TEMA: Anchor the Image, Follow the Text for Multi-Modification Composed Image Retrieval

Zixu Li[1] **Yupeng Hu**[1*] **Zhiheng Fu**[1] **Zhiwei Chen**[1] **Yongqi Li**[2] **Liqiang Nie**[3]
[1] School of Software, Shandong University
[2] Department of Computing, Hong Kong Polytechnic University
[3] School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)
{lizixu.cs, fuzhiheng8, zivczw, liyongqi0, nieliqiang}@gmail.com huyupeng@sdu.edu.cn

Abstract

Composed Image Retrieval (CIR) is an important image retrieval paradigm that enables users to retrieve a target image using a multimodal query that consists of a reference image and modification text. Although research on CIR has made significant progress, prevailing setups still rely simple modification texts that typically cover only a limited range of salient changes, which induces two limitations highly relevant to practical applications, namely **Insufficient Entity Coverage** and **Clause-Entity Misalignment**. In order to address these issues and bring CIR closer to real-world use, we construct two instruction-rich multi-modification datasets, **M-FashionIQ** and **M-CIRR**. In addition, we propose **TEMA**, the Text-oriented Entity Mapping Architecture, which is the first CIR framework designed for multi-modification while also accommodating simple modifications. Extensive experiments on four benchmark datasets demonstrate that TEMA's superiority in both original and multi-modification scenarios, while maintaining an optimal balance between retrieval accuracy and computational efficiency. Our codes and constructed multi-modification dataset (**M-FashionIQ and M-CIRR**) are available at https://github.com/lee-zixu/ACL26-TEMA/

1 Introduction

Composed Image Retrieval (CIR) (Chen et al., 2026; Zhang et al., 2026; Fu et al., 2025; Chen et al., 2025; Huang et al., 2025) uses a "reference image + modification text" query to locate target images that satisfy the user's retrieval intent within large image collections. As shown in Figure 1(a), unlike text only retrieval, CIR leverages the reference image to provide visual priors such as appearance, layout, and style, while the modification text specifies how to modify it relative to this reference anchor. CIR models often need to preserve the subject and style while imposing multiple attribute and relation constraints on multiple entities in order to achieve precise retrieval (Hu et al., 2021a,b; Liu et al., 2018; Li et al., 2026; Xie et al., 2025). This paradigm has substantial application value in multimodal learing (Li et al., 2025; Song et al., 2022; Zhang et al., 2026; Wang et al., 2026; Yang et al., 2026; Yuan and Zhang, 2025), human-computer interaction (Xu et al., 2025; Lin et al., 2026; Wang et al., 2025; Zhang et al., 2026; Xie et al., 2026; Kaifosh and Reardon, 2025), and it has attracted broad attention in recent years (Tang et al., 2025).

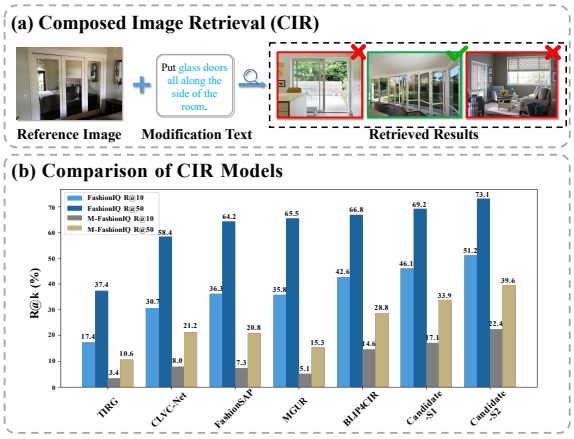

Figure 1: (a) Example of traditional CIR, and (b) Performance comparison of representative baselines on CIR datasets in original and multi-modification scenarios (all models are trained on original FashionIQ).

In recent years, although research on CIR has made notable progress, prevailing setups (Li et al., 2025, 2026; Chen et al., 2026; Qiu et al., 2026) still rely on short modification texts that typically cover only a small number of salient changes (Li et al., 2025; Ray et al., 2023). This reliance gives rise to two limitations that are highly relevant to practical applications. **(1) Insufficient Entity Coverage**. When multiple to-be-modified entities are present, the training signal tends to concentrate on

* Corresponding Author: Yupeng Hu

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 24421–24442
July 2-7, 2026

salient regions and omit some entities. In the modification texts used for CIR, detailed descriptions account for more than 80% on average, and additional portions are occupied by prepositions and conjunctions. The proportion explicitly referring to to-be-modified entities is small and can be easily ignored by models. **(2) Clause-Entity Misalignment**. In real applications, CIR is often used in image retrieval scenarios with stringent requirements for fine-grained details, whereas scenarios with lower requirements can be handled by unimodal image retrieval. It is therefore common for multiple modification clauses to constrain the same entity (e.g., simultaneously modifying the hem, shoulder embellishment, and belt of a dress), or for a single modification clause to constrain multiple entities of the same type (e.g., changing three retrievers in the image into huskies).

However, we regret to observe that existing CIR models struggle to meet multi-modification requirements in practical settings. As shown in Figure 1(b), we convert samples from the FashionIQ validation set into a multi-modification form and evaluate several strong CIR baselines (Vo et al., 2019; Wen et al., 2021; Han et al., 2023; Chen et al., 2024; Liu et al., 2024a,b; Li et al., 2025). We find a pronounced performance drop under multi-modification scenarios, which is likely due to the lack of multi-modification annotations during training and a heightened susceptibility to the limitations of Insufficient Entity Coverage and Clause-Entity Misalignment. To address these issues and realize the two core capabilities of entity coverage and multi-clause aggregation, thereby advancing CIR toward real-world applications, we propose a complementary **data** and **model** solution.

Fresh Data Annotation: Without altering the original reference and target images or evaluation protocols, we expand the modification texts in FashionIQ (Wu et al., 2021) and CIRR (Liu et al., 2021) into instruction-intensive multi-modification versions, constructing the **M-FashionIQ** and **M-CIRR** datasets. The new data replaces the original short texts with **Multi-Modification Texts (MMT)**, generated by MLLM and verified by human annotators, explicitly presenting constraint structures with multiple entities and clauses. This approach provides more comprehensive entity clues and denser training signals for the "Insufficient Entity Coverage" challenge, while offering a test environment more aligned with practical applications for the "Clause-Entity Misalignment" challenge. The aim is to create benchmarks that are closer to real-world scenarios, rather than simply improving performance. The multi-modification annotations, though increasing the complexity of understanding, are more aligned with practical applications and contribute to the real-world deployment of CIR.

Novel Model Architecture: We propose the first CIR framework for multi-modifications while accommodating simple modifications, named **TEMA** (**T**ext-oriented **E**ntity **M**apping **A**rchitecture). To address the "Insufficient Entity Coverage" problem, we design the *MMT Parsing Assistant (PA)*, which enhances the exposure and coverage of modified entities during training through summarization and consistency checks. During inference, the PA is disabled to avoid additional dependencies and delays. To tackle the "Clause-Entity Misalignment" issue, we design an *MMT-oriented Entity Mapping module (EM)* that introduces learnable queries, consolidating multiple clauses of the same entity on the text side and aligning them with the corresponding visual entities on the image side. This stabilizes the modeling of "one-to-many" relationships without explicit alignment annotations. The collaboration of these two modules enables the model to acquire transferable entity coverage and aggregation abilities while remaining robust to multi-granularity multimodal query instructions.

The main contributions are as follows:

- We find that existing CIR models struggle with the multi-modification requirements in practical scenarios. To address this, we construct two instruction-intensive multi-modification datasets, M-FashionIQ and M-CIRR.
- We propose the first CIR framework that accommodates both original and multi-modification scenarios, TEMA, which can learn transferable entity coverage and aggregation abilities during training while maintaining robustness for multi-granularity multimodal query instructions.
- Our proposed TEMA achieves optimal performance in both original (FashionIQ and CIRR datasets) and multi-modification (M-FashionIQ and M-CIRR datasets) CIR scenarios. A large number of quantitative and qualitative experiments validate its superiority.

2 Related Work

Composed Image Retrieval. This task aims to retrieve target images based on a reference image and modification text. Existing CIR methods

can be broadly categorized into traditional models (Vo et al., 2019; Chen et al., 2024, 2020; Lee et al., 2021; Wen et al., 2021) and VLP-based models (Baldrati et al., 2022a,b; Wen et al., 2023; Chen et al., 2024; Yang et al., 2024). Recently, the rapid advancement of Large Vision-Language Models (LVLMs) (He et al., 2024; Sun et al., 2023; Liu et al., 2025; Pu et al., 2025a,b) and visual foundation models (Dong et al., 2026; Wang et al., 2026; Chang et al., 2024; Li et al., 2025) has dramatically enhanced cross-modal understanding (Jiang et al., 2024; Lu et al., 2023; Li et al., 2025, 2024; Liu et al., 2022, 2026; Dong et al., 2025; Bi et al., 2025; Wu et al., 2025) and instruction-following capabilities (Zheng et al., 2025; Bi et al., 2025a,b; Lu et al., 2024; Wang et al., 2025; Sun et al., 2024). However, despite the powerful representation abilities brought by these advancements, existing CIR frameworks are mostly limited to addressing simple modification requests. To bridge this gap, our proposed multi-modification datasets facilitate more comprehensive modification descriptions through MMT, thereby better satisfying users' detailed, instruction-driven retrieval intentions in practical application scenarios.

Multi-object and fine-grained annotations. As user retrieval needs become more complex (Tian et al., 2025; Huang et al., 2024, 2023; Tian et al., 2025; Xu et al., 2025; Lu et al., 2024; Zhou et al., 2025; Lu et al., 2025; Huang et al., 2025; Liu et al., 2024; Sun et al., 2023), modification text annotations must evolve to support multi-object and fine-grained descriptions. Driven by the strong semantic parsing and reasoning capabilities of modern Large Language Models (LLMs), exploring complex, multi-granular textual modifications has become a new trend. While several CIR studies have made progress in this direction, common limitations persist. Works like Cola (Ray et al., 2023), MagicLens (Zhang et al., 2024), and ReT-2 (Caffagni et al., 2025) primarily examine multi-object interference, whereas MIST (Zhou et al., 2025) and early CTI-IR (Zhang et al., 2020) construct training data without considering multi-modification requirements. Even methods like FineCIR (Li et al., 2025), which explicitly parses modification semantics, fail to guarantee that the modification text covers all to-be-modified entities. In contrast, our TEMA explicitly targets multi-modification CIR by introducing MMT together with PA and EM, effectively bridging the gap in explicitly modeling multi-entity to multi-clause alignment.

3 Multi-Modification CIR Datasets Construction

In this section, we introduce the constructed multi-modification dataset. Note that our goal is to create benchmarks that are closer to real-world scenarios rather than simply improving performance.

To promote CIR tasks closer to practical application, we construct two datasets: a fashion-domain dataset M-FashionIQ, and an open-domain dataset M-CIRR. They are built upon the classic CIR datasets FashionIQ and CIRR. Leveraging the automatic annotations generated by MLLM, we incorporate a manual review process to ensure high quality of the datasets. Our empirical results demonstrate that this combined approach effectively captures more nuanced users' modification requests while minimizing false-negative samples, thus enhancing the datasets' suitability for training and testing in multi-modification scenarios.

Data Construction. Since the primary distinction between multi-modification datasets and original CIR datasets lies in the modification text, we select two classical CIR datasets (FashionIQ and CIRR) and re-label the modification text in the original triplets. We note the powerful multimodal comprehension capabilities (Huang et al., 2026, 2025) of Multimodal Large Language Models (MLLMs) (Meta, 2024), and therefore we utilize an MLLM, Llama 3.2 (Meta, 2024) as our primary automatic annotation tool. Specifically, as illustrated in Figure 2 (2) and (3), we extract triplet samples from the original datasets and utilize the reference-target image pairs as the input to Llama 3.2. Simultaneously, we design detailed prompts that necessitate the MMT generated by the Llama 3.2 to faithfully adhere to the original modification texts while articulating refined modification requests that specify the intricacies of transforming the reference image to the target image, as shown in **Q** in Figure 2 (3). This requires the MLLM to understand both the reference and target images, and describe their differences, outputting the candidate MMT.

Moreover, since the two datasets belong to diverse domains, we also design prompts tailored to the specific characteristics of each dataset. For the FashionIQ, we require the MMT generated by Llama 3.2 to focus on various aspects of clothing (*e.g.*, shape, color). In contrast, for the CIRR, we emphasize the different objects present within the open scenario. Such tailored prompts maximize

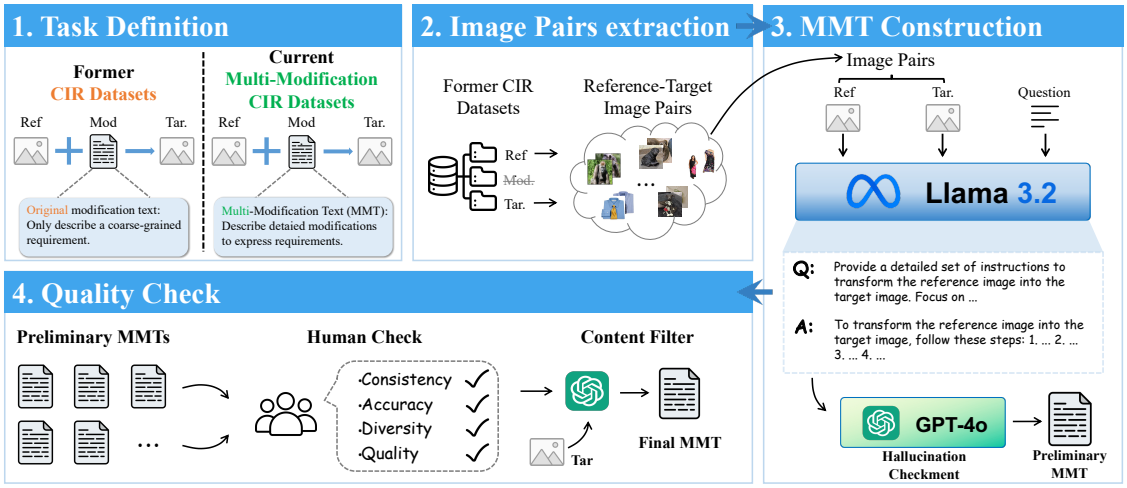

Figure 2: Pipeline of the construction of our proposed multi-modification CIR datasets.

attention on the unique dataset characteristics, ensuring that the generated MMT closely aligns with the authentic modification texts encountered in real retrieval scenarios. Following the above process, we obtain the automatically generated MMT. Besides, we provide the detailed prompts used for both datasets, along with comparative analysis of the impacts on the MMT generated from various prompts (detailed in **Appendix** F).

After obtaining the MMT, we further refine the text output. Considering the hallucination issues (Huang et al., 2023) in current MLLMs, we specifically aim to eliminate hallucinated content embedded within the MMT by the Llama 3.2. Specifically, we use GPT-4o (Brown et al., 2020) (note that other large language models such as Llama-3 (Meta, 2024) can also achieve similar results) to perform a hallucination check on the previously obtained MMT. This process detects and removes any obvious hallucinated content in the text, resulting in the preliminary MMT that can be further processed for the quality check process.

Quality Check. After obtaining the Preliminary MMT, we further adopt a hybrid quality check process involving both human and machine efforts to ensure the overall quality of the MMTs. Specifically, to reduce the workload of human annotators, we first conduct a manual review solely based on textual content, without referencing the associated images. In this stage, a team of 10 research assistants is instructed to examine and revise the texts from four perspectives, including *Consistency*, *Accuracy*, *Diversity*, and *Quality*, which are detailed in **Appendix** B.1. While ensuring the linguistic quality of each MMT, it is equally important to verify whether the modification is faithful to the corresponding reference image. To address this issue, we introduce a *Content Filter* stage following the manual refinement (detailed in **Appendix** B.2).

Positive and negative samples. Essentially, the positive samples in our multi-modification dataset are justifiable since we directly replace the original modification texts with MMT while retaining the original reference and target images in each triplet, and the generated MMT remains faithful to the original modification texts. Furthermore, as MMT provides more precise descriptions of the differences between the reference and target images while encompassing the original modification intent, our empirical evidence indicates that this extension method is effective and mitigates the issue of false negatives that originally existed in the CIR task datasets (detailed in **Appendix** G.1).

We present the dataset statistics in **Appendix** A.1 and compare it with the original CIR dataset. And in **Appendix** A.2, we introduce evaluation metrics.

4 Method

To tackle CIR with multi-modification, we propose a **T**ext-oriented **E**ntity **M**apping **A**rchitecture (**TEMA**), which focuses on understanding modification intentions in MMT, enhancing to-be-modified entity coverage, and exploring clause–entity alignment in multimodal queries to meet fine-grained retrieval needs. As shown in Figure 3, TEMA comprises two main components: *1) MMT Parsing Assistant (PA)*, which includes an *LLM-based text summarizer* and *a Consistency Detector* to extract to-be-modified entities from MMT and perform entity coverage checks to enhance feature exposure (**used only during training**, and detailed in Sec. 4.2); *2) MMT-oriented Entity Mapping (EM)*, which consists of *Textual & Visual Entity Mapping* to aggregate multiple MMT clauses

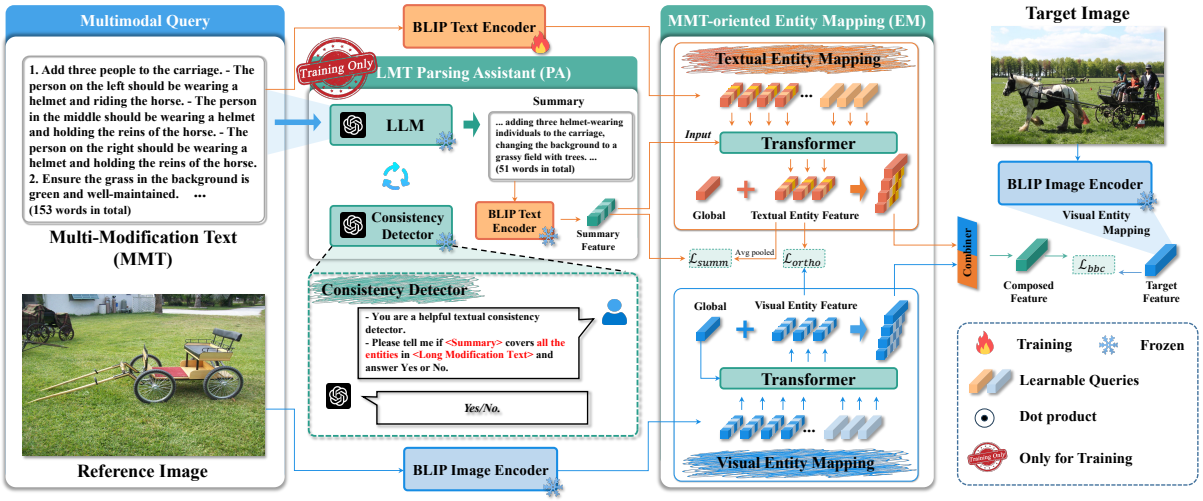

Figure 3: Overall architecture of our proposed TEMA.

related to the same entities, guided by the summary (detailed in Sec. 4.3). We begin with preliminaries in Sec. 4.1, then we elaborate on TEMA's modules.

4.1 Preliminaries

Given a dataset $\mathcal{T} = \{(x_r, t_m, x_t)_n\}_{n=1}^{N}$ of N triplets, where each triplet consists of a reference image x_r, a corresponding MMT t_m, and a target image x_t, the CIR task aims to retrieve x_t based on the composition of x_r and t_m. The model is trained to learn a shared embedding space where the multimodal query (x_r, t_m) is mapped close to its target x_t. Formally, this objective is $\mathcal{F}(x_r, t_m) \rightarrow \mathcal{F}(x_t)$, where $\mathcal{F}(\cdot)$ denotes the learned embedding function for both image and text. The training minimizes the distance between $\mathcal{F}(x_r, t_m)$ and $\mathcal{F}(x_t)$, while ensuring non-matching pairs are pushed apart in the embedding space.

4.2 MMT Parsing Assistant (PA)

Given that MMT contains extensive modification details with sparsely mentioned entities that may be ignored by the model, we propose the PA module to maintain entity focus. It comprises an LLM-based text summarizer for to-be-modified entity parsing and a *Consistency Detector* for entity coverage checking, operating only during training.

LLM-Based Text Summarizer. Specifically, considering the exceptional text comprehension capabilities of LLMs, we leverage an LLM (gpt-3.5-turbo (Brown et al., 2020)) to generate MMT summaries. We use a simple prompt to include all the to-be-modified entities in the summary, as follows,

```
{MMT} denotes the multi-modification text.
Please summarize it into a sentence that
covers all the to-be-modified entities.
```

Consistency Detector. To address potential LLM hallucinations (Farquhar et al., 2024; Huang et al., 2023), we implement a *Consistency Detector* that verifies the summary's entity coverage. Specifically, we use the LLM (gpt-3.5-turbo (Brown et al., 2020)) as a Consistency Detector (with a detailed prompt provided in **Appendix** F) to check whether the summary includes all to-be-modified entities from the MMT, while ensuring no extraneous entities. If inconsistencies are detected, the summary is iteratively refined until it passes verification, yielding the final summary t_s. The summary features are then extracted using a frozen BLIP text encoder $\Phi_{\mathbb{T}}$, formulated as:

$$\mathbf{E}_s = \Phi_{\mathbb{T}}(t_s). \tag{1}$$

We show the quality of the summary generated by the PA module in Section 5.6.

4.3 MMT-oriented Entity Mapping (EM)

Due to the numerous modification details in the MMT, a single to-be-modified entity may correspond to multiple modification clauses. To avoid clause–entity misalignment, we design the MMT-oriented Entity Mapping (EM) module based on PA. It extracts the one-to-many correspondence between entities and MMT clauses, integrating the modification requirements. Specifically, EM incorporates Textual and Visual Entity Mapping components. The textual EM consolidates multiple MMT clauses corresponding to the same to-be-modified entity, guided by the summary. Moreover, to ensure comprehensive entity information preservation in the text tokens generated by EM, we propose a summary-guided distillation strategy, which promotes the generated text tokens to closely align with the to-be-modified entities parsed by PA.

Feature Extracting. Specifically, we first extract the features of the reference image and MMT. Due to the input token limits of CLIP text encoder, we use BLIP (Li et al., 2022), which has been proven effective on CIR task (Liu et al., 2024a,b), to extract the global feature $\mathbf{E}_r^g \in \mathbb{R}^D$ and local feature $\mathbf{E}_r^l \in \mathbb{R}^{C\times D}$ of the reference image x_r, formulated as,

$$\mathbf{E}_r^g = \Phi_{\mathbb{I}}^g\left(x_r\right), \mathbf{E}_r^l = \mathrm{FC}_{\mathbb{I}}\left(\Phi_{\mathbb{I}}^l\left(x_r\right)\right), \quad (2)$$

where D is the hidden dimension. $\Phi_{\mathbb{I}}^g$ and $\Phi_{\mathbb{I}}^l$ are the last and penultimate layers of the BLIP image encoder, respectively. $\mathrm{FC}_{\mathbb{I}}$ is to align the hidden dimension of the local feature and global feature. Similarly, we use BLIP to extract the global feature $\mathbf{E}_m^g$ and local feature $\mathbf{E}_m^l$ of MMT, and the global feature $\mathbf{E}_t^g$ and local feature $\mathbf{E}_t^l$ of target image.

Textual & Visual Entity Mapping. To extract the one-to-many correspondence between to-be-modified entities and MMT clauses, we introduce a set of learnable queries $\mathbf{a}_q = \{a_1, ..., a_k\}$, which, along with the summary feature $\mathbf{E}_s$ (from Eqn (2)) and MMT local features $\mathbf{E}_m^l$, serve as inputs to the transformer model. Since the summary feature includes all to-be-modified entities with minimal details, the learnable queries aggregate the corresponding MMT clauses for the same entity, guided by the summary, formulated as,

$$\hat{\mathbf{a}}_q = \mathrm{Transformer}\left(\left[\mathbf{E}_s, \mathbf{E}_m^l, \mathbf{a}_q\right]\right), \quad (3)$$

where $\hat{\mathbf{a}}_q \in \mathbb{R}^{N\times D}$ denotes the textual entity feature, representing N aggregated entity-clause features from N channels of $\mathbf{a}_q$.

For the reference image, we use a similar aggregation process, but with the global features of the reference image instead of the summary feature. Specifically, we also utilize learnable queries $\mathbf{b}_q = b_1, ..., b_k$ and use the local features $\mathbf{E}_r^l$ and global features $\mathbf{E}_r^g$ of the reference image as inputs to the transformer, adaptively aggregating corresponding feature channels for the same visual entity, formulated as follows,

$$\hat{\mathbf{b}}_q = \mathrm{Transformer}\left(\left[\mathbf{E}_r^g, \mathbf{E}_r^l, \mathbf{b}_q\right]\right), \quad (4)$$

where $\hat{\mathbf{b}}_q \in \mathbb{R}^{N\times D}$ is the visual entity feature.

Multimodal Query Composition. So far, we have obtained the textual and visual entity features. To improve the model's multi-granularity perception of multimodal queries, we concatenate these features with the global features of the reference image and the MMT, resulting in the final entity feature. For the MMT, the final entity feature is $\hat{\mathbf{E}}_m = [\mathbf{E}_m^g, \hat{\mathbf{a}}_q] \in \mathbb{R}^{(1+N)\times D}$. For the reference image, it is $\hat{\mathbf{E}}_r = [\mathbf{E}_r^g, \hat{\mathbf{b}}_q] \in \mathbb{R}^{(1+N)\times D}$, where N is the number of channels in the learnable queries.

Then, following previous CIR methods (Wen et al., 2023; Liu et al., 2024a), we use the same composition module for multimodal query features $\hat{\mathbf{E}}_m$ and $\hat{\mathbf{E}}_r$ to get composed feature $\mathbf{E}_c$. Finally, we introduce the loss functions (including the summary-guided distillation strategy, orthogonal regularization, and batch-based classification loss) and train-inference phases of TEMA in **Appendix** C. And we represented the algorithm of TEMA's processing flow in **Appendix** E.

5 Experiments

In this section, we discuss the detailed experiments.

5.1 Experimental Settings

Evaluation. We use our proposed multi-modification datasets for training and evaluation, while choosing the recall at rank K (R@K) as the evaluation metric, quite similar to the previous CIR task (Chen et al., 2020). The datasets include a fashion-domain dataset, M-FashionIQ, and an open-domain dataset, M-CIRR. For the evaluation of both, we use the validation splits. For M-FashionIQ, we employ R@10, R@50 and their category-wise averages. And M-CIRR assessment included R@k (k=1, 5, 10), R$_{subset}$@k (k=1, 2) and the average (R@5 + R$subset$@1) / 2. In addition, we provide a detailed description of the above datasets (detailed in **Appendix** A).

Implementation Details. We utilize BLIP (Li et al., 2022) as the backbone and freeze the image encoder. We train TEMA using the AdamW optimizer with an initial learning rate of $2e$-5. The batch size is set to 64, and we maintain a dimension of 256. The channel number N of the learnable queries is set to 3 for both M-FashionIQ and M-CIRR. Through a simple grid search, we set κ to 0.6 and μ to 0.2 in Eqn (8). All experiments are accomplished on a single NVIDIA A40 GPU with 48 GB memory.

5.2 Method Comparison

We conducted a comprehensive evaluation of the proposed TEMA model against several significant baselines using the two constructed datasets, *i.e.*, M-FashionIQ and M-CIRR. We also provide results

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

Method	Text Encoder	M-FashionIQ								M-CIRR					
		Dresses		Shirts		Tops&Tees		Avg		R@k			R_{subset}@k		Avg
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	k=1	k=5	k=10	k=1	k=2	
Text-only	BLIP	24.72	47.94	26.88	49.67	27.25	50.22	26.28	49.28	23.50	50.10	67.30	49.28	67.70	49.69
Text+Image	BLIP	38.96	60.57	34.48	56.39	41.50	63.13	38.31	60.03	36.46	70.94	81.57	64.73	80.53	67.84
Traditional Model-Based Methods															
TIRG	LSTM	7.88	14.35	9.66	18.30	10.07	21.51	9.20	18.05	9.68	30.73	47.81	14.92	36.31	22.83
CLVC-Net	LSTM	14.86	27.55	16.18	31.05	17.14	34.27	16.06	30.96	11.60	36.22	50.26	19.67	40.82	27.95
FashionViL	BERT	22.07	47.81	22.32	46.93	29.08	55.62	24.49	50.12	-	-	-	-	-	-
MGUR	RoBERTa	21.42	45.27	16.58	37.59	23.92	49.16	20.64	44.01	-	-	-	-	-	-
VLP Model-Based Methods															
FashionSAP	ALBEF	25.63	51.81	26.89	51.82	32.33	59.97	28.28	54.53	-	-	-	-	-	-
BLIP4CIR	BLIP	40.97	63.28	37.81	59.10	44.19	64.94	40.99	62.44	39.92	74.04	84.07	67.79	82.21	70.92
BLIP4CIR+Bi	BLIP	41.51	63.23	37.51	57.92	43.32	65.00	40.78	62.05	41.24	75.61	84.21	69.46	82.89	72.54
Candidate	BLIP	43.30	65.36	47.96	65.53	50.87	69.23	47.38	66.71	42.03	75.92	84.61	69.58	83.84	72.75
TEMA (Ours)	BLIP	**45.74**	**69.48**	**50.35**	**71.26**	**55.67**	**75.52**	**50.59**	**72.09**	**45.29**	**79.46**	**88.17**	**72.05**	**86.52**	**75.76**

on traditional CIR benchmarks in **Appendix** D.3. Since the source codes for some methods were not accessible, we selected several open-source baselines (MGUR (Chen et al., 2024), BLIP4CIR (Liu et al., 2024a), Candidate (Liu et al., 2024b), etc.). We retrained and tested them according to their original settings on two multi-modification datasets. It is important to note that, due to the token length limitations of the CLIP text encoder, we did not use it as a backbone. The results are presented in Table 1, leading to the following conclusions: 1) Our proposed TEMA achieves superior performance on both M-FashionIQ and M-CIRR, indicating its excellent generalization capabilities and robust comprehension of queries in both fashion and open-domain contexts. 2) TEMA demonstrates a significant performance advantage over the baselines, which also utilize BLIP as their backbone. This superiority may be attributed to the enhancements provided by our PA and EM modules, which improve the model's ability to grasp the nuances of MMT. 3) The performance of BLIP-based models markedly surpasses that of models employing traditional backbones, suggesting that, compared to conventional architectures (such as ResNet and LSTM), VLP-based models are more adept at understanding complex MMT.

5.3 Ablation Study

In this section, we introduce the ablation study of our proposed TEMA with different variants, as shown in Table 2. The compared variants are as follows. • **w/o PA.** We train TEMA without the PA module. • **w/o CD.** We only ablate the Consistency Detector (CD) of PA, i.e., we don't check the summary generated by LLM. • **w/o EM, w/o EM_txt, and w/o EM_img.** We first remove the entire EM, using the summary for composition instead, i.e., w/o EM. Further, we perform EM only for one modality, i.e., w/o EM_txt and w/o EM_img. • **w/o Summ.** We don't use the Summary-guided Distillation strategy in this setup. Instead we only perform the other two losses. • **w/o Ortho, w/o Ortho_txt, and w/o Ortho_img.** To investigate the role of Orthogonal Regularization for entity features, we removed it for both the textual entity feature and visual entity feature. Furthermore, we investigate Orthogonal Regularization by only performing for visual or textual entity features.

Table 2: Ablation study on M-FashionIQ and M-CIRR datasets. We compute Avg-R@10, R@50 for M-FashionIQ, and Avg (mean of R@5 and R_{subset}@1) for M-CIRR, respectively.

Method	M-FashionIQ				M-CIRR	
	R@10	Δ	R@50	Δ	Avg	Δ
MMT Parsing Assistant (PA)						
w/o PA	47.80	-2.79	69.83	-2.26	71.59	-4.17
w/o CD	49.14	-1.45	70.96	-1.13	73.87	-1.89
MMT-oriented Entity Mapping (EM)						
w/o EM	45.41	-5.18	68.18	-3.91	70.99	-4.77
w/o EM_txt	46.11	-4.48	68.25	-3.84	71.20	-4.56
w/o EM_img	46.17	-4.42	68.72	-3.37	71.64	-4.12
Loss Functions						
w/o Summ	49.40	-1.19	71.14	-0.95	74.16	-1.60
w/o Ortho	49.38	-1.21	71.58	-0.51	75.02	-0.74
w/o Ortho_txt	48.14	-2.45	69.96	-2.13	73.39	-2.37
w/o Ortho_img	48.93	-1.66	70.44	-1.65	73.61	-2.15
TEMA	**50.59**	-0.00	**72.09**	-0.00	**75.76**	-0.00

From the ablation results of TEMA in Table 2, we have following four observations. 1) Both w/o PA and w/o CD are inferior to TEMA. In particular, removing PA causes very substantial performance degradation. This is reasonable that the PA module is a powerful training aid, and the summary generated by the PA module serves as a guide for the textual entity feature and facilitates the training of the EM module to aggregate the visual entities and multiple clauses within MMT, respectively, demonstrating the importance of using the MMT

Parsing Assistant. 2) w/o EM, w/o EM_txt, and w/o EM_img all perform worse than TEMA, and removing any of the components of EM resulted in a more substantial drop than the other module ablations, indicating that the entity mapping process indeed improved the MMT comprehension for TEMA, by aggregating complex modification clauses to several to-be-modified entities. 3) Both w/o Summ and w/o Ortho are inferior to TEMA, showing their necessity in TEMA's optimization. 4) In w/o Ortho_txt and w/o Ortho_img, the orthogonal regularization is performed on one modality but causes more drastic performance degradation than w/o Ortho. This may be because such an asymmetric process destroys the alignment of the features of two modalities. We also provide more detailed ablation results in **Appendix** D.1.

5.4 Performance on Traditional CIR

To verify the performance of TEMA on traditional CIR, we conducted additional experiments, training and testing TEMA in the settings of original FashionIQ and CIRR datasets. The performance comparison is shown in Table 3. We can observe that TEMA's performance is better than the previous baselines, which proves the powerful generalization ability of TEMA. It not only performs well on the multi-modification CIR benchmark, but also maintains the performance on CIR. Additionally, we represented the full results in **Appendix** D.3.

Table 3: Performance on traditional CIR benchmarks, including FashionIQ and CIRR.

Methods	Backbone	FashionIQ		CIRR
		R@10	R@50	Avg
CASE	BLIP	48.79	70.68	77.50
Candidate	BLIP	51.17	73.13	**80.90**
CoVR-BLIP	BLIP	48.53	70.25	76.81
TEMA	BLIP	**53.02**	**74.20**	80.18

5.5 Sensitivity Analysis

As shown in Figure 4, we evaluate the sensitivity of our proposed TEMA regarding the hyper-parameter κ in Eqn (8) on M-FashionIQ, and the channel number N of learnable queries on both M-FashionIQ and M-CIRR. As the results show, the performance of TEMA initially improves with the increase of κ, reaching an optimal level, after which it gradually declines as κ continues to rise. This is reasonable, as the summary-guided distillation loss $\mathcal{L}_{summ}$ requires a certain weight to enhance the optimization effect. When the value is too high, it may lead to an imbalance among different loss functions. For the channel number of learnable queries, denoted by N, the performance of TEMA first shows an upward trend on both M-FashionIQ and M-CIRR, reaching the optimal. This is because a certain number of channels are needed to correspond to different to-be-modified entities. However, when the value of N becomes too large, performance begins to fluctuate and decline, as an excess of channels may lead to confusion among the to-be-modified entities, thereby reducing retrieval performance.

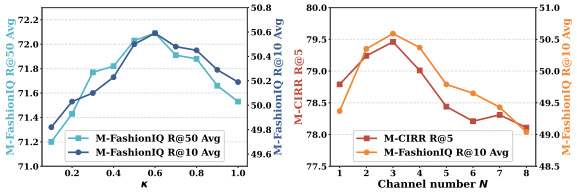

Figure 4: Sensitivity analysis on the hyper-parameter κ and the channel number N of learnable queries.

5.6 Qualitative results for PA module

To investigate the validity of the PA-generated summaries, we present qualitative results from the MMT Parsing Assistant (PA) to verify whether the summary includes all the to-be-modified entities. As shown in Figure 5, we highlight these entities in the MMT, where they appear as subjects in clauses. The summary generated by PA captures all to-be-modified entities in the MMT. For example, in Figure 5(b), the summary accurately identifies entities such as "the breed of the dogs" and "the dog's posture". The key difference from the MMT is the omission of certain detailed descriptions, which condenses the focus on the to-be-modified entities. This approach effectively guides the model, helping it identify the entities while minimizing distractions from lengthy descriptions, conjunctions, and prepositions in the MMT. Consequently, this enhances the model and facilitates the subsequent EM module in aggregating the to-be-modified entities. Additionally, we provide attention visualization results on the summaries in **Appendix** G.2 and more qualitative results in **Appendix** G.3.

6 Conclusion

In this work, we addressed two limitations that were highly relevant to CIR's practical applications, namely Insufficient Entity Coverage and Clause-Entity Misalignment, thereby advancing CIR toward real-world use. We constructed two multi-modification datasets, M-FashionIQ and M-CIRR. In addition, we proposed TEMA, which

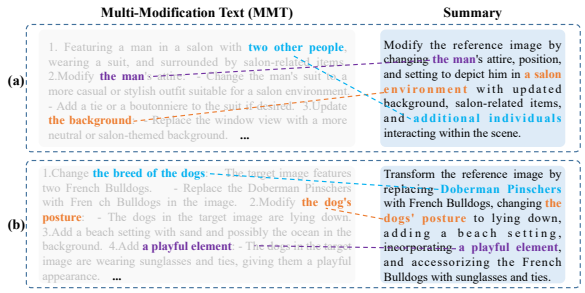

Figure 5: The qualitative results for the PA module, which shows the MMT and corresponding summary. The to-be-modified entities are colored.

was the first CIR framework designed for multi-modification while also accommodating simple modifications. TEMA outperformed previous methods in both original and multi-modification scenarios, showcasing its superiority.

7 Limitations

Although this study constructs the M-FashionIQ and M-CIRR datasets and proposes the TEMA framework, which, through multi-modification text (MMT) parsing and entity mapping, achieves substantial progress in composed image retrieval (CIR) across original and multi-modification scenarios, several limitations remain. First, unlike other data used for pretraining, our CIR datasets for multi-modification scenarios are designed to provide a training and evaluation environment that is closer to real applications, rather than to solely increase model performance. Because the annotations in the constructed datasets are longer, they increase the difficulty for models to understand modification intentions and therefore do not necessarily lead to higher retrieval metrics. Second, the PA module incorporates large language models during training. Although this module is disabled during testing, it still introduces minor computational overhead in training. Finally, consistent with most current CIR studies, the proposed TEMA currently supports only single turn retrieval, and its effectiveness in multi turn interactive CIR scenarios remains to be explored. Future research should address these limitations to enhance the practical utility of our proposed datasets and TEMA model.

8 Ethical Considerations

First, for the constructed datasets, our public release will remove personally identifiable information and prohibits retrieval based on identifiable faces; the license explicitly disallows surveillance uses. We encourage deployments to incorporate abuse detection, rate limiting, keyword and category block lists, and context aware access control policies, and to state permitted uses and prohibitions at the time of releasing data and models. Second, composed image retrieval (CIR) could be repurposed for sensitive settings. We will decline data requests that target such settings and will provide an email address to process takedown requests for problematic content. We will continue to validate and improve this work through careful safety and compliance practices across broader populations and scenarios, thereby ensuring a responsible contribution to the CIR community.

Acknowledgments

This work was supported in partby the National Natural Science Foundation of China, No.:62576195, andNo.:62276155; in part by the Key R&D Program of Shandong Province (Major scientific and technological innovation projects), China, No.: 2025CXGC020101; in part by the China National University Student Innovation & Entrepreneurship Development Program, No.:2025283.

References

Zhiwei Chen, Yupeng Hu, Zhiheng Fu, Zixu Li, Jiale Huang, Qinlei Huang, and Yinwei Wei. 2026. Intent: Invariance and discrimination-aware noise mitigation for robust composed image retrieval. In *AAAI*, volume 40, pages 20463–20471.

Mingyu Zhang, Zixu Li, Zhiwei Chen, Zhiheng Fu, Xiaowei Zhu, Jiajia Nie, Yinwei Wei, and Yupeng Hu. 2026. Hint: Composed image retrieval with dual-path compositional contextualized network. *arXiv preprint arXiv:2603.26341*.

Zhiheng Fu, Zixu Li, Zhiwei Chen, Chunxiao Wang, Xuemeng Song, Yupeng Hu, and Liqiang Nie. 2025. Pair: Complementarity-guided disentanglement for composed image retrieval. In *ICASSP*, pages 1–5. IEEE.

Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Xuemeng Song, and Liqiang Nie. 2025. Offset: Segmentation-based focus shift revision for composed image retrieval. In *ACM MM*, page 6113–6122.

Qinlei Huang, Zhiwei Chen, Zixu Li, Chunxiao Wang, Xuemeng Song, Yupeng Hu, and Liqiang Nie. 2025. Median: Adaptive intermediate-grained aggregation network for composed image retrieval. In *ICASSP*, pages 1–5. IEEE.

Yupeng Hu, Meng Liu, Xiaobin Su, Zan Gao, and Liqiang Nie. 2021a. Video moment localization via deep cross-modal hashing. *IEEE TIP*, 30:4667–4677.

Yupeng Hu, Liqiang Nie, Meng Liu, Kun Wang, Yinglong Wang, and Xian-Sheng Hua. 2021b. Coarse-to-fine semantic alignment for cross-modal moment localization. *IEEE TIP*, 30:5933–5943.

Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *ACM SIGIR*, pages 15–24.

Zixu Li, Yupeng Hu, Zhiwei Chen, Qinlei Huang, Guozhi Qiu, Zhiheng Fu, and Meng Liu. 2026. Retrack: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval. In *AAAI*, volume 40, pages 23373–23381.

Zequn Xie, Chuxin Wang, Yeqiang Wang, Sihang Cai, Shulei Wang, and Tao Jin. 2025. Chat-driven text generation and interaction for person retrieval. In *EMNLP*, pages 5259–5270.

Yanshu Li, Yi Cao, Hongyang He, Qisen Cheng, Xiang Fu, Xi Xiao, Tianyang Wang, and Ruixiang Tang. 2025. M^2IV: Towards efficient and fine-grained multimodal in-context learning via representation engineering. In *Second Conference on Language Modeling*.

Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2022. Transformer tracking with cyclic shifting window attention. In *CVPR*, pages 8791–8800.

Yuanjun Zhang, Fuzel Ahamed Shaik, Suvojit Acharjee, Fahad Khalid, and Mourad Oussalah. 2026. Towards reliable multimodal disaster severity assessment through preference optimization and explainable vision-language reasoning. *Reliability Engineering & System Safety*, page 112674.

Zi-Han Wang, Lam Nguyen, Zhengyang Zhao, Mengyue Yang, Chengwei Qin, Yujiu Yang, and Linyi Yang. 2026. Creativebench: Benchmarking and enhancing machine creativity via self-evolving challenges. *arXiv preprint arXiv:2603.11863*.

Qianyun Yang, Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, and Liqiang Nie. 2026. Stable: Efficient hybrid nearest neighbor search via magnitude-uniformity and cardinality-robustness. *arXiv preprint arXiv:2604.01617*.

Haohan Yuan and Haopeng Zhang. 2025. Understanding llm reasoning for abstractive summarization. *arXiv preprint arXiv:2512.03503*.

Hongshen Xu, Zihan Wang, Zichen Zhu, Lei Pan, Xingyu Chen, Shuai Fan, Lu Chen, and Kai Yu. 2025. Alignment for efficient tool calling of large language models. In *EMNLP*, pages 17787–17803.

Honglin Lin, Chonghan Qin, Zheng Liu, Qizhi Pei, Yu Li, Zhanping Zhong, Xin Gao, Yanfeng Wang, Conghui He, and Lijun Wu. 2026. Scientific image synthesis: Benchmarking, methodologies, and downstream utility. *arXiv preprint arXiv:2601.17027*.

Tongxi Wang, Yang Yu, Qing Wang, and Junlang Qian. 2025. Via score to performance: Efficient human-controllable long song generation with bar-level symbolic notation. *arXiv preprint arXiv:2508.01394*.

Wenyuan Zhang, Xinghua Zhang, Haiyang Yu, Shuaiyi Nie, Bingli Wu, Juwei Yue, Tingwen Liu, and Yongbin Li. 2026. Expseek: Self-triggered experience seeking for web agents. *Preprint*, arXiv:2601.08605.

Zequn Xie, Xin Liu, Boyun Zhang, Yuxiao Lin, Sihang Cai, and Tao Jin. 2026. Hvd: Human vision-driven video representation learning for text-video retrieval. *arXiv preprint arXiv:2601.16155*.

Patrick Kaifosh and Thomas R Reardon. 2025. A generic non-invasive neuromotor interface for human-computer interaction. *Nature*, pages 1–10.

Haomiao Tang, Jinpeng Wang, Yuang Peng, GuangHao Meng, Ruisheng Luo, Bin Chen, Long Chen, Yaowei Wang, and Shu-Tao Xia. 2025. Modeling uncertainty in composed image retrieval via probabilistic embeddings. In *ACL*, pages 1210–1222.

Zixu Li, Zhiwei Chen, Haokun Wen, Zhiheng Fu, Yupeng Hu, and Weili Guan. 2025. Encoder: Entity mining and modification relation binding for composed image retrieval. In *AAAI*, volume 39, pages 5101–5109.

Zixu Li, Yupeng Hu, Zhiwei Chen, Shiqi Zhang, Qinlei Huang, Zhiheng Fu, and Yinwei Wei. 2026. Habit: Chrono-synergia robust progressive learning framework for composed image retrieval. In *AAAI*, volume 40, pages 6762–6770.

Guozhi Qiu, Zhiwei Chen, Zixu Li, Qinlei Huang, Zhiheng Fu, Xuemeng Song, and Yupeng Hu. 2026. Melt: Improve composed image retrieval via the modification frequentation-rarity balance network. *arXiv preprint arXiv:2603.29291*.

Zixu Li, Zhiheng Fu, Yupeng Hu, Zhiwei Chen, Haokun Wen, and Liqiang Nie. 2025. Finecir: Explicit parsing of fine-grained modification semantics for composed image retrieval. *https://arxiv.org/abs/2503.21309*.

Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan Plummer, Ranjay Krishna, and Kate Saenko. 2023. Cola: A benchmark for compositional text-to-image retrieval. *NeurIPS*, 36:46433–46445.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval - an empirical odyssey. In *CVPR*, pages 6439–6448. IEEE.

Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie. 2021. Comprehensive linguistic-visual composition network for image retrieval. In *ACM SIGIR*, pages 1369–1378. ACM.

Yunpeng Han, Lisai Zhang, Qingcai Chen, Zhijian Chen, Zhonghua Li, Jianxin Yang, and Zhao Cao. 2023. Fashionsap: Symbols and attributes prompt for fine-grained fashion vision-language pre-training. In *CVPR*, pages 15028–15038.

Yiyang Chen, Zhedong Zheng, Wei Ji, Leigang Qu, and Tat-Seng Chua. 2024. Composed image retrieval with text feedback via multi-grained uncertainty regularization. *ICLR*.

Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. 2024a. Bi-directional training for composed image retrieval via text prompt learning. In *WACV*, pages 5753–5762.

Zheyuan Liu, Weixuan Sun, Damien Teney, and Stephen Gould. 2024b. Candidate set re-ranking for composed image retrieval with dual multi-modal encoder.

Shuxian Li, Changhao He, Xiting Liu, Joey Tianyi Zhou, Xi Peng, and Peng Hu. 2025. Learning with noisy triplet correspondence for composed image retrieval. In *CVPR*, pages 19628–19637.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *CVPR*, pages 11307–11317.

Zheyuan Liu, Cristian Rodriguez Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, pages 2105–2114. IEEE.

Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020. Image search with text feedback by visiolinguistic attention learning. In *CVPR*, pages 2998–3008. IEEE.

Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. Cosmo: Content-style modulation for image retrieval with text feedback. In *CVPR*, pages 802–812. IEEE.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022a. Effective conditioned and composed image retrieval combining clip-based features. In *CVPR*, pages 21466–21474.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022b. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *CVPR*, pages 4959–4968.

Haokun Wen, Xuemeng Song, Jianhua Yin, Jianlong Wu, Weili Guan, and Liqiang Nie. 2023. Self-training boosted multi-factor matching network for composed image retrieval. *IEEE TPAMI*.

Yanzhe Chen, Huasong Zhong, Xiangteng He, Yuxin Peng, Jiahuan Zhou, and Lele Cheng. 2024. Fashionern: Enhance-and-refine network for composed fashion image retrieval. In *AAAI*, volume 38, pages 1228–1236.

Xingyu Yang, Daqing Liu, Heng Zhang, Yong Luo, Chaoyue Wang, and Jing Zhang. 2024. Decomposing semantic shifts for composed image retrieval. In *AAAI*, volume 38, pages 6576–6584.

Changhao He, Hongyuan Zhu, Peng Hu, and Xi Peng. 2024. Robust variational contrastive learning for partially view-unaligned clustering. In *ACM MM*, pages 4167–4176.

Yuan Sun, Zhenwen Ren, Peng Hu, Dezhong Peng, and Xu Wang. 2023. Hierarchical consensus hashing for cross-modal retrieval. *IEEE TMM*, 26:824–836.

Xu Liu, Yibo Lu, Xinxian Wang, and Xinyu Wu. 2025. Training-free multi-style fusion through reference-based adaptive modulation. *Preprint*, arXiv:2509.18602.

Ruitao Pu, Yang Qin, Xiaomin Song, Dezhong Peng, Zhenwen Ren, and Yuan Sun. 2025a. She: Streaming-media hashing retrieval. In *ICML*.

Ruitao Pu, Yuan Sun, Yang Qin, Zhenwen Ren, Xiaomin Song, Huiming Zheng, and Dezhong Peng. 2025b. Robust self-paced hashing for cross-modal retrieval with noisy labels. In *AAAI*, volume 39, pages 19969–19977.

Haonan Dong, Kehan Jiang, Haoran Ye, Wenhao Zhu, Zhaolu Kang, and Guojie Song. 2026. Neureasoner: Towards explainable, controllable, and unified reasoning via mixture-of-neurons. *Preprint*, arXiv:2604.02972.

Yujia Wang, Yuyan Li, Jiuming Liu, Fang-Lue Zhang, Xinhu Zheng, Neil Dodgson, and 1 others. 2026. Rl-scaniqa: Reinforcement-learned scanpaths for blind 360 {\deg} image quality assessment. *arXiv preprint arXiv:2603.14297*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM TIST*, 15(3):1–45.

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, and 1 others. 2025. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *AAAI*, volume 39, pages 415–423.

Yiyang Jiang, Wengyu Zhang, Xulu Zhang, Xiao-Yong Wei, Chang Wen Chen, and Qing Li. 2024. Prior knowledge integration via llm encoding and pseudo event regulation for video moment retrieval. In *ACM MM*, pages 7249–7258.

Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. 2023. Tf-icon: Diffusion-based training-free cross-domain image composition. In *CVPR*, pages 2294–2305.

Yanshu Li, JianJiang Yang, Ziteng Yang, Bozheng Li, Hongyang He, Zhengtao Yao, Ligong Han, Yingjie Victor Chen, Songlin Fei, Dongfang Liu, and 1 others. 2025. Cama: Enhancing multimodal in-context learning with context-aware modulated attention. *arXiv preprint arXiv:2505.17097*.

Chenglin Li, Qianglong Chen, Zhi Li, Feng Tao, and Yin Zhang. 2024. Videocogqa: A controllable benchmark for evaluating cognitive abilities in video-language models. *arXiv preprint arXiv:2411.09105*.

Peiyang Liu, Xiangyu Xi, Wei Ye, and Shikun Zhang. 2022. Label smoothing for text mining. In *COLING*, pages 2210–2219.

Junlin Liu, Shengnan An, Shuang Zhou, Dan Ma, Shixiong Luo, Ying Xie, Yuan Zhang, Wenling Yuan, Yifan Zhou, Xiaoyu Li, Ziwen Wang, Xuezhi Cao, and Xunliang Cai. 2026. General365: Benchmarking general reasoning in large language models across diverse and challenging tasks. *Preprint*, arXiv:2604.11778.

Haonan Dong, Wenhao Zhu, Guojie Song, and Liang Wang. 2025. AuroRA: Breaking low-rank bottleneck of loRA with nonlinear mapping. In *NeurIPS*.

Jinhe Bi, Danqi Yan, Yifan Wang, Wenke Huang, Haokun Chen, Guancheng Wan, Mang Ye, Xun Xiao, Hinrich Schuetze, Volker Tresp, and 1 others. 2025. Cot-kinetics: A theoretical modeling assessing lrm reasoning process. *arXiv preprint arXiv:2505.13408*.

Yichen Wu, Xu Liu, Chenxuan Zhao, and Xinyu Wu. 2025. Prompt-guided dual latent steering for inversion problems. *Preprint*, arXiv:2509.18619.

Weihua Zheng, Roy Ka-Wei Lee, Zhengyuan Liu, Wu Kui, Aiti Aw, and Bowei Zou. 2025. Ccl-xcot: An efficient cross-lingual knowledge transfer method for mitigating hallucination generation. In *EMNLP Findings*, pages 1768–1788.

Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. 2025a. Llava steering: Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. In *ACL*, pages 15230–15250.

Jinhe Bi, Yifan Wang, Danqi Yan, Aniri, Wenke Huang, Zengjie Jin, Xiaowen Ma, Artur Hecker, Mang Ye, Xun Xiao, Hinrich Schuetze, Volker Tresp, and Yunpu Ma. 2025b. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *Preprint*, arXiv:2502.12119.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. 2024. Mace: Mass concept erasure in diffusion models. In *CVPR*, pages 6430–6440.

Yujun Wang, Jinhe Bi, Yunpu Ma, and Soeren Pirk. 2025. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. *arXiv preprint arXiv:2506.14766*.

Yuan Sun, Yang Qin, Dezhong Peng, Zhenwen Ren, Chao Yang, and Peng Hu. 2024. Dual self-paced hashing for image retrieval. *IEEE TMM*, 26:9619–9629.

Yang Tian, Fan Liu, Jingyuan Zhang, Yupeng Hu, Liqiang Nie, and 1 others. 2025. Core-mmrag: Cross-source knowledge reconciliation for multimodal rag. In *ACL*, pages 32967–32982.

Jincheng Huang, Jialie Shen, Xiaoshuang Shi, and Xiaofeng Zhu. 2024. On which nodes does gcn fail? enhancing gcn from the node perspective. In *Forty-first International Conference on Machine Learning*.

Jincheng Huang, Lun Du, Xu Chen, Qiang Fu, Shi Han, and Dongmei Zhang. 2023. Robust mid-pass filtering graph convolutional networks. In *Proceedings of the ACM Web Conference 2023*, pages 328–338.

Yang Tian, Fan Liu, Jingyuan Zhang, Wei Bi, Yupeng Hu, and Liqiang Nie. 2025. Open multimodal retrieval-augmented factual image generation. *arXiv preprint arXiv:2510.22521*.

Mingzhu Xu, Chenglong Yu, Zexuan Li, Haoyu Tang, Yupeng Hu, and Liqiang Nie. 2025. Hdnet: A hybrid domain network with multi-scale high-frequency information enhancement for infrared small target detection. *IEEE TGRS*.

Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Wai-Kin Kong. 2024. Robust watermarking using generative priors against image editing: From benchmarking to advances. *arXiv preprint arXiv:2410.18775*.

Zihan Zhou, Shilin Lu, Shuli Leng, Shaocong Zhang, Zhuming Lian, Xinlei Yu, and Adams Wai-Kin Kong. 2025. Dragflow: Unleashing dit priors with region based supervision for drag editing. *arXiv preprint arXiv:2510.02253*.

Shilin Lu, Zhuming Lian, Zihan Zhou, Shaocong Zhang, Chen Zhao, and Adams Wai-Kin Kong. 2025. Does flux already know how to perform physically plausible image composition? *arXiv preprint arXiv:2509.21278*.

Jincheng Huang, Yujie Mo, Xiaoshuang Shi, Lei Feng, and Xiaofeng Zhu. 2025. Enhancing the influence of labels on unlabeled nodes in graph convolutional networks. In *Forty-second International Conference on Machine Learning*.

Kaiming Liu, Yunhong Gong, Yu Cao, Zhenwen Ren, Dezhong Peng, and Yuan Sun. 2024. Dual semantic fusion hashing for multi-label cross-modal retrieval. In *IJCAI*, pages 4569–4577.

Yuan Sun, Dezhong Peng, Jian Dai, and Zhenwen Ren. 2023. Stepwise refinement short hashing for image retrieval. In *ACM MM*, pages 6501–6509.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. 2024. MagicLens: Self-supervised image retrieval with open-ended instructions. In *ICML*, pages 59403–59420.

Davide Caffagni, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2025. Recurrence meets transformers for universal multimodal retrieval. *arXiv preprint arXiv:2509.08897*.

Yinan Zhou, Yaxiong Wang, Haokun Lin, Chen Ma, Li Zhu, and Zhedong Zheng. 2025. Scale up composed image retrieval learning via modification text generation. *arXiv preprint arXiv:2504.05316*.

Feifei Zhang, Mingliang Xu, Qirong Mao, and Changsheng Xu. 2020. Joint attribute manipulation and modality alignment learning for composing text and image to image retrieval. In *ACM MM*, pages 3367–3376.

Jincheng Huang, Jie Xu, Xiaoshuang Shi, Ping Hu, Lei Feng, and Xiaofeng Zhu. 2026. Revisiting confidence calibration for misclassification detection in vlms. In *The Fourteenth International Conference on Learning Representations*.

Jincheng Huang, Jie Xu, Xiaoshuang Shi, Ping Hu, Lei Feng, and Xiaofeng Zhu. 2025. The final layer holds the key: A unified and efficient gnn calibration framework. *arXiv preprint arXiv:2505.11335*.

Meta. 2024. The llama 3 herd of models.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

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

Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. 2023. Target-guided composed image retrieval. In *ACM MM*, pages 915–923.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2023. Composed image retrieval using contrastive learning and task-oriented clip-based features. *ACM ToMM*, 20(3):1–24.

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. 2024. Data roaming and quality assessment for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2991–2999.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. 2024. Covr: Learning composed video retrieval from web video captions. In *AAAI*, volume 38, pages 5270–5279.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, and 1 others. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

This is the appendix of "TEMA: Anchor the Image, Follow the Text for Multi-Modification Composed Image Retrieval".

A Multi-Modification Datasets

To evaluate the validity of models in multi-modification scenarios, we constructed two datasets. We now describe each dataset in detail as follows,

- **M-FashionIQ** is based on the classic CIR dataset, the **FashionIQ** (Wu et al., 2021), whose content belongs entirely to the fashion domain. It consists of $77,684$ images which are divided into three categories: *Dresses*, *Shirts*, and *Tops&Tees*. Following the FashionIQ, we treat it as three independent datasets. Following previous CIR methods, we utilize $\sim$46K and $\sim$15K images for training and testing, respectively. Finally, there are 18K triplets for training and $\sim$6K triplets for testing.

- **M-CIRR** is based on the classic open-domain CIR dataset, the **CIRR** (Liu et al., 2021). It contains $21,552$ real images taken from the renowned language reasoning dataset NLVR^2, which is well-known for its natural language reasoning applications. Specifically, we utilize $28,225$ and $4,181$ triplets for training and testing, respectively. In addition, following CIRR, M-CIRR includes a specialized subset designed for fine discrimination. This subset focuses on negative images that exhibit a high degree of visual similarity and is utilized to assess the model's performance in distinguishing false-negative images.

A.1 Dataset Statistics

For evaluation consistency, we note that existing CIR works report validation set results for FashionIQ, and CIRR's test set ground truth is not publicly available. Therefore, we expand modification texts to MMT in both training and validation sets of FashionIQ and CIRR, combining them with the original reference and target images to create new triplet collections, which serve as the training and test sets for M-FashionIQ and M-CIRR, respectively. Specifically, we process a total of $24,016$ queries in the FashionIQ dataset and $32,406$ queries in the CIRR dataset. As illustrated in Table 4, the minimum, maximum, and average lengths of the modification texts in our constructed M-FashionIQ and M-CIRR datasets significantly increase compared to the original datasets. This provides more room for modification texts. For example, M-FashionIQ and M-CIRR contain all three attributes of CR, ME, and FG, while the original FashionIQ does not have any of them. CIRR does not have CR and ME, and its dense labeling with FG is not widely used. These more comprehensive and detailed descriptions better capture users' nuanced composed retrieval needs in real-world scenarios.

Table 4: Comparison of modification text in lengths and attributes between the original CIR datasets and the expanded multi-modification datasets. The length is counted by tokens. In the attributes, **CR** represents Complex Relations (CR), **ME** denotes Multiple Entities (ME), and **FG** means Fine-Grained (FG).

Method	Length of Modification Text			Attribute		
	#Minimal	#Maximal	#Average	CR	ME	FG
FashionIQ	3.0	37.0	24.7	✗	✗	✗
M-FashionIQ	25.0	327.0	152.7	✓	✓	✓
CIRR	2.0	50.0	12.8	✗	✗	✓
M-CIRR	35.0	468.0	319.4	✓	✓	✓

Prompt used for M-FashionIQ

A target image <image> and a reference image <image>, answer is a modification long-text based on the following: {mod}.\
The modification description should focus on the different details of the clothes on different levels, SUCH AS shape, color, details of some parts. \
And be able to change the reference image to the target image. \
1. Focus on a specific aspect of the clothing (e.g., overall style, upper garment, lower garment, accessories, details).\
2. Directly compare the reference and target images.\
3. Describe changes in shape, color, texture, pattern, fit, and any other relevant aspects.\
4. Use clear and specific language.\
Consider the following aspects when formulating your points:\
- Overall style and silhouette changes\
- Specific element modifications (e.g., neckline, sleeves, hemline)\
- Color and shade variations\
- Texture and material changes\
- Pattern and print modifications\
- Accessory additions or removals\
- Fit and proportion adjustments\
- Fine details (e.g., buttons, zippers, stitching)\
Please build on the original modification: {mod}. \
For example:\
To transform the reference image into the target image, the following modifications can be made: \
1. Change the ... of the ... to be more ..., with a more ... design. \
2. the ... of the ... to be more ..., giving it a more ... look. \
3. Ensure that the ... is ..., with ... . \

Prompt used for M-CIRR

A target image <image> and a reference image <image>, analyze these two real-life images carefully. \
Provide a detailed set of instructions to transform the reference image into the target image. Focus on: \
Changes on main subjects or objects: Identify changes of primary elements. \
Changes on background and environment: Describe modifications needed in the setting or surroundings. \
Note changes of objects in number, actions, and so on. \
Note the changes in image layout. \
Changes in details and textures: Point out fine changes in patterns, surfaces, or small elements. \
Added or removed elements: List any objects or features that should be introduced or eliminated. \
Provide clear, step-by-step instructions for each modification. Be specific about the changes required, using descriptive language that could guide an image editing process. \
Ensure your instructions are comprehensive enough to theoretically transform the reference image into the target image. \
Please be loyal to original modification strictly: {mod}.

Figure 6: Prompts used in the process of MMT generation, for both M-FashionIQ and M-CIRR datasets.

A.2 Metrics

In terms of model evaluation, we train the model via the training sets of M-FashionIQ and M-CIRR, and then evaluate the model on the validation sets. Finally, we utilize the same evaluation metric, *i.e.*, recall at rank k (R@k), which is conventionally used in CIR tasks (Wu et al., 2021; Liu et al., 2021; Wen et al., 2023; Baldrati et al., 2022b, 2023).

B More Details for Dataset Construction

We supplemented more details about the dataset construction procedure, including Quality Chcek and Content Filter in Sec 3.

B.1 Quality Check

During the Quality Check process in Sec 3, we examine and revise the texts from four perspectives: **1) Consistency.** The modification text should describe plausible changes to objects or attributes within a given visual scene. Content that refers to irrelevant aspects, such as low-level image parameters (e.g., exposure, white balance), is considered inconsistent and should be removed. Annotators are responsible for ensuring semantic consistency within each text. **2) Accuracy.** The modification should accurately reflect the intended change without introducing speculative or hallucinated content. Annotators are required to verify that the described objects, attributes, or actions are plausible and grounded in real-world context, avoiding exaggeration or factual errors. **3) Diversity.** To enhance the expressiveness and coverage of the dataset and better accommodate diverse user intents, the modification texts should exhibit linguistic and conceptual diversity. Annotators are encouraged to avoid repetitive sentence structures and instead adopt varied phrasings and perspectives to enrich the overall corpus. **4) Quality.** As the initial MMTs are generated by an MLLM, they may suffer from issues such as unnatural phrasing or incoherent logic. Annotators are expected to refine and polish the texts where necessary, while preserving the original intent. A high-quality MMT should be fluent, precise, and reliable for both training and evaluation purposes.

B.2 Content Filter

Intuitively, even a well-formed modification text would be inappropriate if it fails to reflect a change relative to the reference image, e.g., describing

attributes of the target image in isolation. Such cases deviate from the core principle of multi-modification annotations, where the modification should be conditioned on the reference image. To address this issue, we introduce a *Content Filter* stage following the manual refinement. Specifically, we feed the human-corrected MMTs along with the target image into a Large Language Model (LLM) to detect statements that directly describe the target image content without referencing the reference image. These cases indicate a breakdown in the referential grounding and render the reference image ineffective under the multi-modification formulation. We remove such statements from the MMTs to ensure that both the reference image and the MMT collaboratively contribute to the retrieval query.

C Training Strategy

C.1 Loss Functions

Summary-guided Distillation. Furthermore, to ensure that the text tokens generated by the EM module contain all the information about the to-be-modified entities, we employ a summary-guided distillation strategy that aligns the EM module's output with entities parsed by PA. Specifically, for the textual entity feature $\hat{\mathbf{a}}_q$ in Eqn (3), we employ a simple cosine loss to close the semantics between the summary feature $\mathbf{E}_s$ in Eqn (2) and it, formulated as follows,

$$\mathcal{L}_{summ} = 1 - \cos\left(\mathbf{E}_s, \bar{\mathbf{a}}_q\right), \tag{5}$$

where $\bar{\mathbf{a}}_q$ indicates the average-pooled $\hat{\mathbf{a}}_q$.

Orthogonal Regularization. Considering if the N channels of the entity feature accurately represent the aggregation of different to-be-modified entities, they should be semantically independent and orthogonal. Inspired by TG-CIR (Wen et al., 2023), we design an Orthogonal Regularization to minimize the potential semantic overlap between channels, ensuring semantic independence.

$$\mathcal{L}_{ortho} = \left\|\hat{\mathbf{a}}_q^\top \hat{\mathbf{a}}_q - \mathbf{I}\right\|_F^2 + \left\|\hat{\mathbf{b}}_q^\top \hat{\mathbf{b}}_q - \mathbf{I}\right\|_F^2, \tag{6}$$

where $\mathbf{I} \in \mathbb{R}^{N\times N}$ and $\|\cdot\|_F^2$ is the Frobenius norm of matrix.

Batch-based Classification Loss. We apply the universal batch-based classification loss (Chen et al., 2020), which serves as a variant of cross-entropy, to align $\mathbf{E}_c$ with the target image feature $\mathbf{E}_t$, formulated as follows,

$$\mathcal{L}_{bbc} = \frac{1}{B}\sum_{i=1}^{B} -\log\left\{\frac{\exp\left\{\mathrm{s}\left(\bar{\mathbf{E}}_{ci}, \bar{\mathbf{E}}_{ti}\right)/\tau\right\}}{\sum_{j=1}^{B}\exp\left\{\mathrm{s}\left(\bar{\mathbf{E}}_{ci}, \bar{\mathbf{E}}_{tj}\right)/\tau\right\}}\right\}, \tag{7}$$

where as $\bar{\mathbf{E}}_{ci}, \bar{\mathbf{E}}_{ti}$ indicate the average pooled $\mathbf{E}_c, \mathbf{E}_t$ of the i-th triplet, respectively. τ is the temperature coefficient. B is the batch size.

C.2 Train and Inference

During training, we employ both the MMT Parsing Assistant (PA) and the MMT-oriented Entity Mapping (EM), and the final optimization function is formulated as,

$$\mathbf{\Theta}^* = \arg\min_{\mathbf{\Theta}} \left(\mathcal{L}_{bbc} + \kappa\mathcal{L}_{summ} + \mu\mathcal{L}_{ortho}\right), \tag{8}$$

where $\mathbf{\Theta}^*$ is the to-be-optimized parameter for TEMA and κ, μ are the trade-off hyper-parameters.

During the inference phase, the MMT Parsing Assistant (PA) module is forbidden, while the MMT-oriented Entity Mapping (EM) module has learned how to understand the MMT from the PA module during training. We fully compared the efficiency of our proposed TEMA with the SOTA of the CIR task.

D More Quantitative Results

D.1 Ablation Study

To evaluate TEMA's generalization capability and the performance of widely accessible LLMs in multi-modification scenarios, we conducted additional ablation studies on the LLMs integrated into TEMA. As summarized in Table 6, we replaced the MLLM with other easily obtainable models, such as Qwen, LLaMA, and others, to generate MMT summaries. The results show that switching between different LLMs has minimal impact on TEMA's overall performance.

This finding highlights TEMA's strong adaptability to various LLMs and demonstrates that it does not rely solely on proprietary models. Notably, TEMA maintains excellent training outcomes even when leveraging open-source models, such as LLaMA 3 or Qwen2-VL series, to replicate PA summarization and consistency checks. This reinforces the practicality and flexibility of TEMA in utilizing non-proprietary, openly available LLMs.

D.2 Computation Cost Analysis

We determine the computational costs of our proposed TEMA compared to the sub-optimal model

Table 5: Full validation results on traditional CIR benchmarks, including FashionIQ and CIRR.

Method	Year	Text Encoder	FashionIQ								CIRR					Avg
			Dresses		Shirts		Tops&Tees		Avg		R@k			R_{subset}@k		
			R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	k=1	k=5	k=10	k=1	k=2	
CASE (Levy et al., 2024)	2024	BLIP	47.44	69.36	48.48	70.23	50.18	72.24	48.79	70.68	48.00	79.11	87.25	75.88	90.58	77.50
Candidate (Liu et al., 2024b)	2024	BLIP	48.14	71.34	50.15	71.25	55.23	76.80	51.17	73.13	**50.55**	81.75	**89.78**	**80.04**	**91.90**	**80.90**
CoVR-BLIP (Ventura et al., 2024)	2024	BLIP	44.55	69.03	48.43	67.42	52.60	74.31	48.53	70.25	49.69	78.60	86.77	75.01	88.12	76.81
TEMA (Ours)		BLIP	**49.66**	**71.98**	**52.90**	**73.55**	**56.49**	**77.07**	**53.02**	**74.20**	49.15	**82.18**	88.81	78.17	90.32	80.18

Table 6: Ablation study on different LLMs in TEMA.

LLMs	M-FashionIQ		M-CIRR
	R@10	R@50	Avg
gpt-4o-mini	49.67	70.93	73.68
Qwen2-VL	**51.12**	72.59	**75.83**
Llama-2	48.33	70.24	72.73
Llama-3	50.57	72.63	75.31
Claude3.5-sonnet	50.98	**73.02**	75.59

Candidate (Liu et al., 2024b). Specifically, we choose FLOPs, train time, test time, and GPU memory, as shown in Table 7. All experiments were performed on a single NVIDIA A40 GPU and the batch size is set to 64. The FLOPs represent the number of floating-point operations. The train time describes the time it takes for the model to optimize to the optimum, while the test time is the time it takes for the inference on one sample. It is worth noting that our proposed TEMA model is superior on all indicators, demonstrating its overall effectiveness.

Table 7: Comparison of TEMA and the sub-optimal model Candidate on computation cost. The better results are in bold.

Method	Backbone	FLOPs	Train	Test	Memory
Candidate	BLIP	5.79G	16h	16.7ms/sample	47.3G
TEMA(Ours)	BLIP	**3.68G**	**2.83h**	**7.9ms/sample**	**43.9G**

D.3 Evaluation on Traditional CIR Benchmarks

In Table 5, we supplemented TEMA's performance on the original CIR datasets (FashionIQ and CIRR). The results show that TEMA is superior to existing SOTA on most metrics, sufficiently demonstrating TEMA's extensibility.

E Algorithm of TEMA's Training and Inference Process

To complement the method section in the full paper and more clearly illustrate the TEMA processing flow, we provide the complete TEMA training and inference processes in the form of pseudocode, which are presented in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 TEMA Training

Require: Triplets $\mathcal{T} = \{(x_r, t_m, x_t)\}_{n=1}^{N}$; frozen BLIP encoder $\Phi_{\mathbb{I}}, \Phi_{\mathbb{T}}$; η; batch size B; hyper-parameters κ, μ; learnable query number N
Ensure: Trained parameters Θ^*
Initial parameters Θ
for epoch $= 1$ to E **do**
  for each mini-batch $\{(x_r^i, t_m^i, x_t^i)\}_{i=1}^{B}$ **do**
    MMT Parsing Assistant (Sec 4.2)$\rightarrow$
    for $i = 1$ to B **do**
      Generating the summary t_s^i
      Check t_s^i using Consistency Detector
    end for
    $E_s^i = \Phi_{\mathbb{T}}(t_s^i)$
    MMT-oriented Entity Mapping (Sec 4.3)$\rightarrow$
    Feature Extraction (Sec 4.3)$\rightarrow$:
    $E_{r,g}^i, E_{r,l}^i = \Phi_{\mathbb{I}}(x_r^i)$;
    $E_{m,g}^i, E_{m,l}^i = \Phi_{\mathbb{T}}(t_m^i)$;
    $E_{t,g}^i, E_{t,l}^i = \Phi_{\mathbb{I}}(x_t^i)$
    Textual Entity Mapping (Sec 4.3)$\rightarrow$:
    $\hat{\mathbf{a}}_q = \text{Transformer}\left(\left[\mathbf{E}_s, \mathbf{E}_m^l, \mathbf{a}_q\right]\right)$
    Visual Entity Mapping (Sec 4.3)$\rightarrow$:
    $\hat{\mathbf{b}}_q = \text{Transformer}\left(\left[\mathbf{E}_r^g, \mathbf{E}_r^l, \mathbf{b}_q\right]\right)$
    Multimodal Query Composition (Sec 4.3)$\rightarrow$
    $\hat{\mathbf{E}}_m = [\mathbf{E}_m^g, \hat{\mathbf{a}}_q], \hat{\mathbf{E}}_r = [\mathbf{E}_r^g, \hat{\mathbf{b}}_q]$
    Then we get composed feature $\mathbf{E}_c$:
    $\mathbf{E}_c =$ **Combiner**$(\hat{\mathbf{E}}_m, \hat{\mathbf{E}}_r)$
    Summary-guided Distillation (Eqn 2)$\rightarrow$
    $\mathcal{L}_{summ} = 1 - \cos\left(\mathbf{E}_s, \bar{\mathbf{a}}_q\right),$
    Orthogonal Regularization (Eqn 26)$\rightarrow$
    $\mathcal{L}_{ortho} = \left\|\hat{\mathbf{a}}_q^\top \hat{\mathbf{a}}_q - \mathbf{I}\right\|_F^2 + \left\|\hat{\mathbf{b}}_q^\top \hat{\mathbf{b}}_q - \mathbf{I}\right\|_F^2,$
    Batch-based Classification Loss (Eqn 7)$\rightarrow$
    $\mathcal{L}_{bbc} = \frac{1}{B}\sum_{i=1}^{B} -\log\{\frac{\exp\{s(\bar{\mathbf{E}}_{ci}, \bar{\mathbf{E}}_{ti})/\tau\}}{\sum_{j=1}^{B}\exp\{s(\bar{\mathbf{E}}_{ci}, \bar{\mathbf{E}}_{tj})/\tau\}}\}$
    Overall Object (Eqn 8)$\rightarrow$
    $\mathcal{L} = \mathcal{L}_{\text{bbc}} + \mu\,\mathcal{L}_{\text{ortho}}$
    $\Theta \leftarrow \text{OptimizerUpdate}(\Theta, \nabla_\Theta \mathcal{L})$
  end for
end for
return Θ^*

F More Analysis on Prompts

In this section, we provided a more detailed analysis of prompts. In Appendix F.1, we present the detailed prompts used for MMT generation in the M-FashionIQ and M-CIRR datasets. In Appendix F.2, we analyzed the effect of different prompts on MMT generation.

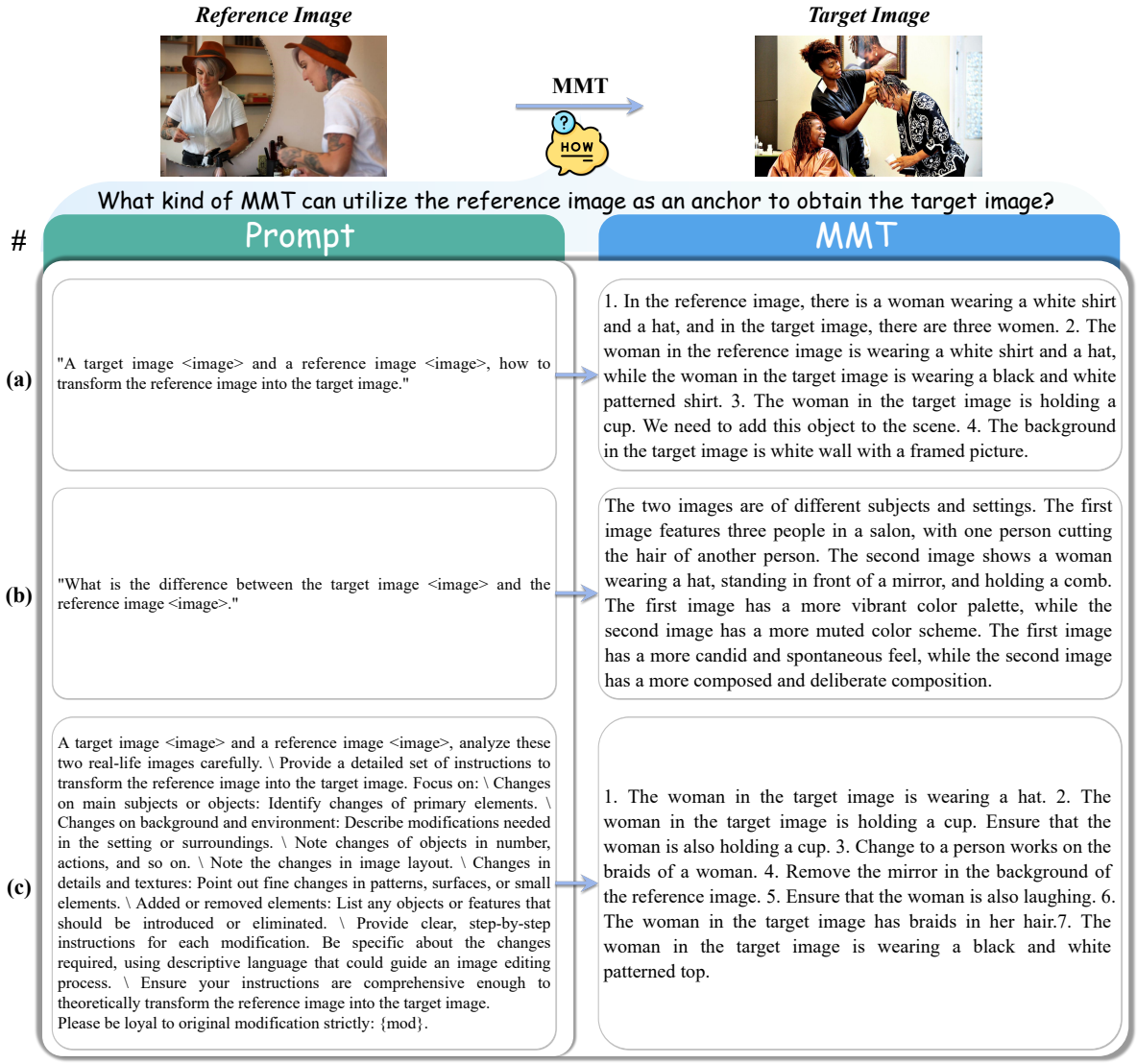

Figure 7: Generated MMTs using various prompts for BLIP-3.

Algorithm 2 TEMA Inference

Require: Queries (x_r, t_m); candidate images $\{x\}$; frozen BLIP encoder $\Phi_{\mathbb{I}}, \Phi_{\mathbb{T}}$;
Ensure: Ranked retrieval results

Feature Extraction:
$E_{r,g}, E_{r,l} = \Phi_{\mathbb{I}}(x_r),\ E_{m,g}, E_{m,l} = \Phi_{\mathbb{T}}(t_m)$
Textual & Visual Entity Mapping:
$\hat{\mathbf{a}}_q = \text{Transformer}\left([\mathbf{E}_s, \mathbf{E}_m^l, \mathbf{a}_q]\right)$
$\hat{\mathbf{b}}_q = \text{Transformer}\left([\mathbf{E}_r^g, \mathbf{E}_r^l, \mathbf{b}_q]\right)$
Composed Feature:
$\mathbf{E}_c = \textbf{Combiner}(\hat{\mathbf{E}}_m, \hat{\mathbf{E}}_r)$
For each candidate image x:
$\mathbf{E}_t(x) = \Phi_{\mathbb{I}}(x)$
Obtain the similarity:
$s(x) = \text{sim}(\mathbf{E}_c, \mathbf{E}_t(x))$
Rank by $s(x)$

F.1 Detailed Prompts for MMT Generation

In Figure 6, we showed the prompts used for generating the MMT of both M-FashionIQ and M-CIRR datasets, utilizing BLIP-3 (Xue et al., 2024).

For M-FashionIQ, we employed specifically designed prompts for BLIP-3. Based on the characteristics of the original FashionIQ dataset (which primarily focuses on various garment details such as the presence/absence of straps and sleeve lengths), these prompts were crafted to direct BLIP-3's output toward attention on different components of garments, holistic perception of the clothing items, and comparative analysis between reference and target images. This process enabled us to generate fine-grained MMT that captures detailed modification descriptions.

For M-CIRR, the details are more complex because the original CIRR dataset is open-domain, which generally involves several different objects as well as background elements. This also confirms the necessity of the MMT, which is only detailed

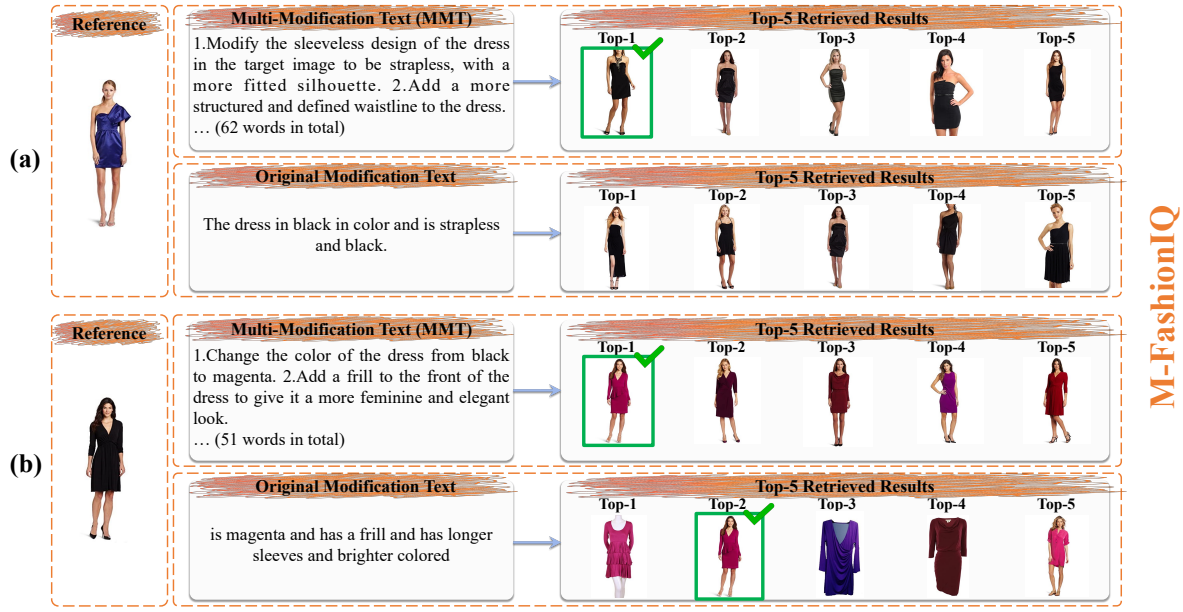

Figure 8: The mitigation on the false-negative samples when using MMT. We showed the top-5 retrieved results on both multi-modification datasets and original CIR datasets. The target images are framed in green.

enough to clearly describe the real modifications based on the reference and target images. Specifically, we require that the output of BLIP-3 focuses on changes in main subjects or objects, backgrounds and environments, details and textures, and so on.

F.2 Analysis of Different Prompts

Additionally, we require that the output of BLIP-3 be faithful to the original short modification text and that the MMT expands upon it, including details not noted in the original modification text. For example, the original modification text may refer to one object in the whole scene, however, the target image actually changes more than one object based on the reference image, allowing the MMT to describe the complete and detailed modification, avoiding the incompleteness of the original modification text.

For a more detailed analysis, we report different MMTs using various designed prompts for BLIP-3, as shown in Figure 7. In this figure, for prompts (a) and (c), both of them well captured the detailed modification from the reference image to the target image and elaborated on different perspectives. However, the MMT based on prompt (c) is more precise and comprehensive, focusing on one-to-many mappings and multiple constraints. Thus, we choose prompt (c) for our pipeline.

G More Qualitive Results

G.1 Mitigation on False-negative Samples

Ground-truth labeling in the CIR datasets are often insufficient due to the presence of numerous visually similar images and the limited descriptive capability of short modification text. For a given multimodal query, there are candidate images that differ subtly from the ground-truth yet still meet the query, these images are all labeled as negative samples, which we refer to as false-negative samples.

To validate the advantages of our proposed M-FashionIQ and M-CIRR in reducing false-negative samples, we selected a straightforward baseline, BLIP4CIR (Liu et al., 2024a), which incorporates multimodal query features to derive compositional features for retrieving target images. We conducted experiments in the MMT scenario for multi-modification datasets, and the original modification text scenario for original CIR datasets. As illustrated in Figure 8 and Figure 9, we present the top-5 retrieval results for both scenarios, with the target images highlighted in green boxes.

In the fashion domain dataset M-FashionIQ, as illustrated by the CIR example (bottom) in Figure 8(a), the top-5 results retrieved by the model satisfy the multimodal query, however, they are all classified as negative samples. This occurs because the short modification text in CIR fails to account for the inclusion of “a necklace” in the

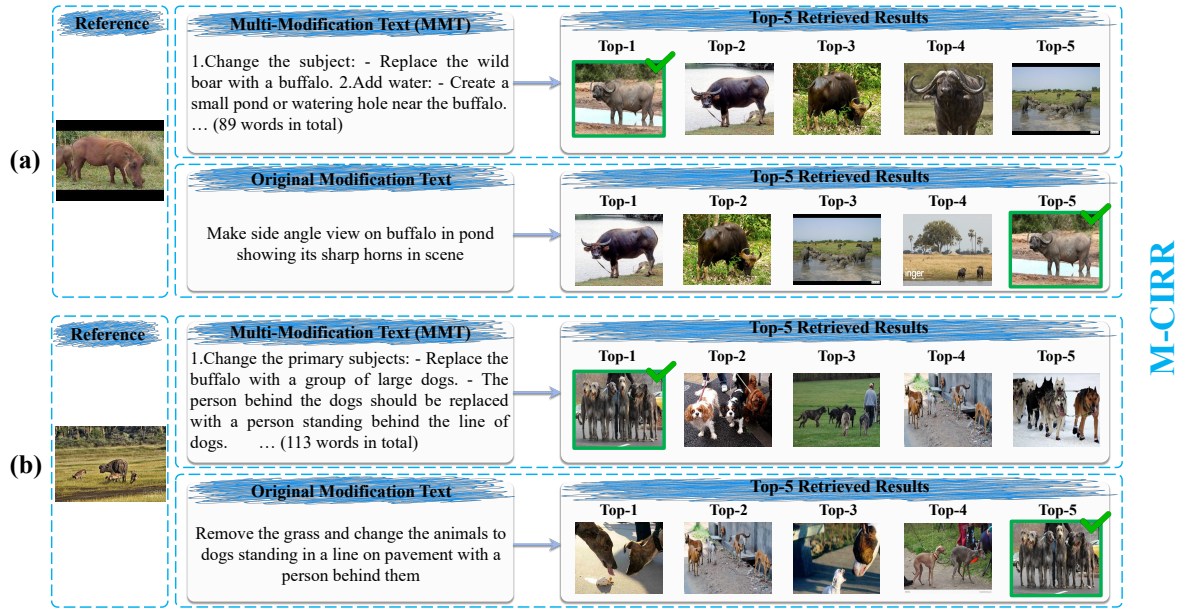

Figure 9: The mitigation on the false-negative samples when using MMT. We showed the top-5 retrieved results on both multi-modification datasets and original CIR datasets. The target images are framed in green.

target image. Conversely, the multi-modification annotation example (top) in Figure 8(a) highlights this detail modification after employing MMT relabeling, resulting in the target image becoming the only positive sample and significantly reducing the number of false-negative samples. A similar situation is observed in Figure 8(b), where the ranking of the target image in the retrieval results improves from second to first position following relabeling by MMT.

For open domain dataset M-CIRR, as the case shown in Figure 9(a), the original short modification text only required for "side angle view on buffalo", "in pond", and "sharp horns". However, the ground-truth (retrieved successfully using the MMT) included more requirements, such as "standing in the water", "dirt path", and "trees in the background". In the multi-modification scenario, the MMT encapsulated these details that are not present in the original short modification text, and therefore correctly retrieved the target image, weakening the impact caused by the false negative issue. Similarly, in Figure 9(b), the ranking of the target image in the retrieval results improves from fifth to first position after relabeling by MMT. These results demonstrate that MMT provides more detailed descriptions, causing the original false-negative samples to no longer satisfy the new multimodal query. Thus, these samples are converted into true-negative samples, alleviating the issue of false negatives in the multi-modification datasets and reducing their impact on model training.

G.2 Attention Visualization for Summaries

The summary generated by the PA module serves as a simplified representation in MMT, encompassing all the to-be-modified entities in the reference image. To evaluate the quantity of the summaries, we employed Grad-CAM to visualize its attention in relation to the reference image.

For the fashion domain dataset M-FashionIQ, as the case illustrated in Figure 10(a), the to-be-modified entities in MMT include "neckline", "sleeve length", "shoes", "skirt", and "belt". The summary concisely consolidates these entities into a single sentence while omitting specific modification details for each entity. The attention visualization demonstrates the summary's focus on different regions of the reference image. We observed that all to-be-modified entities are well-attended, validating the correctness and accuracy of the summary content. Similarly, the summary in Figure 10(b) highlights all the to-be-modified entities, including "sleeve pattern", "skirt", and "neckline".

For the open domain dataset M-CIRR, taking the case in Figure 11(a) as an example, the MMT includes the to-be-modified entities "sheep" and "person", with many modification details. In contrast, the summary succinctly and accurately expresses the to-be-modified entities while omitting detailed information. Through the attention visualization, we observe that both the "sheep" distributed across

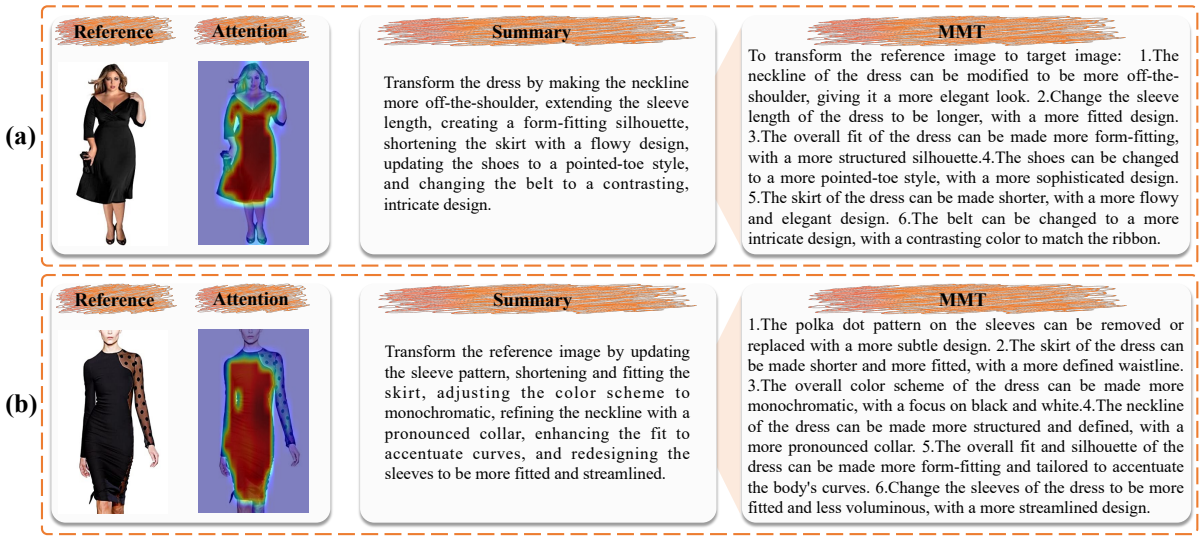

Figure 10: Attention visualization results for the reference image on M-FashionIQ by the PA-generated summary.

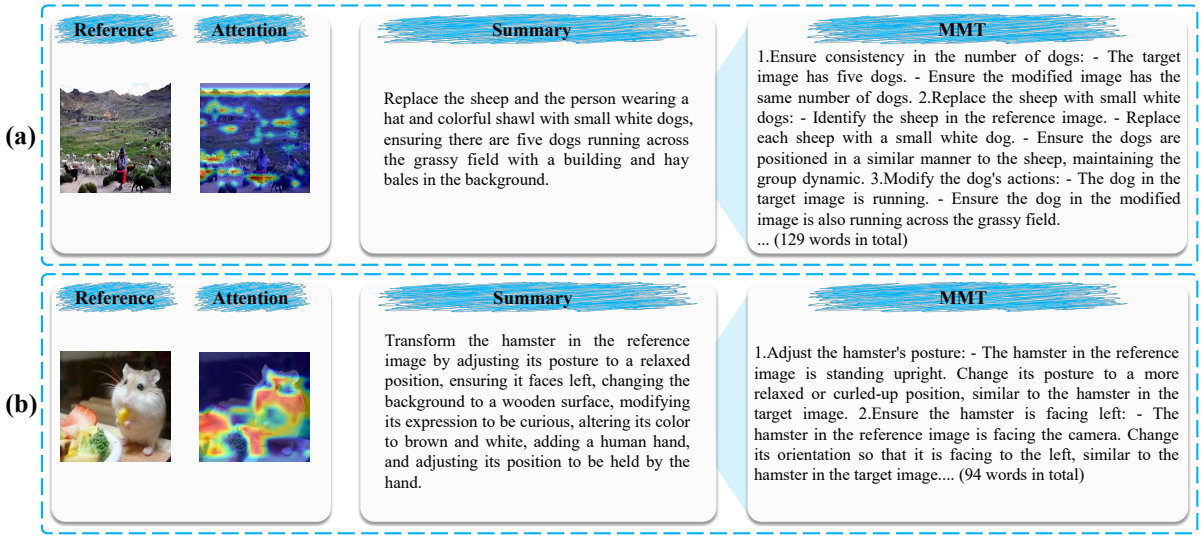

Figure 11: Attention visualization results for the reference image on M-CIRR by the PA-generated summary.

different regions and the single "person" are well-attended to. Similarly, the summary in Figure 11(b) addresses the full range of to-be-modified entities in MMT, including the "hamster", "background", and "expression". This validated the effectiveness of our generated summary in encompassing all the to-be-modified entities in the MMT, thereby enhancing the model's focus on these entities.

G.3 Case Study

To intuitively validate the performance of our proposed TEMA on multi-modification datasets, we present several examples demonstrating TEMA's retrieval results, along with the comparison to sub-optimal model candidates, as shown in Figure 12. Specifically, Figure 12(a) and (b) demonstrated the results on M-FashionIQ, while Figure 12(c) and (d) present the results on M-CIRR.

For cases in Figure 12(a), (c), and (d), TEMA successfully retrieved the target images at top-1, whereas Candidate failed to rank the target images in the first position. For the case in Figure 12(d), the Candidate model even failed to retrieve the target image within the top-5 results. These examples demonstrate the superior performance of our proposed TEMA and its effectiveness. Notably, in Figure 12(b), both TEMA and Candidate failed to retrieve the target image in the first position, which may be attributed to insufficient annotation in the original CIR dataset. While our proposed TEMA framework mitigated the false-negative issues inherent in CIR, these problems persisted in a small number of triplets, resulting in model failure in these cases.

Figure 12: Qualitative examples of our proposed TEMA compared to the sub-optimal model Candidate.