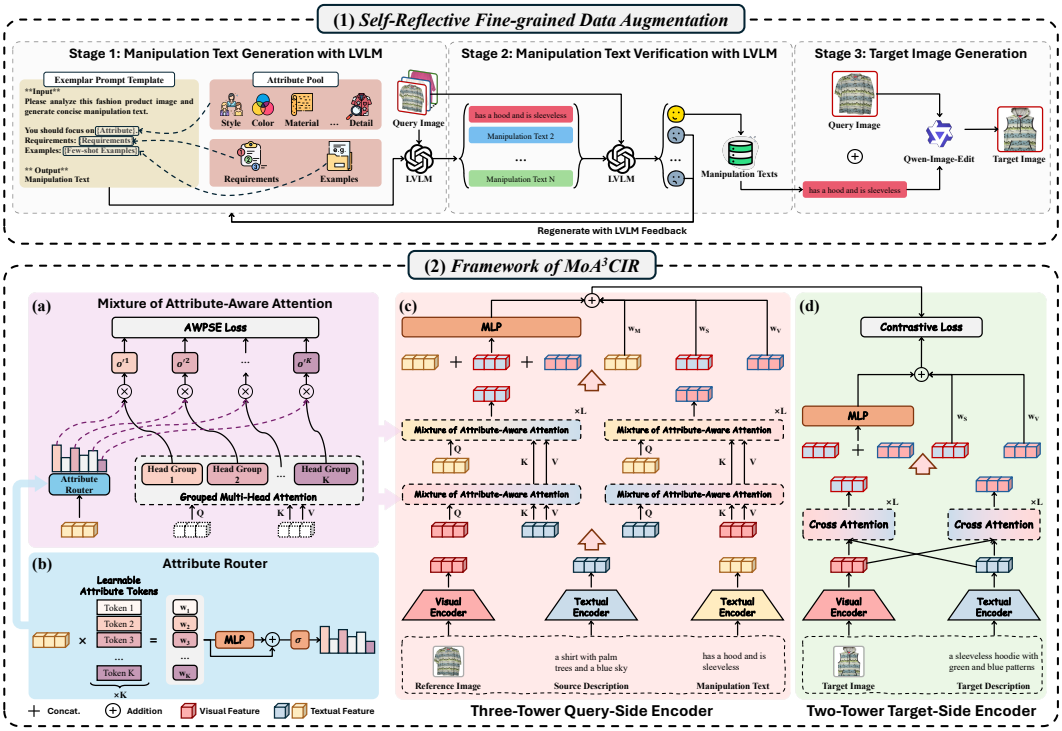

Figure 2: Our proposed self-reflective data augmentation pipeline (1) and the framework of **MoA^{3}CIR** (2).

2 Method

2.1 Preliminary

Task Definition. We formulate e-commerce CIR as follows. Given a hybrid-modality query $Q = \{I_q, D_q, M\}$ consisting of a reference image I_q, a source description D_q, and a manipulation text M specifying the desired attribute modifications, and a candidate set $\mathcal{D} = \{(I_t^j, D_t^j)\}_{j=1}^{N_\mathcal{D}}$ of $N_\mathcal{D}$ items where each item is associated with both a target image I_t and a target description D_t, the goal is to retrieve the top-k items from $\mathcal{D}$ that best match the user's modification intent, with $k \ll N_\mathcal{D}$. This formulation differs from conventional CIR in two key aspects: both query and target sides incorporate authentic textual descriptions, and the manipulation typically targets a small set of specific attributes (e.g., color, material, style) rather than arbitrary scene-level changes.

Challenges. The above formulation poses three challenges that prior CIR methods fail to address jointly. *(C1) Utilization of authentic item descriptions:* how to effectively incorporate query- and target-side descriptions to capture non-visual attributes (e.g., material, brand, function) that cannot be inferred from images alone. *(C2) Multi-attribute confounding under contrastive learning:* real item pairs differ along many uncontrolled axes simultaneously, causing standard contrastive losses to learn entangled rather than attribute-specific representations. *(C3) Fine-grained attribute modeling:* existing fusion modules collapse the query into a single embedding without explicit attribute decomposition, making them blind to the distinction between, e.g., "change to red" and "change to blue". These three challenges motivate the three components of **MoA^{3}CIR**: a five-tower architecture coupled with the ECCIR benchmark for (C1), a self-reflective augmentation pipeline with two-phase training for (C2), and the MoA^3 mechanism for (C3).

2.2 Framework Overview

MoA^{3}CIR adopts an asymmetric five-tower design comprising a three-tower query-side encoder that processes reference image, source description, and manipulation text, and a two-tower target-side

encoder that fuses target image and description. Both sides are projected into a shared embedding space, as illustrated in Figure 2 (2). Training uses AWPSE loss for attribute specialization and InfoNCE loss for holistic alignment.

2.3 Query-Side Encoder

We adopt CLIP-based encoders to extract token-level representations from the reference image I_q, source description D_q, and manipulation text M, yielding visual patch tokens $\mathbf{V}_q \in \mathbb{R}^{L_{v_q} \times d}$, source description tokens $\mathbf{D}_q \in \mathbb{R}^{L_{d_q} \times d}$, and manipulation text tokens $\mathbf{M} \in \mathbb{R}^{L_m \times d}$, where L_{v_q}, L_{d_q}, L_m denote sequence lengths and d is the feature dimension.

To capture fine-grained correspondences between manipulation text and target attributes, we stack L MoA^3 layers, each containing two MoA^3 modules that process visual and textual sequences. For visual tokens $\mathbf{V}_q^{(\ell-1)}$ at layer ℓ, processing proceeds in two steps:

Visual-Textual Cross-Modal Fusion. Visual tokens attend to source-description tokens:

$$\mathbf{V}_q^{'(\ell)} = \text{MoA}^3(\mathbf{V}_q^{(\ell-1)}, \mathbf{D}_q^{(\ell-1)}). \tag{1}$$

Manipulation Injection. Enhanced visual tokens are then guided by manipulation text:

$$\mathbf{V}_q^{(\ell)} = \text{MoA}^3(\mathbf{V}_q^{'(\ell)}, \mathbf{M}). \tag{2}$$

so that visual queries selectively integrate the attribute changes specified by $\mathbf{M}$. Source description tokens undergo the same two-step processing:

$$\begin{aligned} \mathbf{D}_q^{'(\ell)} &= \text{MoA}^3(\mathbf{D}_q^{(\ell-1)}, \mathbf{V}_q^{(\ell-1)}), \\ \mathbf{D}_q^{(\ell)} &= \text{MoA}^3(\mathbf{D}_q^{'(\ell)}, \mathbf{M}). \end{aligned} \tag{3}$$

After L layers, we aggregate attribute-enhanced tokens through mean pooling:

$$\mathbf{v}_q, \mathbf{d}_q, \mathbf{m} = \text{MeanPool}(\mathbf{V}_q^{(L)}, \mathbf{D}_q^{(L)}, \mathbf{M}), \tag{4}$$

where $\mathbf{v}_q, \mathbf{d}_q, \mathbf{m} \in \mathbb{R}^d$. These are fused to form final query representation by hybrid aggregation:

$$\mathbf{c}_q = \text{MLP}([\mathbf{v}_q; \mathbf{d}_q; \mathbf{m}]) + w_{v_q}\mathbf{v}_q + w_{d_q}\mathbf{d}_q + w_m\mathbf{m}, \tag{5}$$

where $w_{\{v_q, d_q, m\}}$ are learnable weights and $[\cdot;\cdot;\cdot]$ denotes concatenation.

2.4 Mixture of Attribute-Aware Attention

The MoA^3 mechanism provides explicit attribute-aware fusion through *grouped multi-head attention* and *learnable attribute query tokens* that yield dynamic routing weights.

Grouped Multi-Head Attention (GMA). As shown in Figure 2 (2) (a), we partition H attention heads into K groups, each with $N_h{=}H/K$ heads. Given query tokens $\mathbf{X}_q \in \mathbb{R}^{L_q \times d}$ and key/value tokens $\mathbf{X}_{kv} \in \mathbb{R}^{L_{kv} \times d}$, the i-th group uses independent projections $\mathbf{W}_Q^i, \mathbf{W}_K^i, \mathbf{W}_V^i \in \mathbb{R}^{d \times (N_h d_h)}$ with $d_h{=}d/H$, computing:

$$\mathbf{o}^i = \text{Attn}_i\big(\mathbf{X}_q\mathbf{W}_Q^i,\, \mathbf{X}_{kv}\mathbf{W}_K^i,\, \mathbf{X}_{kv}\mathbf{W}_V^i\big), \tag{6}$$

where Attn$_i$ is scaled dot-product attention over the N_h heads of group i (followed by a group-specific output projection) and $\mathbf{o}^i$ represents attribute feature of group i. Treating each group as an *attribute expert*, we perform soft routing with activation weights $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$:

$$\mathbf{Y} = \big[\alpha_1\mathbf{o}^1; \ldots; \alpha_K\mathbf{o}^K\big]\, \mathbf{W}_{\text{proj}}, \tag{7}$$

where $\mathbf{W}_{\text{proj}}$ projects the concatenated outputs to dimension d, and $\mathbf{Y} \in \{\mathbf{V}_q^{'(\ell)}, \mathbf{V}_q^{(\ell)}, \mathbf{D}_q^{'(\ell)}, \mathbf{D}_q^{(\ell)}\}$ denotes enhanced visual and textual features. This *soft* routing yields stable gradients and allows multiple experts to contribute jointly.

Learnable Attribute Query Tokens. To compute $\boldsymbol{\alpha}$, we introduce learnable attribute tokens $\mathbf{Q}_{\text{attr}}=\{\mathbf{q}_1,\ldots,\mathbf{q}_K\} \in \mathbb{R}^{K\times d}$ that act as semantic probes (shown in Figure 2(2)(b)) and each corresponds to one expert in GMA. From manipulation tokens $\mathbf{M}$, we compute $\bar{\mathbf{m}}=\text{MeanPool}(\mathbf{M})$ and obtain two signals, namely a similarity-based path:

$$s_i = \langle \bar{\mathbf{m}}, \mathbf{q}_i \rangle / (\|\bar{\mathbf{m}}\| \, \|\mathbf{q}_i\| \cdot \tau), \tag{8}$$

and a predictive path $\mathbf{w}_{\text{pred}} = \text{MLP}_{\text{weight}}(\bar{\mathbf{m}}) \in \mathbb{R}^K$. The final routing weights are:

$$\alpha_i = \sigma(s_i + w_{\text{pred},i}) \in (0,1), \tag{9}$$

yielding per-expert activations rather than a single distribution. This allows multiple attribute experts to activate jointly when the manipulation text targets multiple attributes, which a softmax router would prevent by forcing competition among experts. The complete module:

$$\text{MoA}^3(\mathbf{X}_q, \mathbf{X}_{kv}; \mathbf{M}) = \text{GMA}\big(\mathbf{X}_q, \mathbf{X}_{kv}, \, \boldsymbol{\alpha}(\mathbf{M})\big), \tag{10}$$

where $\boldsymbol{\alpha}(\mathbf{M})$ is computed from the manipulation tokens $\mathbf{M}$ via the attribute router, shared across all MoA^3 modules within a layer.

We further prove in Appendix C (Proposition C.1 and Theorem C.4) that the gradient gating induced by MoA^3 provably promotes attribute specialization among expert groups, a property that standard multi-head attention provably lacks (Remark C.5).

2.5 Target-Side Encoder

The target-side processes target image I_t and target description D_t through the same CLIP encoders to obtain $\mathbf{V}_t \in \mathbb{R}^{L_{v_t}\times d}$ and $\mathbf{D}_t \in \mathbb{R}^{L_{d_t}\times d}$. Unlike the query side, targets do not have modification intent. Thus we adopt a lightweight two-stream architecture with L layers of bidirectional cross-attention, shown in Figure 2(2)(d), omitting MoA^3 modules. We then pool and aggregate:

$$\begin{aligned} \mathbf{v}_t, \mathbf{d}_t &= \text{MeanPool}\Big(\mathbf{V}_t^{(L)}, \mathbf{D}_t^{(L)}\Big), \\ \mathbf{c}_t &= \text{MLP}\big([\mathbf{v}_t; \mathbf{d}_t]\big) + w_{v_t}\mathbf{v}_t + w_{d_t}\mathbf{d}_t. \end{aligned} \tag{11}$$

2.6 Two-Phase Training Strategy

We decouple attribute specialization from holistic retrieval optimization through a two-phase training strategy: first leveraging augmented data for fine-grained expert learning, then optimizing on real data for holistic retrieval.

Phase 1: Attribute Specialization with AWPSE loss. Standard contrastive losses optimize holistic alignment, but may overlook fine-grained attribute-level correspondences. To address this, we train on augmented data using the Asymmetric Weighted Point-Set Embedding (AWPSE) loss, inspired by Uesaka et al. [36]. We represent the query as a weighted point set $\mathcal{Q} = \{(\alpha_i, \mathbf{o}'^i)\}_{i=1}^K$, where $\mathbf{o}'^i$ denotes the i-th expert output from the final MoA^3 layer. We define an asymmetric similarity function where query retains point-set structure while target uses a unified embedding:

$$\text{s}(\mathcal{Q}, \mathbf{c}_t) = \sum_{i=1}^{K} \alpha_i \cdot k\big(\mathbf{o}'^i, \, \pi_i(\mathbf{c}_t)\big), \tag{12}$$

with π_i projecting $\mathbf{c}_t$ into the i-th attribute subspace and $k(\cdot,\cdot)$ a similarity kernel. The contrastive objective is then:

$$\mathcal{L}_{\text{AWPSE}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\text{s}(\mathcal{Q}_i, \mathbf{c}_{t_i})/\tau)}{\sum_{j=1}^{B}\exp(\text{s}(\mathcal{Q}_i, \mathbf{c}_{t_j})/\tau)}, \tag{13}$$

where B is batch size and τ is temperature. Each projection π_i is implemented as a learnable linear transformation $\pi_i(\mathbf{c}_t) = \mathbf{W}_i^{\text{proj}}\mathbf{c}_t + \mathbf{b}_i^{\text{proj}}$ where $\mathbf{W}_i^{\text{proj}} \in \mathbb{R}^{(d/K)\times d}$ projects the target embedding into the i-th attribute subspace matching the dimensionality of expert output $\mathbf{o}'^i \in \mathbb{R}^{d/K}$. The kernel function $k(\cdot,\cdot)$ computes the cosine similarity between the expert output and the projected target: $k(\mathbf{o}'_i, \pi_i(\mathbf{c}_t)) = (\mathbf{o}'_i \cdot \pi_i(\mathbf{c}_t))/(\|\mathbf{o}'_i\| \, \|\pi_i(\mathbf{c}_t)\|)$.

Critically, our augmentation pipeline naturally provides hard negatives by grouping variants from the same source within batch, where samples differ solely in manipulated attributes.

Phase 2: Holistic Alignment with InfoNCE Loss. After establishing attribute specialization, we transition to real data and optimize the standard InfoNCE loss:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\mathbf{c}_{q_i}^{\top}\mathbf{c}_{t_i}/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{c}_{q_i}^{\top}\mathbf{c}_{t_j}/\tau)}. \tag{14}$$

To preserve the attribute-specialized routing and group-wise attention learned in Phase 1, we freeze the router, group-specific attention projections, and attribute query tokens in all MoA^3 modules. We keep the final aggregation MLPs and modality-level fusion layers trainable to adapt the global query-target alignment to real data.

Table 1: Performance comparison on ECCIR and Shoes datasets. The best results are bold, while the second-best results are underlined.

Method	Venue	ECCIR				Shoes			
		R@1	R@5	R@10	Avg.	R@1	R@10	R@50	Avg.
FashionERN [4]	AAAI2024	36.45	70.18	82.27	62.97	22.73	55.59	81.71	53.34
Prog. Lrn. [47]	SIGIR2022	38.62	72.85	83.54	65.00	22.88	58.83	84.16	55.29
CAFF [38]	CVPR2024	39.18	73.47	84.06	65.57	25.21	60.17	80.79	55.39
TG-CIR [40]	ACM MM2023	41.35	75.68	85.84	67.62	25.89	63.20	85.07	58.05
CCIN [35]	CVPR2025	42.98	74.25	83.03	66.75	25.95	65.76	86.54	59.42
DQU-CIR [41]	SIGIR2024	43.83	77.62	87.48	69.65	31.47	69.19	88.52	63.06
OFFSET [5]	ACM MM2025	44.76	78.29	87.58	70.21	31.52	69.96	<u>89.21</u>	63.56
SPRC [2]	ICLR2024	47.02	79.41	88.22	71.55	28.47	66.18	87.35	60.67
TME [20]	CVPR2025	46.82	80.02	88.64	71.83	29.15	67.52	88.18	61.62
CIR-LVLM [34]	AAAI2025	52.72	86.26	<u>92.71</u>	77.23	31.40	70.20	88.91	63.51
MoA^{3}CIR	NeurIPS2026	**58.13**	**89.39**	**95.16**	**80.89**	**33.83**	**70.96**	**89.66**	**64.82**
$\sim$ w.o. Target Description		<u>53.24</u>	<u>86.42</u>	92.27	<u>77.31</u>	<u>32.76</u>	<u>70.21</u>	89.07	<u>64.01</u>

Table 2: Performance comparison on FashionIQ dataset. The best results are bold, while the second-best results are underlined.

Method	Venue	Dress		Shirt		Toptee		Avg.	
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
DWC [11]	AAAI2024	32.67	57.96	35.53	60.11	40.13	66.09	36.11	61.39
SSN [39]	AAAI2024	34.36	60.48	38.13	61.83	44.26	69.05	38.92	63.89
SADN [39]	ACM MM2024	40.01	65.10	43.67	66.05	48.04	70.93	43.91	67.36
CaLa [14]	SIGIR2024	42.38	66.08	46.76	68.18	50.93	73.42	46.69	69.22
CoVR-BLIP [37]	AAAI2024	44.55	69.03	48.43	67.42	52.60	74.31	48.53	70.25
CASE [16]	AAAI2024	47.44	69.36	48.48	70.23	50.18	72.24	48.79	70.68
Re-ranking [26]	TMLR2024	48.14	71.43	50.15	71.25	55.23	76.80	51.17	73.13
FashionERN [4]	AAAI2024	43.93	68.77	52.70	75.07	56.09	78.38	50.91	74.07
DQU-CIR [41]	SIGIR2024	51.90	74.37	53.57	73.21	54.65	75.56	53.37	74.38
SPRC [2]	ICLR2024	49.18	72.43	55.64	73.89	59.35	78.58	54.72	74.97
CCIN [35]	CVPR2025	49.38	72.58	55.93	74.14	57.93	77.56	54.41	74.76
TME [20]	CVPR2025	49.73	71.69	56.43	74.44	59.31	78.94	55.15	75.02
MAPNet [32]	ICCV2025	51.17	74.12	56.37	75.17	59.56	79.30	55.70	76.20
CIR-LVLM [34]	AAAI2025	50.42	73.57	**58.59**	**75.86**	56.21	76.14	55.07	75.19
MoA^{3}CIR	NeurIPS2026	**53.39**	**74.99**	<u>56.58</u>	<u>75.26</u>	**62.56**	**81.76**	**57.51**	**77.34**
$\sim$ w.o. Target Description		<u>52.71</u>	<u>74.45</u>	55.82	75.09	<u>61.99</u>	<u>81.17</u>	<u>56.84</u>	<u>76.91</u>

3 Experiments

3.1 Experimental Setup

3.1.1 Datasets

We evaluate **MoA^{3}CIR** on three fashion-domain benchmarks: **ECCIR**, our constructed dataset; **FashionIQ** [43], a widely-adopted benchmark with three categories (Dress, Shirt, Toptee), where we follow the Original-Split; and **Shoes** [9], focusing on footwear retrieval. Since FashionIQ and Shoes originally provide only images without textual descriptions, we augment them with image-generated descriptions following the pipeline from DQU-CIR [41] to adapt to our framework.

3.1.2 Evaluation Metrics

We report dataset-specific metric: R@10 and R@50 for FashionIQ across three categories plus category-averaged results; R@1, R@10, and R@50 with average for Shoes; and R@1, R@5, and R@10 with average for ECCIR.

3.1.3 Implementation Details

MoA^{3}CIR employs CLIP as the vision-language backbone, where we adopt ViT-H-14 as the visual encoder with feature dimension d=1024. We use K=8 learnable attribute query tokens with N_h=2 heads per group (total H=16 heads). On the query side, we stack L=6 MoA^3 layers, which introduce only 29.41M additional parameters. All learnable tokens are randomly initialized. We train with AdamW optimizer, setting learning rates to 5×10^{-6} for the CLIP backbone and 5×10^{-5} for other components. The batch size is set to 64 and the dropout rate is 0.2.

For augmentation, we generate N_c=20 manipulation-text candidates per image and keep N_d=10 verified ones for target synthesis. We apply this pipeline to all three datasets, yielding augmented data equal to 50% of each dataset's original size. Training proceeds in two phases: AWPSE on augmented data and InfoNCE on real data with frozen MoA^3 modules, both for 5 epochs.

3.2 Experimental Results

Across all benchmarks, **MoA^{3}CIR** achieves consistent gains over multiple baselines.

3.2.1 Quantitative Results

Performance on ECCIR. Table 1 shows that **MoA^{3}CIR** achieves strong results on ECCIR, with R@1 of 58.13% and average recall of 80.89%. This outperforms the second-best method CIR-LVLM by 5.41 pp at R@1 and 3.66 pp on average recall, i.e., relative gains of 10.26% and 4.74%, respectively. **MoA^{3}CIR** also surpasses competitors at R@5 and R@10, indicating robust ranking quality. We attribute these gains mainly to incorporating item textual descriptions that capture non-visual attributes and to our attribute-aware modeling.

Generalization on Public Benchmarks. As shown in Table 2, on FashionIQ dataset, **MoA^{3}CIR** achieves averages of 57.51% at R@10 and 77.34% at R@50, surpassing MAPNet by 1.81 pp and 1.14 pp, respectively. Across categories, our method ranks first or second at R@10, demonstrating strong generalization. On Shoes, as depicted in Table 1, **MoA^{3}CIR** obtains an average recall of 64.82%, improving over OFFSET by 1.26 pp. In contrast to the surprising gains on ECCIR, the modest improvements on these benchmarks arise because their image-generated descriptions [41] primarily capture visual features, lacking the non-visual attributes present in ECCIR's descriptions.

Comparison with LVLM-based Method. As shown in Table 3, our **MoA^{3}CIR** achieves competitive performance with significantly fewer parameters compared to recent baselines. Across three benchmarks, our method achieves consistent gains over LVLM-based method CIR-LVLM while being more efficient with **8×** faster inference and **7×** smaller model size, underscoring its potential for resource-constrained scenarios.

Table 3: Comparison of model parameters and latency.

Method	**Architecture**	**Param.**	**Latency** (ms)
DQU-CIR	CLIP-based	1B	96
SPRC	Q-Former-based	2B	207
TME	Q-Former-based	2B	212
CIR-LVLM	LVLM-based	7B	813
MoA^{3}CIR	CLIP-based	1B	101

Comparison under Identical Input. To verify that our gains stem from architectural innovations rather than solely from additional modality, we compare against baselines using identical inputs (without target descriptions) since they are unable to process target descriptions. As shown in the "w.o. Target Description" rows in Tables 1 and 2, **MoA^{3}CIR** still outperforms baselines across all datasets.

For instance, we surpass CIR-LVLM by 0.52 pp at R@1 on ECCIR, and outperform MAPNet by 1.14 pp on average R@10 on FashionIQ.

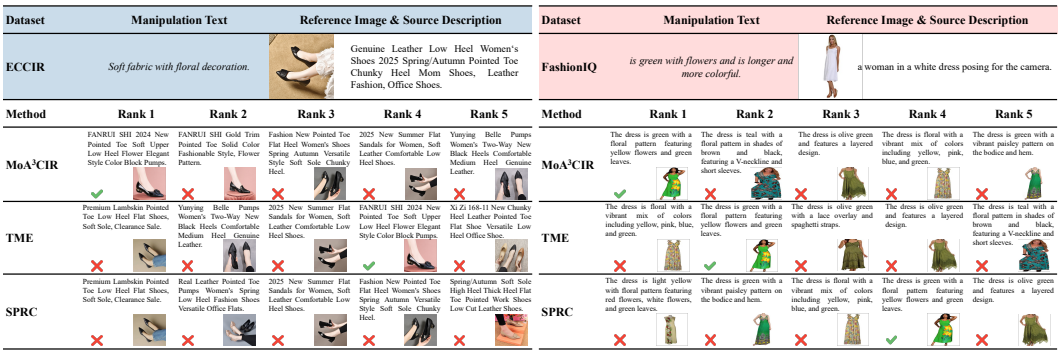

Figure 3: Case study on ECCIR and FashionIQ for three methods. Note that only **MoA^{3}CIR** can utilize target description.

3.2.2 Qualitative Analysis

Effectiveness of MoA^3. We analyze the routing weights $\boldsymbol{\alpha}$ over the K groups and stress that no expert is predefined. Instead, experts self-organize through grouped attention and learned routing. We provide theoretical justification for this specialization in Appendix C. As illustrated in Figure 4(a), without Phase-1, the routing weights are more uniform across experts; with Phase-1 training on augmented data (Figure 4(b)), the distributions become sharper and more consistent across queries with similar attribute modifications. Moreover, different manipulation types tend to concentrate on different experts: color queries (red) assign higher routing weights to experts 6 and 7; pattern queries (blue) emphasize expert 3; and style queries (green) emphasize experts 1, 2, and 6. This selective activation contrasts with standard attention, where all heads uniformly process information without explicit attribute specialization.

Visualization Cases. Figure 3 visualizes the comparative prediction results of our **MoA^{3}CIR** against TME and SPRC. As shown in the first example, our approach accurately identifies the nuanced attributes "soft and floral", leading to correct retrieval, which demonstrates both our model's fine-grained attribute comprehension and the significance of item descriptions.

3.3 Ablation Study

As shown in Table 4, we perform ablations on ECCIR to isolate the contribution of each component: the MoA^3 mechanism, AWPSE loss, Phase-1 training, and Phase-2 training. Further ablation studies on public datasets are in Appendix J.

Table 4: Ablation Study on key factors in **MoA^{3}CIR**. A.S. denotes the abbreviation for ablation study. We report results on ECCIR (Recall@K) and keep the average for reference.

A.S.	Key Factor	ECCIR			Avg.
		R@1	R@5	R@10	
1	w.o. MoA^3 module	54.28	86.62	91.43	77.45
2	w.o. AWPSE loss	55.54	87.94	93.29	78.93
3	w.o. Phase-1 Training	55.01	87.21	92.52	78.24
4	w.o. Phase-2 Training	48.12	80.79	87.46	72.12
	MoA^{3}CIR	**58.13**	**89.39**	**95.16**	**80.89**

Effect of MoA^3. Replacing MoA^3 with standard cross-attention (A.S. 1) causes a marked drop. R@1 decreases by 3.85 pp, confirming the importance of grouped attention and adaptive routing for fine-grained attribute correspondence.

Effect of AWPSE Loss. Using InfoNCE instead of AWPSE in Phase 1 (A.S. 2) consistently degrades performance (average 1.96 pp), proving AWPSE's point-set formulation explicitly encourages attribute-level specialization.

Effect of Phase 1. Skipping Phase 1 (A.S. 3) reduces average performance by 2.65 pp, underscoring that the first-phase training is critical for attribute expert specialization.

Effect of Phase 2. Skipping Phase 2 (A.S. 4) leads to a drastic decline because the model trained solely on augmented data does not adapt to the real data distribution.

Effect of Augmentation Quantity in Phase-1. To determine the suitable amount of augmented data for phase-1 training, we conduct experiments with augmentation ratios ranging from 0% to 100%. As shown in Table 5, incorporating augmented samples consistently improves model performance, with gains becoming more moderate at higher ratios. Considering the computational overhead of generating large-scale augmented data, we adopt 50% as our default setting, which provides a favorable trade-off between retrieval performance and generation cost.

Table 5: Effect of augmentation ratio in phase-1 training.

Aug. Ratio	ECCIR			Avg.
	R@1	R@5	R@10	
0%	55.01	87.21	92.52	78.24
25%	57.01	88.83	94.35	80.06
50%	58.13	89.39	95.16	80.89
75%	58.48	89.95	95.72	81.38
100%	**58.67**	**90.26**	**96.08**	**81.67**

Effect of Freezing MoA^3 Modules in Phase-2. To validate freezing MoA^3 modules during phase-2, we compare against keeping them trainable. As shown in Table 6, the trainable variant drops to 79.64 on average recall. This degradation arises from catastrophic forgetting when sequentially training across augmented and real datasets with distinct distributions, confirming that freezing preserves attribute-specialized knowledge from phase-1.

Table 6: Effect of freezing MoA^3 modules in phase-2 training.

Method	ECCIR			Avg.
	R@1	R@5	R@10	
MoA^{3}CIR	**58.13**	**89.39**	**95.16**	**80.89**
$\sim$ w.o. MoA^3 Freezing	56.37	87.91	94.64	79.64

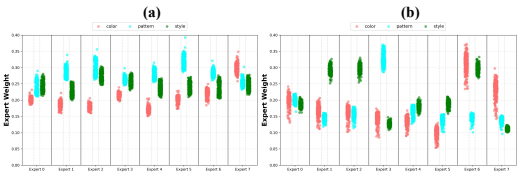

Figure 4: Router distribution across manipulation types. **(a)** Phase 2 only (without Phase 1). **(b)** Complete two-phase training.

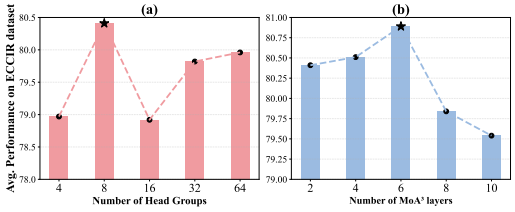

Figure 5: Ablation study on **(a)** the number of head groups in the MoA^3 module and **(b)** the number of MoA^3 layers.

Hyperparameter Analysis. Figure 5 indicates that performance peaks at K=8 attribute groups and L=6 MoA^3 layers. Fewer groups ($K \leq 4$) limit specialization capacity, while more groups ($K \geq 16$) lead to over-segmentation. Similarly, fewer layers ($L \leq 4$) provide insufficient refinement, whereas deeper ones show diminishing returns.

4 Limitation

In this section, we present several limitations. First, the self-reflective augmentation pipeline relies on LVLM and image-editing models, incurring a non-trivial one-time cost; we plan to release the augmented data so subsequent work can benefit without re-running it. Second, although **MoA^{3}CIR** generalizes to open-domain CIR even without textual descriptions, its full benefits are most pronounced

when item descriptions are readily available. Third, ECCIR currently covers fashion and accessories from a single platform; extending it to broader product domains is future work. We foresee no specific negative societal consequences beyond those shared by general image retrieval systems.

5 Conclusion

We present **MoA^{3}CIR**, a CIR framework that closes key gaps between research and real e-commerce usage by (i) integrating item descriptions on both query and target sides through an asymmetric five-tower architecture and constructing the dual-modal ECCIR benchmark with authentic descriptions, (ii) proposing a self-reflective data augmentation pipeline that generates controlled attribute modifications coupled with a two-phase training strategy, and (iii) introducing the **MoA^3** mechanism to explicitly model fine-grained attributes through grouped multi-head attention with learnable attribute tokens and dynamic routing. Extensive experiments show state-of-the-art results on ECCIR, FashionIQ, and Shoes, with ablations and analyses confirming the contribution of each component and the interpretability of attribute-specialized experts.

Acknowledgments and Disclosure of Funding

We thank Kuaishou Technology for supporting this research work.

Funding. This work was supported by Kuaishou Technology. The authors received no other third-party funding or support for this work during the 36 months preceding submission.

Competing interests. The authors' affiliation with Kuaishou Technology is disclosed in the author information. The authors declare no additional financial relationships or competing interests relevant to this work during the 36 months preceding submission.

References

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

[2] Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval. *arXiv preprint arXiv:2310.05473*, 2023.

[3] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21466–21474, 2022.

[4] Yanzhe Chen, Huasong Zhong, Xiangteng He, Yuxin Peng, Jiahuan Zhou, and Lele Cheng. Fashionern: enhance-and-refine network for composed fashion image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1228–1236, 2024.

[5] Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Xuemeng Song, and Liqiang Nie. Offset: Segmentation-based focus shift revision for composed image retrieval. *arXiv preprint arXiv:2507.05631*, 2025.

[6] Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

[7] Zhangchi Feng, Richong Zhang, and Zhijie Nie. Improving composed image retrieval via contrastive learning with scaling positives and negatives. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1632–1641, 2024.

[8] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023.

[9] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. *Advances in neural information processing systems*, 31, 2018.

[10] Xiaoyu He, Zibin Zheng, and Yuren Zhou. Mmes: Mixture model-based evolution strategy for large-scale optimization. *IEEE Transactions on Evolutionary Computation*, 25(2):320–333, 2021.

[11] 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*, volume 38, pages 2303–2311, 2024.

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

[13] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[14] Xintong Jiang, Yaxiong Wang, Mengjian Li, Yujiao Wu, Bingwen Hu, and Xueming Qian. Cala: Complementary association learning for augmenting comoposed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2177–2187, 2024.

[15] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

[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*, volume 38, pages 2991–2999, 2024.

[17] Guanyu Li, Zhiheng Xi, Zhihao Zhang, Boyang Hong, Tao Gui, Qi Zhang, and Xuanjing Huang. LoRACoE: Improving large language model via composition-based LoRA expert. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31290–31304, Suzhou, China, November 2025.

[18] 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.

[19] 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.

[20] Shuxian Li, Changhao He, Xiting Liu, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Learning with noisy triplet correspondence for composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19628–19637, 2025.

[21] You Li, Fan Ma, and Yi Yang. Imagine and seek: Improving composed image retrieval with an imagined proxy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3984–3993, 2025.

[22] Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *arXiv preprint arXiv:2405.11273*, 2024.

[23] Zihan Liang, Yufei Ma, ZhiPeng Qian, Huangyu Dai, Zihan Wang, Ben Chen, Chenyi Lei, Yuqing Ding, and Han Li. Uniecs: Unified multimodal e-commerce search framework with gated cross-modal fusion. *arXiv preprint arXiv:2508.13843*, 2025.

[24] Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *IEEE Transactions on Multimedia*, pages 1–14, 2026.

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

[26] Zheyuan Liu, Weixuan Sun, Damien Teney, and Stephen Gould. Candidate set re-ranking for composed image retrieval with dual multi-modal encoder. *arXiv preprint arXiv:2305.16304*, 2023.

[27] Yufei Ma, Zihan Liang, Huangyu Dai, Ben Chen, Dehong Gao, Zhuoran Ran, Wang Zihan, Linbo Jin, Wen Jiang, Guannan Zhang, Xiaoyan Cai, and Libin Yang. MoDULA: Mixture of domain-specific and universal LoRA for multi-task learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. EMNLP.

[28] Zhipeng Qian, Zihan Liang, Yufei Ma, Ben Chen, Huangyu Dai, Yiwei Ma, Jiayi Ji, Chenyi Lei, Han Li, and Xiaoshuai Sun. CSMCIR: CoT-enhanced symmetric alignment with memory bank for composed image retrieval. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 4411–4424, San Diego, California, United States, July 2026.

[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.

[30] Carlos Riquelme Ruiz, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*. ICLR, 2021.

[31] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[32] Jiangming Shi, Xiangbo Yin, Yeyun Chen, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Multi-schema proximity network for composed image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19999–20008, 2025.

[33] Lingyun Song, Xuequn Shang, Ruizhi Zhou, Jun Liu, Jie Ma, Zhanhuai Li, and Mingxuan Sun. A multi-group multi-stream attribute attention network for fine-grained zero-shot learning. *Neural Netw.*, 179(C), November 2024.

[34] Zelong Sun, Dong Jing, Guoxing Yang, Nanyi Fei, and Zhiwu Lu. Leveraging large vision-language model as user intent-aware encoder for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7149–7157, 2025.

[35] Likai Tian, Jian Zhao, Zechao Hu, Zhengwei Yang, Hao Li, Lei Jin, Zheng Wang, and Xuelong Li. Ccin: Compositional conflict identification and neutralization for composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3974–3983, 2025.

[36] Toshimitsu Uesaka, Taiji Suzuki, Yuhta Takida, Chieh-Hsin Lai, Naoki Murata, and Yuki Mitsufuji. Weighted point set embedding for multimodal contrastive learning toward optimal similarity metric. In *The Thirteenth International Conference on Learning Representations*. ICLR, 2025.

[37] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5270–5279, 2024.

[38] Yongquan Wan, Wenhai Wang, Guobing Zou, and Bofeng Zhang. Cross-modal feature alignment and fusion for composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8384–8388, 2024.

[39] Yifan Wang, Wuliang Huang, Lei Li, and Chun Yuan. Semantic distillation from neighborhood for composed image retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5575–5583, 2024.

[40] Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. Target-guided composed image retrieval. In *Proceedings of the 31st ACM international conference on multimedia*, pages 915–923, 2023.

[41] Haokun Wen, Xuemeng Song, Xiaolin Chen, Yinwei Wei, Liqiang Nie, and Tat-Seng Chua. Simple but effective raw-data level multimodal fusion for composed image retrieval. In *Proceedings of the 47th International ACM SIGIR conference on research and development in information retrieval*, pages 229–239, 2024.

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

[43] 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.

[44] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*, 2024.

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

[46] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. *arXiv preprint arXiv:2403.11549*, 2024.

[47] Yida Zhao, Yuqing Song, and Qin Jin. Progressive learning for image retrieval with hybrid-modality queries. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1012–1021, 2022.

A Related Work

A.1 Composed Image Retrieval

CIR leverages Vision-Language Pre-trained (VLP) models [29, 18, 19] to combine a reference image with manipulation text for retrieval. CLIP4CIR [3] pioneered contrastive pre-training for visual-textual fusion. Subsequent works advanced CIR along multiple directions: DQU-CIR [41] introduces late fusion for efficient retrieval, and SPRC [2] introduces sentence-level prompts to improve compositionality. However, most methods are unable to utilize item textual descriptions in e-commerce scenarios, which capture complementary non-visual attributes beyond images, and do not specifically model fine-grained attributes, due to their architectural designs. To address these gaps, we propose a five-tower **MoA^{3}CIR** framework that leverages item textual descriptions, incorporating the MoA^3 mechanism alongside tailored data augmentation and a two-phase training strategy.

A.2 CIR Datasets

The primary e-commerce CIR benchmarks include FashionIQ and Shoes. FashionIQ [43] pioneered fashion CIR with human-annotated manipulation texts but suffers from annotation quality issues [12] (e.g., contradictory or vague manipulations) and lacks item textual descriptions. Shoes [9] provides another fashion benchmark but remains limited in scale. Recent synthetic approaches leverage diffusion models [8, 21] or LLMs [12, 7] to construct triplets by synthesizing image-grounded captions; however, these texts are derived from the visual content itself and therefore do not introduce complementary non-visual attributes. Our ECCIR offers real e-commerce data with authentic item descriptions on both query and target sides, yielding high annotation quality, larger scale, and better aligns with practical applications.

A.3 Mixture of Experts

Mixture of Experts (MoE) enables dynamic computation through selective expert activation [31, 15] and has been extended to vision [30], NLP [27, 45, 6, 13], and multi-modal learning [22]. While our MoA^3 shares the high-level idea of selective expert activation, it differs from existing MoE designs in expert structure, routing signal, and design objective. We organize the related MoE designs into three categories.

FFN-based MoE for capacity scaling. A dominant line of MoE replicates feed-forward layers (FFN) as experts and dispatches individual tokens through hidden-state-based routers, primarily aimed at scaling model capacity under a fixed inference budget [31, 15, 13, 6]. MoE-LLaVA [24] duplicates FFNs in LVLM decoder blocks with token-level routing for efficient multimodal scaling, and the concurrent CIR work MMES [10] adopts a similar FFN-MoE design for retrieval. The router in this category is driven by load balancing and capacity allocation, with no explicit notion of *which conceptual axis* (e.g., color, material, style) is being modulated by user intent.

Adapter- or LoRA-based MoE for parameter-efficient transfer. A second line treats lightweight modules as experts within the parameter-efficient fine-tuning paradigm. MoE-Adapters [46] dispatches inputs across domain-specific adapters for continual learning of vision–language models, where each expert stores task- or domain-specific knowledge. MoLE [44] learns gating functions over independently pre-trained LoRA experts for compositional knowledge merging, while LoRACoE [17] decomposes LoRA matrices to the rank level and predicts per-rank importance for multi-task adaptation. In all these works, routing serves *adapter or rank composition* rather than attribute-level semantic fusion, and experts are inserted around a frozen backbone rather than being part of the cross-modal interaction itself.

Group-based attention for attribute modeling. The closest design philosophy to ours is group-based attribute attention. MGMSA [33] partitions categories into groups via offline K-means clustering and assigns each group an independent convolutional stream with static spatial/channel attention for fine-grained zero-shot classification. However, its attribute-to-group assignment is fixed and predetermined, with no dynamic routing mechanism, and the design is unimodal, lacking any cross-modal manipulation signal.

What sets MoA^3 apart. Our MoA^3 differs from all the above along three axes that are jointly absent from prior work. *(i) Expert granularity:* experts are groups of attention heads inside a cross-

modal attention layer, rather than FFN, adapter, or LoRA modules. *(ii) Routing signal:* experts are dynamically activated based on *manipulation-text semantics* via learnable attribute query tokens, directly answering the CIR-specific question "which attributes does the user want to modify?". It's a question none of the above formulations addresses, since their routers are driven by hidden-state load balancing, task indicators, or fixed offline clusters. *(iii) Design objective:* MoA^3 targets *attribute-level cross-modal fusion* for fine-grained CIR, rather than capacity scaling, continual learning, or PEFT-style adapter composition. We further provide a theoretical analysis in Appendix C showing that the gradient gating induced by MoA^3 provably promotes attribute specialization, a property that standard multi-head attention provably lacks.

On direct experimental comparison. A natural question is whether MoA^3 should be benchmarked head-to-head against the MoE designs above. Such a comparison is not directly applicable, because each of these methods is designed for a fundamentally different task: FFN-MoE for LLM capacity scaling, adapter/LoRA-MoE for continual learning or multi-task PEFT, and MGMSA for zero-shot classification. Adapting them to CIR would require non-trivial reimplementation as a cross-modal fusion module, in which case the resulting numbers would primarily reflect the chosen adaptation rather than the original MoE mechanism. The most relevant concurrent work, MMES [10], applies an FFN-MoE design to CIR but is closed-source at the time of submission; we will include the comparison once code becomes available. Within the supervised CIR setting, our experiments cover all open-source state-of-the-art baselines on ECCIR, FashionIQ, and Shoes (Tables 1–2), where MoA^3 consistently outperforms prior fusion designs.

B Dataset Construction and Augmentation

B.1 ECCIR Dataset Construction

Existing e-commerce CIR datasets lack item textual descriptions that are ubiquitous on real e-commerce platforms. These authentic descriptions provide essential non-visual attributes that cannot be inferred from images or generated by LVLMs, creating a fundamental gap between research and practical deployment. Moreover, these datasets suffer from quality issues including ambiguous manipulation texts and noisy annotations [12], as shown in Figure 7. To bridge this gap, we introduce ECCIR, the first dual-modal CIR benchmark featuring paired image-text data on both query and target sides.

Data Collection. We collect authentic item images and descriptions from a large-scale e-commerce platform based on real user browsing logs, with all private information removed, spanning eight categories including women's/men's/children's wear, shoes, underwear and accessories (bags, hats, scarves). The category distribution is shown in Figure 8. Importantly, these authentic descriptions provide richer information than LVLM-generated captions. This dataset features strict train-test separation ensuring no overlapping items, with detailed statistics shown in Table 8.

Description Quality Validation. Since item descriptions are seller-authored and may carry noise, we apply a three-stage quality validation pipeline before release. First, we strip user-identifiable information and sensitive terms to ensure privacy compliance. Second, we apply automatic spelling correction on the raw texts to fix obvious typographical errors. Third, we conduct manual spot-checks via majority voting among three annotators on a stratified sample across all eight categories, yielding a qualification rate above 98%. This pipeline filters out approximately 12% of raw descriptions due to low-quality or irrelevant content (e.g., descriptions consisting only of marketing slogans without substantive attribute information). For FashionIQ and Shoes, which lack authentic seller descriptions, we follow the same image-captioning pipeline as DQU-CIR [41] without additional filtering, in order to ensure a fair comparison with prior work.

Data Annotation. As illustrated in Figure 6, we annotate samples through a three-step pipeline. First, we compute pairwise similarity using UniECS [23] and retain similar-but-distinct pairs that are more suitable for CIR. We cap each item at five times to maintain a balanced distribution. Second, we employ Qwen2.5-32B-VL [1] to generate manipulation texts describing inter-item differences, using two modalities as input. Prompts emphasize colloquial phrasing and fine-grained details, with few-shot examples to mimic authentic user queries. Third, we conduct manual verification to remove ambiguous and noisy samples, retaining only cases approved by both annotators. Details are provided in Appendix F and G.

Dataset Advantages. ECCIR offers three key advantages: (1) *Dual-modal design*: incorporates textual descriptions alongside images on both sides, reflecting real-world e-commerce settings. (2) *Authentic data*: constructed from genuine browsing logs across diverse categories. (3) *High-quality annotations*: weakly similar pairs and colloquial texts verified by human experts yield cleaner supervision and better alignment with real user behavior.

Robustness to Description Noise. Beyond data-level filtering, we evaluate **MoA^{3}CIR**'s robustness to residual noise that may persist in real deployments. We inject controlled noise into target descriptions at training time by either inserting random distractor tokens (e.g., "whenever", "teh", "whti", "thereof") or applying random masking, at perturbation ratios of 20% and 40%. As shown in Table 7, performance degrades only marginally: at 40% perturbation, average recall on ECCIR drops by 1.38 pp from the clean setting yet remains 2.20 pp above the no-description baseline. This indicates that MoA^3's routing mechanism downweights uninformative tokens via dynamic activation weights, providing model-level robustness on top of the data-level validation pipeline above.

B.2 Self-Reflective Data Augmentation

Existing CIR samples often differ in multiple uncontrolled attributes, which weakens fine-grained supervision. Therefore, we propose a self-reflective augmentation pipeline that yields controlled attribute modifications, as shown in Figure 2(1). The augmented data provide targeted supervision for specializing attribute experts in our training strategy. The complete algorithm is detailed in Algorithm 1 in Appendix.

Table 7: Robustness of **MoA^{3}CIR** to description noise on ECCIR. Baseline denotes **MoA^{3}CIR** without target descriptions.

Setting	ECCIR Avg. Recall
Baseline (w.o. Target Description)	77.31
Clean (0%)	80.89
20% random perturbation	79.92
40% random perturbation	79.51

Stage 1: Manipulation Text Generation. We employ Qwen2.5-32B-VL [1] to generate attribute-specific manipulation texts. For each query, we sample one or two attributes from a predefined pool (color, material, style, etc.) and use a prompt with requirements and few-shot examples to produce N_c diverse candidates. **Stage 2: Self-Reflective Verification.** The same LVLM verifies each candidate text for appropriateness and feasibility, outputting binary decisions. Failed candidates receive feedback for regeneration. **Stage 3: Target Image Generation.** We use Qwen-Image-Edit [42] to synthesize target images for verified texts. Variants from the same query differ only in the manipulated attributes while preserving other contents, creating natural hard negatives that facilitate attribute-level learning. The full set of prompting templates and detailed qualitative examples are provided in Appendix H and I.

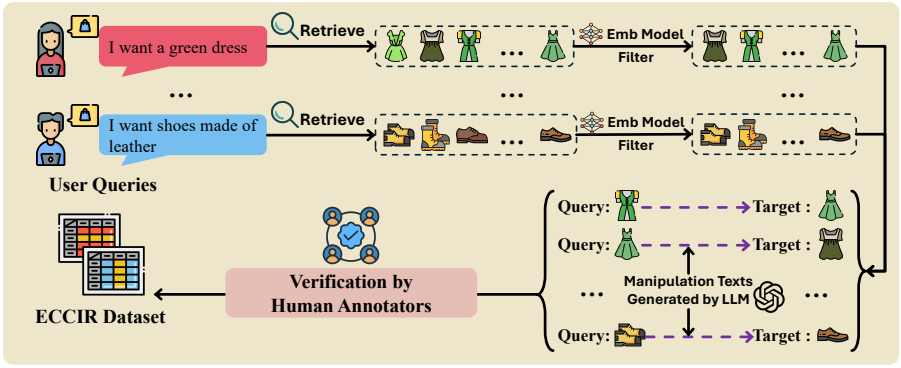

Figure 6: Construction pipeline for ECCIR dataset.

C Theoretical Analysis of MoA^3 Mechanism

In this section, we provide a rigorous theoretical analysis of why MoA^3 mechanism induces attribute specialization among expert groups. We formalize the mechanism, establish the gradient gating property, and prove convergence to specialized representations under mild assumptions.

C.1 Formal Setup and Notation

Standard Multi-Head Attention. Given query tokens $\mathbf{X}_q \in \mathbb{R}^{L_q \times d}$ and key-value tokens $\mathbf{X}_{kv} \in \mathbb{R}^{L_{kv} \times d}$, standard multi-head attention (MHA) with H heads computes:

$$\text{MHA}(\mathbf{X}_q, \mathbf{X}_{kv}) = \text{Concat}(\mathbf{h}_1, \ldots, \mathbf{h}_H)\mathbf{W}^O, \tag{15}$$

where $\mathbf{h}_i = \text{Softmax}\left(\frac{\mathbf{X}_q \mathbf{W}_i^Q (\mathbf{X}_{kv}\mathbf{W}_i^K)^\top}{\sqrt{d_h}}\right)\mathbf{X}_{kv}\mathbf{W}_i^V$ with $d_h = d/H$. Critically, all heads contribute equally to the output through uniform concatenation.

MoA^3 Formulation. We partition H attention heads into K disjoint groups $\{\mathcal{G}_1, \ldots, \mathcal{G}_K\}$, each containing $N_h = H/K$ heads. Let $\boldsymbol{\theta}_i = \{\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V, \mathbf{W}_i^O\}$ denote the parameters of group i, and let $\mathbf{o}^i \in \mathbb{R}^{L_q \times d}$ denote its output. Given manipulation text with pooled embedding $\bar{\mathbf{m}} = \text{MeanPool}(\mathbf{M})$ and learnable attribute query tokens $\mathbf{Q}_{\text{attr}} = [\mathbf{q}_1; \ldots; \mathbf{q}_K] \in \mathbb{R}^{K \times d}$, the routing weight for expert i is:

$$\alpha_i = \sigma\left(s_i + w_{\text{pred},i}\right), \quad \text{where } s_i = \frac{\langle \bar{\mathbf{m}}, \mathbf{q}_i \rangle}{\|\bar{\mathbf{m}}\|\|\mathbf{q}_i\| \cdot \tau}, \tag{16}$$

with $\sigma(\cdot)$ being the sigmoid function, τ a temperature parameter, and $\mathbf{w}_{\text{pred}} = \text{MLP}(\bar{\mathbf{m}}) \in \mathbb{R}^K$ a predictive pathway. The MoA^3 output is:

$$\mathbf{Y} = \left[\alpha_1 \mathbf{o}^1; \ldots; \alpha_K \mathbf{o}^K\right]\mathbf{W}_{\text{proj}}. \tag{17}$$

C.2 Gradient Gating Property

We first establish that routing weights multiplicatively gate gradient flow to each expert.

Proposition C.1 (Gradient Gating)**.** *Let $\mathcal{L}$ be a differentiable loss function. For the MoA^3 output in* (17)*, the gradient with respect to expert parameters $\boldsymbol{\theta}_i$ satisfies:*

$$\nabla_{\boldsymbol{\theta}_i}\mathcal{L} = \alpha_i \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{o}^i} \cdot \frac{\partial \mathbf{o}^i}{\partial \boldsymbol{\theta}_i}. \tag{18}$$

Proof. Applying the chain rule to (17):

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_i} = \frac{\partial \mathcal{L}}{\partial \mathbf{Y}} \cdot \frac{\partial \mathbf{Y}}{\partial \mathbf{o}^i} \cdot \frac{\partial \mathbf{o}^i}{\partial \boldsymbol{\theta}_i}. \tag{19}$$

From (17), $\frac{\partial \mathbf{Y}}{\partial \mathbf{o}^i} = \alpha_i \cdot \mathbf{W}_{\text{proj}}^{(i)}$, where $\mathbf{W}_{\text{proj}}^{(i)}$ denotes the projection slice for expert i. Thus:

$$\nabla_{\boldsymbol{\theta}_i}\mathcal{L} = \alpha_i \cdot \underbrace{\frac{\partial \mathcal{L}}{\partial \mathbf{Y}} \cdot \mathbf{W}_{\text{proj}}^{(i)}}_{\frac{\partial \mathcal{L}}{\partial \mathbf{o}^i}} \cdot \frac{\partial \mathbf{o}^i}{\partial \boldsymbol{\theta}_i}, \tag{20}$$

which establishes (18). □

Implication. The routing weight α_i acts as a multiplicative gate on gradient magnitude. When $\alpha_i \approx 1$, expert i receives full gradient signal; when $\alpha_i \approx 0$, gradient flow to expert i is suppressed. This creates a positive feedback loop: experts receiving larger gradients for certain samples become increasingly specialized for those samples.

C.3 Attribute Specialization Theorem

We now formalize conditions under which MoA^3 induces attribute-specific expert specialization.

Definition C.2 (Attribute-Indexed Training Distribution)**.** Let $\mathcal{A} = \{a_1, \ldots, a_K\}$ be a set of K attribute categories. For each training sample (I_q, D_q, M, I_t, D_t), let $\phi : \mathcal{M} \to \mathcal{A}$ map manipulation text M to its primary attribute. Define:

- $\mathcal{D}$: the overall training distribution
- $\mathcal{D}_k = \mathcal{D} \mid \phi(M) = a_k$: the conditional distribution for attribute a_k
- $p_k = P(\phi(M) = a_k)$: the attribute prior probability

Definition C.3 (Activation Concentration Matrix)**.** Define the activation concentration matrix $\mathbf{A} \in \mathbb{R}^{K \times K}$ with entries:

$$A_{ik} = \mathbb{E}_{(I_q, D_q, M, I_t, D_t) \sim \mathcal{D}_k}[\alpha_i(M)] \tag{21}$$

representing the expected activation of expert i on samples with attribute a_k.

Theorem C.4 (Specialization Convergence)**.** *Assume the following conditions hold:*

- ***(C1) Attribute coverage:*** *$p_k > 0$ for all $k \in \{1, \ldots, K\}$.*
- ***(C2) Semantic separability:*** *There exist attribute prototypes $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathbb{R}^d$ such that for manipulation text M with $\phi(M) = a_k$, we have $\mathbb{E}[\bar{\mathbf{m}}] = \boldsymbol{\mu}_k$, and $\langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle < \|\boldsymbol{\mu}_i\|^2$ for all $i \neq j$.*
- ***(C3) Contrastive training with hard negatives:*** *The model is trained with AWPSE loss on augmented data where hard negatives differ from positives only in the manipulated attributes.*

Then, as training converges, the activation matrix $\mathbf{A}$ achieves diagonal dominance:

$$A_{ik} \xrightarrow{\text{training}} \begin{cases} \alpha^* & \text{if } i = k \\ \epsilon_{ik} & \text{if } i \neq k \end{cases} \tag{22}$$

where $\alpha^* > 1/2$ and $\epsilon_{ik} \ll \alpha^*$ for all $i \neq k$. That is, expert i becomes predominantly activated for attribute a_i.

Proof. The proof proceeds in three steps.

Step 1: Router Gradient Analysis. Consider the gradient of $\mathcal{L}_{\text{AWPSE}}$ with respect to attribute token $\mathbf{q}_i$. From Equations (12)-(13) in the main text, the AWPSE loss involves the similarity:

$$\mathrm{s}(\mathcal{Q}, \mathbf{c}_t) = \sum_{j=1}^{K} \alpha_j \cdot k(\mathbf{o}'^j, \pi_j(\mathbf{c}_t)), \tag{23}$$

where $k(\cdot, \cdot)$ is a kernel function and π_j projects into the j-th attribute subspace. For a positive pair with attribute a_k, the contrastive objective encourages increasing $\mathrm{s}(\mathcal{Q}, \mathbf{c}_t^+)$ while decreasing $\mathrm{s}(\mathcal{Q}, \mathbf{c}_t^-)$ for negative targets. This yields:

$$\nabla_{\mathbf{q}_i} \mathcal{L} \propto -\sum_{\text{pos}} \frac{\partial \alpha_i}{\partial \mathbf{q}_i} k(\mathbf{o}'^i, \pi_i(\mathbf{c}_t^+)) + \sum_{\text{neg}} \frac{\partial \alpha_i}{\partial \mathbf{q}_i} k(\mathbf{o}'^i, \pi_i(\mathbf{c}_t^-)). \tag{24}$$

Step 2: Alignment Dynamics. From (16), $\frac{\partial \alpha_i}{\partial \mathbf{q}_i} \propto \sigma'(\cdot) \cdot \bar{\mathbf{m}} / \|\mathbf{q}_i\|$. Under **(C3)**, hard negatives share the same manipulation text as positives but differ in target attributes. Thus the kernel $k(\mathbf{o}'^i, \pi_i(\mathbf{c}_t))$ measures how well expert i's output aligns with attribute-i information in the target.

For expert k processing attribute-k samples: the positive term in (24) dominates when $k(\mathbf{o}'^k, \pi_k(\mathbf{c}_t^+))$ is large (correct attribute match). The gradient direction $\propto \bar{\mathbf{m}} \approx \boldsymbol{\mu}_k$ updates $\mathbf{q}_k$ toward the attribute-k prototype. For expert $i \neq k$: the negative term dominates since hard negatives require suppressing α_i to avoid false matches. This pushes $\mathbf{q}_i$ away from $\boldsymbol{\mu}_k$, encouraging orthogonalization.

Step 3: Convergence to Diagonal Dominance. By **(C2)**, the attribute prototypes $\{\boldsymbol{\mu}_k\}$ are separable. The gradient dynamics from Step 2 drive:

$$\mathbf{q}_k \to \gamma_k \boldsymbol{\mu}_k + \boldsymbol{\delta}_k, \quad \text{where } \gamma_k > 0 \text{ and } \|\boldsymbol{\delta}_k\| \ll \gamma_k \|\boldsymbol{\mu}_k\|. \tag{25}$$

At convergence, for a sample with $\phi(M) = a_k$ (so $\bar{\mathbf{m}} \approx \boldsymbol{\mu}_k$):

$$s_k = \frac{\langle \bar{\mathbf{m}}, \mathbf{q}_k \rangle}{\|\bar{\mathbf{m}}\| \|\mathbf{q}_k\|} \cdot \frac{1}{\tau} \approx \frac{1}{\tau} \cdot \frac{\gamma_k \|\boldsymbol{\mu}_k\|^2}{\|\boldsymbol{\mu}_k\| \cdot \gamma_k \|\boldsymbol{\mu}_k\|} = \frac{1}{\tau}. \tag{26}$$

For $i \neq k$, since $\mathbf{q}_i$ has been pushed away from $\boldsymbol{\mu}_k$:

$$s_i = \frac{\langle \boldsymbol{\mu}_k, \mathbf{q}_i \rangle}{\|\boldsymbol{\mu}_k\| \|\mathbf{q}_i\|} \cdot \frac{1}{\tau} \leq \frac{1}{\tau} \cdot \cos \theta_{ki}, \tag{27}$$

where $\theta_{ki} = \angle(\boldsymbol{\mu}_k, \mathbf{q}_i) > \theta_{\min} > 0$ for some minimum angle induced by the orthogonalization dynamics.

Applying the sigmoid nonlinearity with appropriate τ:

$$\alpha_k = \sigma(\frac{1}{\tau} + w_{\text{pred},k}) \to \alpha^* > 1/2, \quad \alpha_i = \sigma(\frac{1}{\tau} \cos \theta_{ki} + w_{\text{pred},i}) \to \epsilon_{ik} \ll \alpha^*, \tag{28}$$

establishing the diagonal dominance in (22). □

C.4 Comparison with Standard Attention

Remark C.5 (Limitations of Standard Attention). In standard multi-head attention, the gradient for head i's parameters is:

$$\nabla_{\boldsymbol{\theta}_i} \mathcal{L}_{\text{MHA}} = \frac{\partial \mathcal{L}}{\partial \mathbf{h}_i} \cdot \frac{\partial \mathbf{h}_i}{\partial \boldsymbol{\theta}_i}, \tag{29}$$

without the α_i gating factor present in MoA^3 (cf. Proposition C.1). This has three consequences:

1. **Uniform gradient distribution:** All heads receive comparable gradient magnitudes across all samples, regardless of attribute relevance.
2. **No specialization incentive:** The training dynamics provide no mechanism for heads to specialize for different attributes.
3. **Entangled representations:** Attribute-specific information remains distributed and entangled across all heads.

This theoretical distinction explains the empirical observations in Figure 5: standard attention cannot produce the attribute-specific routing patterns that MoA^3 achieves.

Summary. Our theoretical analysis establishes that MoA^3's gradient gating mechanism (Proposition C.1), combined with AWPSE training on attribute-controlled augmented data, provably induces expert specialization (Theorem C.4). This property cannot emerge from standard attention (Remark C.5), providing rigorous theoretical justification for MoA^3's superior performance in fine-grained attribute retrieval.

D Noisy Annotation Examples in FashionIQ

Figure 7 illustrates common annotation quality issues in FashionIQ that motivated ECCIR construction. These examples show (1) vague descriptions (e.g., “weird" and “less emo" are confusing for the target pattern modification.), (2) contradictory text and entity mismatch (e.g., “is not clothing" for a dress and a dress in reference image while a wallet in target image), and (3) Incomplete manipulation execution where manipulation text does not align with actual image differences. Such noisy annotations impede reliable model training and evaluation.

Reference Image	Manipulation Text	Target Image	Explanation
	*is weird*; *is less emo*		The reference image changes greatly in pattern, yet the manipulation text describes it with only the vague attributes ‘*weird*’ and ‘*less emo*’.
	*is a long dress with a combination of black and white*; *is not clothing and has smiley faces*		The reference image depicts a long dress while the target image shows a wallet, representing distinct entities. Additionally, the manipulation text exhibits an internal contradiction by stating the entity ‘is *a long dress’* yet asserting it ‘is *not clothing’*.
	*red*; *red and longer mermaid skirted*		Despite the manipulation text specifying a ‘*red*’ color and ‘*longer mermaid*’ style, the skirt in the target image exhibits neither of these characteristics.

Figure 7: Noisy Annotation Examples in FashionIQ.

E Statistics of ECCIR Dataset

ECCIR contains 110K total samples (92K training, 18.6K testing), making it $5\times$ larger than FashionIQ and $8\times$ larger than Shoes. This scale reflects our quality-first principle, prioritizing authentic item descriptions and human-verified manipulation texts over sheer quantity. Figure 8 shows the category distribution across eight e-commerce categories. Table 8 presents detailed comparisons with existing benchmarks, highlighting ECCIR's unique dual-modal design where both query and target sides contain authentic textual descriptions that capture non-visual attributes (e.g., material, brand, function) unavailable in previous datasets.

Table 8: Dataset statistics on FashionIQ, Shoes and ECCIR.

Dataset	Query Side		Target Side		Split	Num.	Attr. Num.	Avg. M.T. Len.	Cate. Num.	Desc. Source
	Img.	Desc.	Img.	Desc.						
FashionIQ	✓	–	✓	–	Train	18000	37375	11.68	3	N/A
	✓	–	✓	–	Test	6016	12294	11.77		
Shoes	✓	–	✓	–	Train	10000	19518	5.19	1	N/A
	✓	–	✓	–	Test	4658	9126	5.32		
ECCIR	✓	✓	✓	✓	Train	92160	118381	5.74	8	Online Logs
	✓	✓	✓	✓	Test	18591	24636	5.72		

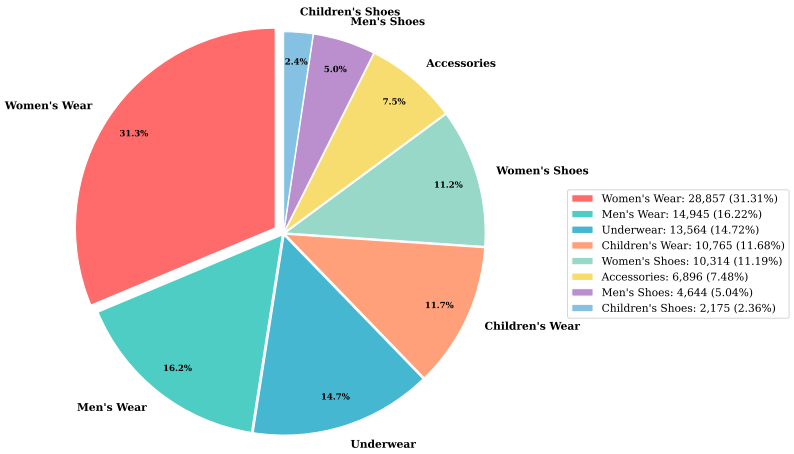

Figure 8: Category distribution of ECCIR train set. Test set follows the same distribution.

F Samples of ECCIR Dataset

Figure 9 illustrates samples from ECCIR, which contains real-world e-commerce data, including item images, descriptions, and manipulation texts. Notably, these descriptions often contain semantic information that can not be directly conveyed by images. For example, in the second one, the description includes seasonal information, while the third and fourth mention specific brands ("yaloo" and "huili"). These descriptions are integral to e-commerce platforms as they provide additional context and clarification that images alone cannot deliver.

Source Description	Reference Image	Manipulation Text	Target Image	Target Description
Plus-size rain boots for men, waterproof and non-slip.		*A **black low-top** version.*		New style men's rain boots, fishing casual water shoes, oil-proof, low-top short rain boots.
men's spring/autumn white denim jacket, korean style slim fit baseball jacket, cool brand workwear jacket.		*Do you have a **pink distressed denim** jacket with **no letters**?*		spring/autumn ripped denim jacket for teen boys, korean style lightweight loose denim coat.
yaloo women's lightweight short hooded slim fit down jacket with 90% white duck down, warm and lightweight coat.		*Do you have **gradient colors**?*		yaloo autumn/winter new lightweight white duck down jacket women gradient color short hooded fashion coat.
changxiang huili kids shoes white sports shoes autumn winter leather waterproof boy's rotating button white shoes soft sole.		***Mesh Netting** version please.*		changxiang huili boys' summer mesh white sneakers with rotating buttons, breathable upper, soft sole running shoes for kids.
Men's Spring and Autumn Washed and Distressed Jeans, American Style INS Retro Fashion Brand, Slightly Flared Pants.		*With **hole** and a **lighter color**.*		American high-street style, slim-flare jeans for men and women, vibe trendy brand, cool and edgy long pants.

Figure 9: Examples of ECCIR dataset.

G Prompting Template for ECCIR

The prompting template used for constructing ECCIR is shown in Figure 10, which guides the following reasoning steps:

Item Description Refinement (Step 0): This step generates refined descriptions by combining image and original description, focusing on color, material, and decorative features.

Query and Target Sides Analysis (Steps 1-2): These steps extract key features from both the reference and target items.

Key Differences Identification (Step 3): This step compares the extracted features and identifies 1-3 prominent differences.

Manipulation Text Generation (Step 4): In this final step, a colloquial manipulation text is generated to describe the changes, simulating real user search behavior.

Prompting Template for ECCIR

You are an assistant helping annotators create manipulation texts for Composed Image Retrieval in e-commerce scenarios. Annotators will provide two similar but distinct fashion items:

Input:

- **Reference Image** (with **Source Description**): The item the user currently has or has seen
- **Target Image** (with **Target Description**): The item the user wants to find

User's Search Intent: *"I want an item similar to the reference, but with specific characteristics from the target item"*

CRITICAL: The manipulation text must describe HOW the target differs from the reference, i.e., what changes the user wants from their reference baseline.

Analysis Steps (Follow This Order Strictly):

STEP 0: Optimize Item Descriptions
Based on both the image content and the provided description, create enhanced visual descriptions for each item.

Requirements:

- Limit each description to **30 words or less**
- Combine visual details from the image with semantic information from the description
- Focus on: **color, material, design style, structural features, decorative details**
- Include practical details from descriptions (e.g., age group, function, special craftsmanship)
- Remove marketing language and retain descriptive information
- Use intuitive, everyday visual vocabulary

Outputs:

1. **Reference Image Optimized Description**: Based on the image and source description "{source_description}", generate a detailed visual description
2. **Target Image Optimized Description**: Based on the image and target description "{target_description}", generate a detailed visual description

STEP 1: Analyze Key Features of Reference Image
Using the optimized description and image content, identify these key features:

- **Color**: Primary and secondary colors
- **Material**: Surface material and texture characteristics
- **Details**: Decorations, patterns, craftsmanship details
- **Style**: Formal/casual, seasonal appropriateness, occasion suitability

STEP 2: Analyze Key Features of Target Image
Using the optimized description and image content, describe the **same dimensions** of features:

- **Color**: Primary and secondary colors
- **Material**: Surface material and texture characteristics
- **Details**: Decorations, patterns, craftsmanship details
- **Style**: Formal/casual, seasonal appropriateness, occasion suitability

STEP 3: Identify Key Differences & Assess Suitability
Compare the two items based on the optimized descriptions:

- Are the item categories consistent? (e.g., both dresses, both shoes)
- **What are the 1-3 most salient differences?**
- Can these differences be expressed easily with everyday words?
- Is the degree of difference moderate? (Not too subtle, not too drastic)

Suitability Decision: Determine if this pair is appropriate for CIR annotation.

STEP 4: Generate User Manipulation Text
Based on the key differences identified in Step 3, generate a natural, colloquial manipulation text.

Core Principle: The manipulation text describes **what makes the Target different from the Reference**, expressing the user's desire to modify from the Reference baseline toward the Target's characteristics.

Critical Requirements:

- **Describe ONLY the differential features** (what distinguishes Target from Reference)
- **Do NOT describe features both items share**
- Focus on the **1-3 most salient differences** from Step 3
- Express as a real user would naturally request

Manipulation Text Specifications

Core Principles:

1. **Colloquial & Natural**: Simulate real user search behavior with complete questions or statements
2. **Highlight Differentials**: Focus exclusively on Target's distinguishing characteristics
3. **Everyday Language**: Avoid terminology; use ordinary user expressions
4. **Appropriate Length**: Neither too brief nor verbose. Be realistic for actual searches
5. **Detail When Obvious**: For clear visual differences, appropriate detail is acceptable

Important Clarification:

- **Reference Image** = User's current/known item (the baseline)
- **Target Image** = Item demonstrating the desired change direction
- **Manipulation Text** = Describes **"Reference → Target transformation"**

Common Expression Patterns:

Remember: ALL expressions describe Target's characteristics, NOT Reference's!

1. Question Form (Recommended)

- "Is there a white version of this style?"
- "Does this style come in leather material?"
- "Is there one with a matte finish in the same design?"
- "Similar but in a smaller size available?"

2. Needs Expression Form

- "Want the white version"
- "Need it in leather material"
- "Looking for one with matte finish"
- "Looking for the same style in a larger size"

3. Comparison Expression Form

- "Same as this but in white"
- "Same style but in canvas material"
- "Similar design but with a glossy finish"

4. Scenario-Based Expression Form

- "Suitable for formal occasions version"
- "More casual version of this style"
- "Similar style suitable for summer"
- "Fall/winter warm version of this"

5. Detail Description Form (for obvious differences)

- "Want this with lace decoration"
- "Want the same style with bow decoration"
- "Need a mesh tulle skirt version"

Final Output Format:

```
Reference Image Optimized Description: [Your description]
Target Image Optimized Description: [Your description]
Key Features of Reference: [Analysis]
Key Features of Target: [Analysis]
Key Differences (1-3): [List differences]
Suitability Assessment: [SUITABLE / NOT SUITABLE] [Reason]
Manipulation Text: [Your generated text]
```

Figure 10: Prompting template for ECCIR

H Prompting Template for Self-reflective Data Augmentation

The prompting template used for self-reflective data augmentation is shown in Figure 11, which guides the following stages:

Stage-1 (Figure 11(a)): This template guides Qwen2.5-32B-VL to generate attribute-specific manipulation texts for selected item, with few-shot examples demonstrating desired output formats.

Stage-2 (Figure 11(b)): This template implements self-reflective verification to filter inappropriate candidates, where LVLM outputs binary decisions with brief explanations. If a candidate fails verification, the LVLM also provides feedback to guide regeneration, enabling iterative refinement until high-quality manipulation texts are produced for the next step (image generation).

(a) Stage-1: Manipulation Text Generation with LVLM

Please analyze this fashion item image and generate concise manipulation texts for composed image retrieval.

Number of Manipulation Texts to Generate: {Nc} candidates

Selected Attribute(s): {attr1} [and {attr2} if two attributes sampled]

Attribute Requirements:
{requirements1} [for attr1]

[*If two attributes sampled:*]
{requirements2} [for attr2]

Few-shot Examples for {attr1}:
{examples1}

[*If two attributes sampled:*]
Few-shot Examples for {attr2}:
{examples2}

Output Format Requirements:

- Generate exactly {Nc} different manipulation texts
- Each text should describe **one or two** specific attribute changes
- Follow the pattern: "[change 1]" or "[change 1] and [change 2]"
- Use **colloquial, natural language** mimicking real user queries
- Keep each text concise (5-15 words)
- Separate each manipulation text with `<SEP>`
- Do not include numbering, formatting, or explanations

Example Output:
has long sleeves.`<SEP>`is cropped and fitted.`<SEP>`has a high neckline and is form-fitting.`<SEP>`is looser fit and more casual.`<SEP>`...

Please carefully observe the image and generate {Nc} manipulation texts:

(b) Stage-2: Self-Reflective Verification with LVLM Feedback

You are a professional fashion expert performing self-reflective verification of manipulation texts for composed image retrieval.

Task: Analyze this fashion item image and evaluate the following manipulation text.

CRITICAL UNDERSTANDING:
The manipulation text is **NOT** a description of the provided image. It is a **modification instruction** for a Diffusion Model to generate a target image that differs from this reference image **ONLY** in the specified attribute(s).

Manipulation Text to Verify:
`{instruction}`

Target Attribute Category: {attr}

Verification Criteria:

1. **Appropriateness:** Does this manipulation text make logical sense for this item type?
2. **Feasibility:** Can a diffusion model realistically apply this modification to generate a valid target image?
3. **Attribute Specificity:** Does the text clearly specify changes in the target attribute(s)?
4. **Controllability:** Can this modification be applied while preserving all other item characteristics?

CRITICAL OUTPUT FORMAT:
You must respond in **EXACTLY** this three-part format:

Part 1: Output ONLY ONE word:

- "**GOOD**" if the manipulation text meets **ALL** criteria above
- "**BAD**" if the manipulation text fails **ANY** criteria above

Part 2: Provide a brief reason (maximum 15 words)

Part 3: *(ONLY if verdict is BAD)* Provide specific feedback for regeneration

Example Outputs:

```
GOOD
Attribute-specific material modification, controllable and preserves overall style
```

```
BAD
Too vague, does not specify which attribute to modify
FEEDBACK: Specify exact attribute change (e.g., "leather material" instead of "different texture")
```

```
BAD
Modification would require changing multiple unrelated attributes simultaneously
FEEDBACK: Simplify to single attribute change (focus on color OR pattern, not both)
```

Now perform self-reflective verification and provide output in the required format:

Figure 11: Prompting templates for stage-1 (a) and stage-2 (b) of self-reflective data augmentation pipeline. Stage-1 employs flexible one-or-two attribute sampling to generate attribute-specific manipulation texts. Stage-2 performs self-reflective verification with binary decisions and regeneration feedback, enabling iterative refinement until high-quality candidates are produced for target image generation (Stage-3).

Algorithm 1 Self-Reflective Data Augmentation with LVLM Feedback

Require: Query image I_q, selected attributes $\{\text{attr}_1, [\text{attr}_2]\}$, attribute requirements $\{\mathcal{R}_1, [\mathcal{R}_2]\}$, few-shot examples $\{\mathcal{E}_1, [\mathcal{E}_2]\}$, number of initial candidates N_c, number of desired verified texts N_d, maximum retries per candidate $K_{\max}$

Ensure: Set of verified manipulation texts $\mathcal{M}_{\text{verified}}$ and corresponding target images $\mathcal{I}_{\text{target}}$

{**Stage 1: Manipulation Text Generation with LVLM**}
$\mathcal{M}_{\text{candidates}} \leftarrow \emptyset$
for $i = 1$ to N_c **do**
  $\mathcal{P}_{\text{gen}} \leftarrow \text{Prompt}_{\text{Stage1}}(I_q, \text{attr}, \mathcal{R}, \mathcal{E})$ {Construct prompt with attributes and examples}
  $m_i \leftarrow \text{LVLM}_{\text{generate}}(\mathcal{P}_{\text{gen}})$
  $\mathcal{M}_{\text{candidates}} \leftarrow \mathcal{M}_{\text{candidates}} \cup \{m_i\}$
end for

{**Stage 2: Self-Reflective Verification and Regeneration with LVLM Feedback**}
$\mathcal{M}_{\text{verified}} \leftarrow \emptyset$
for all $m \in \mathcal{M}_{\text{candidates}}$ **do**
  $k \leftarrow 0$
  isVerified $\leftarrow$ False
  $\mathcal{F} \leftarrow \emptyset$
  while $\neg$isVerified and $k < K_{\max}$ **do**
    {LVLM performs self-reflective verification}
    $\mathcal{P}_{\text{verify}} \leftarrow \text{Prompt}_{\text{Stage2}}(I_q, m, \text{attr})$
    $(v, r, [f]) \leftarrow \text{LVLM}_{\text{verify}}(\mathcal{P}_{\text{verify}})$ {v: verdict, r: reason, f: ***feedback***}
    if $v =$ ''GOOD'' **then**
      $\mathcal{M}_{\text{verified}} \leftarrow \mathcal{M}_{\text{verified}} \cup \{m\}$
      isVerified $\leftarrow$ True
    else
      {***Self-Reflective Regeneration Loop***}
      $\mathcal{F} \leftarrow \mathcal{F} \cup \{f\}$
      {Reconstruct prompt augmented with ***LVLM feedback***}
      $\mathcal{P}_{\text{regen}} \leftarrow \text{Prompt}_{\text{Stage1}}(I_q, \text{attr}, \mathcal{R}, \mathcal{E}, \mathcal{F})$
      $m \leftarrow \text{LVLM}_{\text{regenerate}}(\mathcal{P}_{\text{regen}})$
      $k \leftarrow k + 1$
    end if
  end while
  if $k = K_{\max}$ and $\neg$isVerified **then**
    Discard candidate after maximum retries
  end if
end for

{**Stage 3: Target Image Generation**}
$\mathcal{M}_{\text{final}} \leftarrow \text{Sample}(\mathcal{M}_{\text{verified}}, N_d)$
$\mathcal{I}_{\text{target}} \leftarrow \emptyset$
for all $m \in \mathcal{M}_{\text{final}}$ **do**
  $I_{\text{target}} \leftarrow \text{DiffusionModel}(I_q, m)$ {Qwen-Image-Edit}
  $\mathcal{I}_{\text{target}} \leftarrow \mathcal{I}_{\text{target}} \cup \{I_{\text{target}}\}$
end for
return $\mathcal{M}_{\text{final}}, \mathcal{I}_{\text{target}}$

I Augmented Samples for FashionIQ, Shoes, and ECCIR

Figure 12 illustrates several augmented samples generated by our pipeline across three datasets. Each row shows a reference image followed by multiple target images synthesized through Qwen-Image-Edit, where each target differs from the reference in only the specified attribute while preserving all other characteristics. Human evaluation on 500 samples by 2 annotators (both must agree) confirms 96% modify only the targeted attributes while preserving other characteristics.
Since the augmented targets lack textual descriptions, we generate them using the same pipeline as DQU-CIR (mentioned in Section 3.1.1) to adapt to our framework.

Dataset	Source Description	Reference Image	Manipulation Text	Target Image	Target Description
FashionIQ	The shirt is light blue with a tropical palm tree pattern. It has short sleeves and a button-down collar.		*has **zippers** instead of buttons and has a **ruffled** collar.*		Tropical Palm Tree Print Quarter-Zip Pullover with Ruffled Collar for Women - Blue and White Striped Design.
			*has an **embroidered design** and includes a **drawstring** waist.*		Women's Short-Sleeve Button-Up Dress with Tropical Palm Tree Print and Drawstring Waist Detail.
Shoes	merrell women's firefly 2 mesh clog.		*is **taller** in heel height.*		a women's black perforated slip-on high heel shoe with a textured sole.
			*has a **smoother texture**.*		a black slip-on clog shoe with a textured sole and orange logo detail.
ECCIR	hcgn/Hucheng Guniang Yunsan Acetic Acid Dress.		*Need this in a **V-neck style** to add more versatility.*		Women's White V-Neck Long Dress with Blue Ink Landscape Print.
			*Can I get this in a **sleeveless** version with the same printed design?*		Women's White Sleeveless Dress with Blue Mountain and Boat Print Design.

Figure 12: Augmented samples for FashionIQ, Shoes, and ECCIR.

J Ablation Study on Public Datasets

We conduct ablation studies on FashionIQ and Shoes to further justify key components in **MoA^{3}CIR**. As shown in Table 9, the performance trends align with those on ECCIR, though the impact of removing target descriptions is slightly smaller. This discrepancy stems from the nature of descriptions: FashionIQ and Shoes use image-generated descriptions that primarily capture visual features, while ECCIR employs authentic e-commerce descriptions containing rich non-visual attributes (e.g., material, brand, function).

Table 9: Ablation study on key factors in **MoA^{3}CIR** on Shoes and FashionIQ. A.S. denotes the abbreviation for ablation study.

A.S.	Key Factor	Shoes			FashionIQ		Avg.
		R@1	R@10	R@50	R@10	R@50	
1	w.o. MoA^3 module	31.24	68.37	87.59	55.68	75.42	63.66
2	w.o. AWPSE loss	32.45	69.82	88.91	56.72	76.57	64.89
3	w.o. Target Description	32.76	70.21	89.07	56.84	76.91	65.16
4	w.o. Phase-1 Training	32.18	69.14	88.35	56.29	76.18	64.43
5	w.o. Phase-2 Training	28.15	62.08	83.24	51.15	69.82	58.89
	MoA^{3}CIR	**33.83**	**70.96**	**89.66**	**57.51**	**77.34**	**65.86**

K Generalization to Open-Domain CIR

Although **MoA^{3}CIR** is designed for fine-grained attribute retrieval in e-commerce, a natural concern is whether the proposed mechanism generalizes to open-domain CIR scenarios involving real-life images and scene-level manipulations. We address this by evaluating **MoA^{3}CIR** on CIRR [25], a widely used open-domain CIR benchmark constructed from real-life images.

Setup. CIRR provides neither query-side nor target-side textual descriptions. To preserve our architecture while ensuring a strictly fair comparison with prior methods, we set all description inputs to empty strings throughout both Phase-1 and Phase-2 training. This guarantees that **MoA^{3}CIR** receives identical inputs to SPRC [2] and CIR-LVLM [34], isolating the contribution of the MoA^3 mechanism itself. For Phase-1 attribute specialization, we reuse the augmented data generated for FashionIQ following the pipeline in Sec. B.2, since CIRR shares the open-domain image distribution.

Main Results. Table 10 reports recall on the CIRR test set. With K=8, our 1B **MoA^{3}CIR** outperforms both SPRC (2B) and CIR-LVLM (7B) on all metrics despite using fewer parameters. Increasing K to 16 or 32 yields only marginal further improvements, indicating that K=8 already provides sufficient expert capacity even for open-domain CIR. This directly addresses the concern that more diverse attributes in CIRR might require substantially larger K.

Table 10: Comparison on the CIRR test set. **MoA^{3}CIR** is evaluated with empty description inputs, identical to baselines, isolating the contribution of MoA^3.

Method	**Param.**	**R@1**	**R@5**	**R@10**	**R@50**
SPRC [2]	2B	51.96	82.12	89.74	97.69
CIR-LVLM [34]	7B	53.64	83.76	90.60	97.93
MoA^{3}CIR (K=8)	1B	54.16	84.03	91.14	98.07
MoA^{3}CIR (K=16)	1B	**54.37**	84.58	91.71	98.29
MoA^{3}CIR (K=32)	1B	54.28	**84.65**	**91.92**	**98.49**

Ablation: Source of Gains. To further isolate where the gains come from, we compare four variants of **MoA^{3}CIR** on CIRR (Table 11): with or without source descriptions (generated using the DQU-CIR image-captioning pipeline), and with or without Phase-1 attribute specialization. Two observations follow. *(i) The architectural design itself is effective.* Without Phase-1 and without source descriptions, our 1B model already reaches comparable performance to CIR-LVLM's 7B model (53.52 vs. 53.64 at R@1), confirming that the grouped expert attention with dynamic routing alone provides competitive gains. *(ii) Phase-1 expert specialization transfers across domains.* Adding Phase-1 brings $+0.64$ pp at R@1 (53.52 $\rightarrow$ 54.16), demonstrating that the attribute specialization induced by augmented data transfers from fashion to scene-level manipulations on real-life images. The marginal effect of source descriptions ($+0.37$ pp at R@1) is consistent with our analysis on ECCIR: the core gain stems from MoA^3, not from the additional textual modality.

Table 11: Ablation on CIRR. "Source desc." uses DQU-CIR-style image captions; "$-$Phase-1" removes attribute specialization training on augmented data.

Variant	**Param.**	**R@1**	**R@5**	**R@10**	**R@50**
MoA^{3}CIR ($-$source desc., $-$Phase-1)	1B	53.52	83.65	90.58	97.96
MoA^{3}CIR ($-$source desc.)	1B	54.16	84.03	91.14	98.07
MoA^{3}CIR (w. source desc., $-$Phase-1)	1B	53.87	83.98	90.82	98.12
MoA^{3}CIR (w. source desc.)	1B	**54.53**	**84.41**	**91.49**	**98.24**

These results confirm that the grouped expert and dynamic routing mechanisms in MoA^3, although designed with fine-grained attribute modeling in mind, are not restricted to the fashion domain. The mechanism transfers to open-domain CIR with scene-level manipulations, and the gains are attributable to the architectural design rather than to the additional textual modality.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction enumerate three contributions including the ECCIR benchmark, the self-reflective data augmentation pipeline, and the MoA^3 mechanism, each of which is supported by experimental and theoretical results in Sections 3 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated Limitation section (Section 4) discusses the one-time computational cost of the augmentation pipeline, the dependence on item descriptions in e-commerce settings, and the current restriction of ECCIR to fashion-related categories.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results (Proposition C.1, Theorem C.4, Remark C.5) are stated with explicit assumptions and accompanied by complete proofs in Appendix C, with the corresponding informal claims cross-referenced in the main text (Section 2.3).

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 3.1.3 specifies the backbone, hyperparameters, optimizer, and training schedule; the augmentation pipeline is fully described in Section 2 and Appendix B with prompts (Appendix G–H) and pseudocode (Algorithm 1).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Both the ECCIR dataset and the source code (including the augmented data used for Phase-1 training) will be open-sourced upon acceptance to facilitate reproduction.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Data splits, evaluation metrics, optimizer choice, learning rates, batch size, and dropout are reported in Section 3.1; further training-strategy details are provided in Section 2.5 and the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Following common practice in supervised CIR work (e.g., SPRC, CIR-LVLM, MAPNet), we report single-run recall values, since multi-seed runs over multiple datasets and augmentation rounds are computationally prohibitive; the consistency of trends across

three benchmarks and across ablation settings nonetheless supports the robustness of our findings.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 3.1.3 reports the GPU type and approximate runtime for both the augmentation pipeline and model training, and the Limitation section additionally discloses the total one-time cost of the augmentation step.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics; ECCIR is collected from a real e-commerce platform with user-identifiable information removed and sensitive terms filtered prior to release.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Limitation section briefly discusses societal considerations, noting that the work targets composed image retrieval and shares the standard considerations of general image retrieval systems, without introducing additional foreseeable negative impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: For the released ECCIR dataset, we apply a multi-step filtering pipeline (privacy stripping, sensitive-term removal, spelling correction, and human spot-checking) before release, as described in Appendix B.1; the synthetic images generated by the augmentation pipeline are restricted to fashion items and pose no obvious misuse risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All existing assets used in this work including CLIP, Qwen2.5-VL, Qwen-Image-Edit, FashionIQ, Shoes, and CIRR are properly cited in the paper, and we adhere to their respective licenses and terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The newly introduced ECCIR dataset will be released upon acceptance, together with documentation describing its schema, construction pipeline, category distribution, and intended use, as detailed in Appendix B and E.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Human annotation is involved in the construction and verification of ECCIR (manual verification of manipulation texts and majority-vote spot-checking of descriptions); in-house annotators were compensated above the local minimum wage, and the annotation guidelines are provided in Appendix G.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve human subjects research beyond paid annotation labor on publicly observable e-commerce listings; no IRB-equivalent approval was required under our institution's guidelines.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Large vision–language models (Qwen2.5-VL for manipulation-text generation and self-reflective verification, Qwen-Image-Edit for target-image synthesis) are core components of the proposed augmentation pipeline; their roles, prompting templates, and verification protocol are documented in Section 2 and Appendices G–H.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.