MAI: A MULTI-TURN AGGREGATION-ITERATION MODEL FOR COMPOSED IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-Turn Composed Image Retrieval (MTCIR) addresses a real-world scenario where users iteratively refine retrieval results by providing additional information until a target meeting all their requirements is found. Existing methods primarily achieve MTCIR through a "multiple single-turn" paradigm, wherein methods incorrectly converge on shortcuts that only utilize the most recent turn's image, ignoring attributes from historical turns. Consequently, retrieval failures occur when modification requests involve historical information. We argue that explicitly incorporating historical information into the modified text is crucial to addressing this issue. To this end, we build a new retrospective-based MTCIR dataset, **FashionMT**, wherein modification demands are highly associated with historical turns. We also propose a Multi-turn Aggregation-Iteration (**MAI**) model, emphasizing efficient aggregation of multimodal semantics and optimization of information propagation in multi-turn retrieval. Specifically, we propose a new Two-stage Semantic Aggregation (TSA) paradigm coupled with a Cyclic Combination Loss (CCL), achieving improved semantic consistency and modality alignment by progressively interacting the reference image with its caption and the modified text. In addition, we design a Multi-turn Iterative Optimization (MIO) mechanism that dynamically selects representative tokens and reduces redundancy during multi-turn iterations. Extensive experiments demonstrate that the proposed MAI model achieves substantial improvements over state-of-the-art methods.

1 INTRODUCTION

Image retrieval remains a longstanding task in computer vision Sain et al. (2023); Levy et al. (2024a), gaining continuous attention in practical applications such as e-commerce in recent years Jin et al. (2023); Park et al. (2019). However, relying solely on images may fall short of practical needs, as users often modify these images better to match their requirements Chen et al. (2020); Guo et al. (2018). In response, Composed Image Retrieval (CIR) has been introduced to locate target images by combining reference images and modified text Wen et al. (2023); Shoib et al. (2023). Due to the interactive nature of retrieval scenario Xu & Sundar (2014); Adhikari et al. (2018), multi-turn systems can leverage more user feedback, fulfilling user needs better than single-turn systems Agnolucci et al. (2023); Chen et al. (2023a). Therefore, Multi-turn Composed Image Retrieval (MTCIR), which aims to retrieve the most suitable target image by allowing users to iteratively select images and provide modification feedback, as illustrated in Figure 1, has garnered increasing attention in recent years Guo et al. (2018); Yuan & Lam (2021); Liu et al. (2024b).

Due to the lack of dedicated datasets for the MTCIR task Pal et al. (2023), existing methods typically construct multi-turn datasets by concatenating single-turn CIR datasets Wu et al. (2021); Guo et al. (2018), using the target image from the historical turn as the reference image for the next turn. However, datasets constructed in this manner exhibit the following limitations: **(i) Lack of historical context.** Modified text in existing MTCIR datasets lacks image information from historical turns, resembling "multiple single-turn" retrievals and deviating from real-world scenarios. **(ii) Small data scale.** Existing single-turn datasets face challenges due to their limited scale Saito et al. (2023); Feng et al. (2024); Zhao et al. (2024b). Moreover, this concatenation method further diminishes the size of multi-turn datasets, lagging behind current trends Chen et al. (2023b); Baldrati et al. (2023).

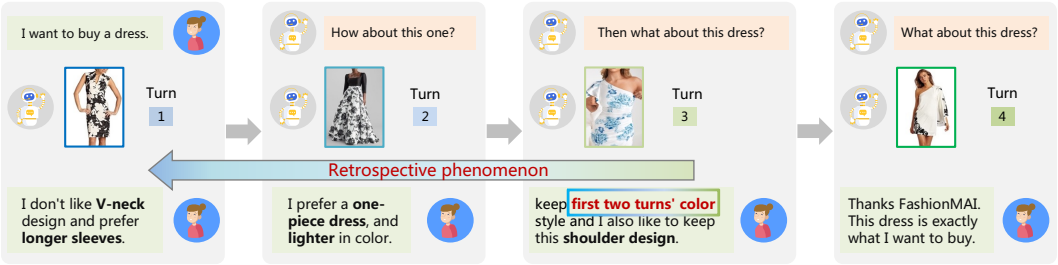

Figure 1: The definition of Multi-Turn Composed Image Retrieval (**MTCIR**). The retrospective phenomenon is common in the MTCIR task, wherein a user's new turn modification request often involves the attributes of images from historical turns.

The deficiencies outlined in existing MTCIR datasets have hindered the development of methods in this domain. Existing methods typically employ a "multiple single-turn" paradigm for multi-turn retrieval. However, this paradigm causes methods to incorrectly converge on shortcuts that only utilize the most recent turn's image, neglecting attributes from previous turns. Consequently, retrieval failures arise when modification requests involve attributes or modifications from previous images. Additionally, existing methods lack designs to leverage inherent multimodal information in images Chen et al. (2023b); Li et al. (2024a), and to store multi-turn information effectively.

To address these issues and align with existing MTCIR datasets, we construct a new dataset, **FashionMT**, tailored for e-commerce scenarios characterized by typical multi-turn interactions. FashionMT has the following characteristics: **(i) Retrospective-based.** It simulates real-world MTCIR scenarios, where the modified text in each new turn may involve information from historical reference images, such as preserving certain attributes. This necessitates retrieval algorithms to utilize multi-turn historical information retrospectively. **(ii) Massive and diverse.** FashionMT contains 14 times more fashion images and 30 times more categories than MT FashionIQ Yuan & Lam (2021). Our Modification Generation Framework generates multi-turn transactions nearly 27 times larger than MT FashionIQ, offering rich multimodal data, including images, text, attributes, etc.

We further propose a multi-turn key information-aware approach, the Multi-turn Aggregation-Iteration (**MAI**) model, which focuses on two challenges in MTCIR: **(i) multimodal semantics aggregation** and **(ii) multi-turn information optimization.** Specifically, MAI introduces a new Two-stage Semantic Aggregation (TSA) paradigm coupled with a Cyclic Combination Loss (CCL). TSA introduces captions as a transition, progressively aggregating the image with its caption and then with the modified text. The CCL's cyclic structure further enhances semantic consistency and modality alignment. We also provide theoretical insights into the rationale behind introducing captions for two-stage fusion. Furthermore, we design a parameter-free Multi-turn Iterative Optimization (MIO) mechanism that dynamically selects representative tokens with high semantic diversity, effectively reducing the storage space for historical information tokens.

Our contributions are summarized as follows:

- We build the first dataset specifically designed for multi-turn composed image retrieval, named FashionMT, characterized by its retrospective-based nature and massive diversity.
- We propose the Multi-turn Aggregation-Iteration (MAI) model, focusing on efficient aggregation and iterative optimization of multimodal semantics in multi-turn composed image retrieval.
- We provide theoretical insights that our modality fusion approach effectively bridges the modality and semantic gaps, which informs the design of our loss function.
- Extensive experiments demonstrate that our proposed MAI model obtains substantial improvements and achieves state-of-the-art performance.

2 Related Work

Single-turn Composed Image Retrieval. In existing works on composed image retrieval, the focus has mainly been on single-turn retrieval Wen et al. (2023); Chen et al. (2024c), which can be

categorized based on the amount of training data into fully trained on all data Goenka et al. (2022); Levy et al. (2024b) and zero-shot Karthik et al. (2024); Gu et al. (2023); Chen & Lai (2023) or few-shot Wu et al. (2023) settings. Currently, composed image retrieval methods can be broadly categorized into two paradigms Bai et al. (2024): late fusion Chen et al. (2024b); Zhang et al. (2024) or pseudo-word embedding methods Baldrati et al. (2023); Liu et al. (2024b); Suo et al. (2024). In the first paradigm type, Baldrati et al. (2022) propose a simple yet effective fusion model, Combiner, to combine features extracted by the CLIP Radford et al. (2021) model. In the second paradigm type, Saito et al. (2023) propose an LLAVA-like Liu et al. (2024a) method to convert visual features into tokens for a text encoder. Bai et al. (2024) propose a method similar to the BLIP-2 Li et al. (2023) to learn sentence-level prompts, achieving state-of-the-art results. However, the above methods are limited to single-turn retrieval scenarios and are challenging to apply directly to the more user-demand-oriented MTCIR tasks.

MTCIR Methods. Several recent methods have emerged in the fusion of visual and textual inputs across multiple exchanges of information Zhu et al. (2024); Li et al. (2024b); Hu et al. (2024). Due to the inherent multi-turn nature of dialogues Zhang et al. (2022); Yu et al. (2019), a common application scenario is multi-turn dialogue systems Zolkepli et al. (2024); Zheng et al. (2022). In the prevalent retrieval tasks of the fashion domain, multi-turn retrieval has emerged as a more comprehensive approach compared to single-turn retrieval, offering enhanced user interaction and feedback to better cater to user needs Zhang et al. (2019); Agnolucci et al. (2023). There have been several groundbreaking studies in multi-turn composed image retrieval in recent years. Guo et al. (2018) propose modeling images and text using CNN networks, capturing sequential information with RNNs, and employing reinforcement learning for constraint. Yuan & Lam (2021) construct the first multi-turn composed retrieval dataset based on the single-turn retrieval dataset FashionIQ Wu et al. (2021). Pal et al. (2023) introduce a memory network to retain historical retrieval information and further develop a multi-turn retrieval dataset based on the single-turn dataset Shoes Guo et al. (2018). However, the above methods fail to leverage the multimodal content naturally present in fashion images, such as captions and titles. Additionally, these methods do not consider optimizing the storage overhead of multi-turn representations.

Fashion Datasets. In the past few years, a large number of datasets have been proposed for retrieval Corbiere et al. (2017); Ge et al. (2019); Rostamzadeh et al. (2018); Han et al. (2017); Zhan et al. (2021). Due to the inherent inclusion of a vast amount of data and extensive user interactions in the e-commerce domain, existing fashion datasets exhibit a large scale. The Product1M Zhan et al. (2021) contains 1,182,083 cosmetic samples. The M5Product Dong et al. (2022) encompasses 6,131,064 samples with 5 modalities. A massive amount of data also contributes to the model acquiring capabilities closer to practical usage Chen et al. (2023b). In the composed image retrieval task, FashionIQ Wu et al. (2021) and Shoes Guo et al. (2018) represent pioneering works, being more user-friendly compared to direct image retrieval. However, the MTCIR task still lacks a dedicated custom dataset. Constructing modifications by concatenating single-turn datasets fails to capture the historical context crucial for multi-turn scenarios. Our proposed FashionMT offers essential and timely data support to advance this task.

3 THE FASHIONMT DATASET

Table 1: Comparison with other MTCIR datasets. *MT* stands for Multi-turn. In a *transaction*, there are multiple *turns*. *Length* denotes the modified text's average length. *Categories* denotes the number of finest subcategories, while *Product type* lists the typical product categories.

Datasets	# Images	# Transactions	# Turns	# Categories	Length	Product type
MT FashionIQ Yuan & Lam	74,381	11,505	26,506	3	10.7	shirt, top-tee, dress
MT Shoes Pal et al.	15,661	4,097	11,346	10	5.2	boots, sneakers, clogs, etc.
FashionMT (ours)	1,067,688	247,911	743,733	95	24.3	shirt, top-tee, dress, shoes, coat, pants, bag, ornament, etc.

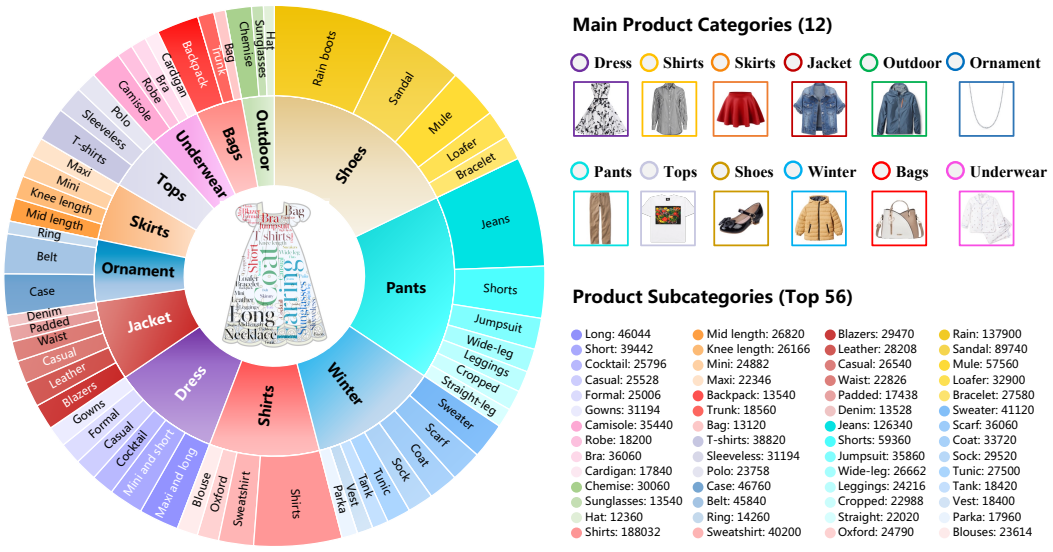

Figure 2: Top categories and distribution in our proposed FashionMT dataset. We have listed 12 main product categories and the top 56 product subcategories to provide a clearer presentation.

3.1 Data Collections and Construction

Our data primarily originates from two sources: (i) Gathering images and associated text from existing single-turn composed image retrieval datasets Wu et al. (2021); Guo et al. (2018); Han et al. (2017). (ii) Crawling images and related text from multiple e-commerce platforms. We clean the scraped images, including removing damaged, unclear, and non-product images.

Inspired by the manual annotation process of modified text Wu et al. (2021); Liu et al. (2021), we propose a Modification Generation Framework (MGF) to automate the construction of our dataset by capturing the distinctions between reference and target image pairs. The framework consists of the following steps: (i) Image Selection: Selecting $N+1$ images from a product subcategory for N turns in a transaction. (ii) Caption Generation: Generating captions for these images using an image captioning model. (iii) Base Modification Generation: Employing a large language model (LLM) to describe the differences between image captions from adjacent turns. (iv) Retrospective Modification Generation: Determining the specific turns requiring retrospective analysis and generating corresponding modified text based on the intersection of attributes between the most recent image and images from previous turns.

Specifically, we generate the captions using the prompt: "*Question: Describe the product. Answer:*". For generating base modified text, the prompt is: "*The reference depicts* {*REF*}*, and the target depicts* {*TAR*}*. Describe the modifications to transform the reference into the target*", where *REF* and *TAR* represent the captions of the reference and target images within a single turn, respectively.

To better align with retrospective needs in real-world scenarios, we have established two scenarios for generating retrospective-based modified text: **rollback** and **combination**. In the rollback setting, similar to base transaction generation, modifications are generated between a specified reference and the target by rolling back. An example under this setup is: "*Compared to the most recent turn, I still prefer the item from the second turn. Building on that, I like...*". In the combination setting, users combine attributes from multiple images in historical turns to formulate modification requests. An example under this setup would be: "*I like ... from the first turn, and ... from the second turn*". In this setup, the modified text consists of two parts: the initial segment encompasses common attributes earmarked for retention, prefaced by the prompt "*Keep the* {*Attr*} *in the* {*ID*} *turn*" where *Attr* represents common properties like color, logo, pattern, etc., and *ID* signifies the turns sharing commonalities with the target. Meanwhile, the subsequent segment delineates additional modification requisites, prefaced by the prompt "*the reference images depict REFs, the target depicts TAR. Describe the distinctiveness of the target:*".

3.2 Dataset Statistics

The data distribution of FashionMT is illustrated in Figure 2. Detailed information and a comparison with existing datasets, MT FashionIQ Yuan & Lam (2021) and MT Shoes Pal et al. (2023), are presented in Table 1. FashionMT significantly surpasses existing datasets in both scale and richness, featuring 14 times more images than MT FashionIQ and nearly 10 times more categories than MT Shoes. By leveraging the Modification Generation Framework, FashionMT enables the efficient construction of high-quality transactions, resulting in a dataset that is 27 times larger than MT FashionIQ. Additionally, FashionMT provides more detailed modified text, with an average length twice that of MT FashionIQ. As a dataset tailored specifically for MTCIR task, FashionMT offers more comprehensive and realistic data support. For more details on our proposed dataset, including setup explanations and quality control, please refer to Section 7.3.

4 Approach

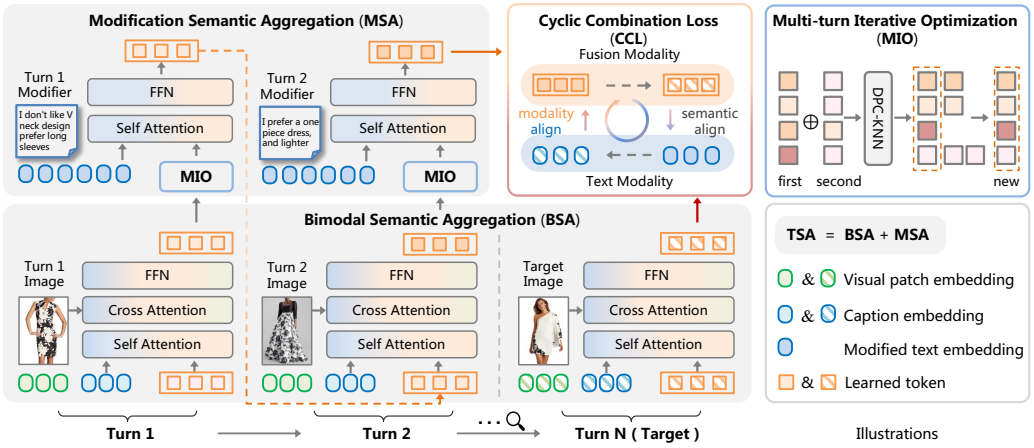

Figure 3: The architecture of the MAI model. For each turn, images, captions, and modified text are progressively aggregated through BSA and MSA, with MIO preserving core information across turns, and CCL constraining the training process. For simplicity, we illustrate two retrieval turns.

4.1 Problem Formulation

In our task setup, we provide the previous $N-1$ turns' multimodal data, which consists of predefined reference images with captions and modified text, aiming to retrieve the most suitable target image based on the final modified text. We represent the image patch embedding, caption embedding, and modified text embedding of the n-th turn as $v_n \in \mathcal{V}$, $c_n \in \mathcal{C}$, and $m_n \in \mathcal{M}$ respectively. Furthermore, we distinguish among the reference image embedding, reference caption embedding, target image embedding, and target caption embedding as v_n^r, c_n^r, v_n^t, and c_n^t respectively.

4.2 Multi-turn Aggregation-Iteration (MAI) Model

The architecture of MAI is depicted in Figure 3. We will introduce Bimodal Semantic Aggregation (BSA) and Modification Semantic Aggregation (MSA), which are part of the Two-stage Semantic Aggregation (TSA), along with the Multi-turn Iterative Optimization (MIO).

Bimodal Semantic Aggregation (BSA). In the n-th turn, we first conduct a lexical analysis on the modified text to determine if there is a rollback operation. We established a template for automatically generating modified text with Rollback instructions to facilitate benchmark construction. The template includes phrases such as: "Compared to this one I prefer the {}, and", "I would rather choose the {}, and", where {} denotes rollback turn descriptions, such as "Turn 2: White off-shoulder lace short sleeve." Rollback operations are executed when modified text match the template. If so, the reference image is designated as the image from the specified rollback turn. If not,

the default reference image for the n-th turn is adopted. We extract visual patch embeddings v_n of images using a frozen visual encoder. The effectiveness of the Q-Former architecture Li et al. (2023) in integrating vision-text embeddings has been validated in prior studies Bai et al. (2024); Hu et al. (2024). Hence, our BSA transfers this framework to adapt to the MTCIR task. Through learned tokens t_n, BSA initially learns the bimodal semantics of images and their corresponding captions before interacting with modified text. This strategy employs captions as a transition, enhancing modality relevance during interaction with modified text, as elaborated in Section 4.3. A fixed text encoder extracts caption embeddings c_n, interacting with learned tokens t_n in BSA's self-attention layers. In the cross-attention layers, they engage with visual patch embeddings v_n. After BSA, learned tokens aggregate multimodal semantics from reference images and captions, denoted as $t_n^{r,\mathrm{BSA}}$. As the target side lacks modified text, this embedding is directly used for training loss constraints and inference distance measurement.

Multi-turn Iterative Optimization (MIO). Despite learned tokens being more space-efficient than visual embeddings Li et al. (2023), storing these tokens for each turn still results in significant space consumption. Additionally, fashion images encompass various attributes such as color, style, size, etc Tian et al. (2023); Chen et al. (2023b); Han et al. (2023). Multi-turn retrieval often revolves around the same subcategory of products, resulting in similar attributes across the images involved in multiple turns. Therefore, we propose a parameter-free mechanism to optimize and retain the key attributes throughout multi-turn interactions.

Specifically, we concatenate $t_{n-1}^{r,\mathrm{MIO}}$ from the previous turn with $t_n^{r,\mathrm{BSA}}$ from the current turn to obtain $t_n^{r,\mathrm{MIO}}$. Our objective is to preserve key semantic tokens while discarding redundant ones from the learned tokens t_n^c. This process involves several steps. (i) Clustering. We apply an efficient density peaks clustering based on k-nearest neighbors (DPC-kNN) algorithm Du et al. (2016). The learned tokens t_n^c are clustered into k groups and the clustering operation is formulated as follows:

$$\mathrm{cluster}(t_n^c, k) = \arg\min_{C} \sum_{i=1}^{k} \sum_{v \in C_i} ||v - c_i||^2 \tag{1}$$

where C represents the clusters, C_i represents the i-th cluster, and c_i represents the centroid of the i-th cluster. (ii) Density Estimation. After clustering, the density of each cluster is estimated based on the distances between the tokens within the cluster and learned tokens with lower densities are filtered out to enhance clustering efficiency. The density estimation is calculated as follows:

$$\mathrm{density}(v) = \exp(-\frac{1}{k} \sum_{u \in \mathrm{Nei}(v)} ||v - u||^2) \tag{2}$$

where $\mathrm{Nei}(v)$ represents the neighboring tokens of v. (iii) Pruning. Tokens with low density are eliminated to ensure that only the most semantically significant tokens are retained. To achieve this, each token is assigned a score, computed as the product of its density and its distance to the nearest neighbor. The top k tokens with the highest scores are then selected as the optimized tokens.

$$\mathrm{score}(v) = \mathrm{density}(v) \times \mathrm{dist}(v) \tag{3}$$

where $\mathrm{dist}(v)$ represents the distance of token v to its nearest neighbor. The final tokens, denoted as $t_n^{r,\mathrm{MIO}}$, are obtained by selecting the tokens with the top-k scores. Through the optimization process described above, MIO effectively preserves learned tokens carrying discriminative semantics while discarding tokens with relatively less semantic importance, thereby saving computational resources.

Modification Semantic Aggregation (MSA). During the MSA stage, we engage the tokens $t_n^{r,\mathrm{MIO}}$, which encapsulate reference semantics, with the modified text embedding m_n. By employing a frozen text encoder to extract embeddings m_n, we concatenate them with learned tokens $t_n^{r,\mathrm{MIO}}$ before feeding them into the self-attention layer. Subsequently, we employ a linear and normalization layer on the learned tokens to map them, producing a reference-side embedding t_n^r. This embedding concurrently embodies multimodal semantics from the reference and modified text.

It is important to note that in the combination setting, due to the involvement of multiple historical images, BSA aggregates the bimodal embeddings by concatenating the learned tokens from previous turns with their corresponding image captions. Subsequently, these embeddings are semantically aggregated with the modified text in the MSA.

4.3 THEORETICAL INSIGHTS

In this section, we justify the rationale behind introducing captions for bimodal semantic aggregation and explain how our approach outperforms a naive solution. In the MTCIR task, we hypothesize that the transition from the initial image to the final target image occurs by gradually introducing specific attributes related to the modified text in each turn, denoted as $v_n^t - v_n^r \sim \mathcal{N}(m_n, \frac{1}{N}I)$. Ideally, the visual increments $v_n^t - v_n^r$ should correspond to the textual modifications m_n, i.e., $v_n^r + m_n = v_n^t$, which can be supervised using the following similarity loss:

$$\mathcal{L}_{\text{sim}} = \frac{1}{B}\sum_{i=1}^{B}\big(1 - \frac{|v_n^{ri}| + |m_n^i|}{2} \cdot |v_n^{ti}|\big) \tag{4}$$

where for simplicity, given an embedding x, $|x|$ stands for the normalized form $\frac{x}{||x||}$.

However, the effectiveness of the aforementioned supervision is constrained by: (i) the inherent modality gap between texts and images; (ii) the semantic disparity between the additional textual attributes and visual items. To mitigate these gaps, we propose leveraging image captions to (i) align with the modality of modified texts; (ii) match the semantics of visual items. We can adopt a naive method Huang et al. (2023), replacing the visual embeddings in Eq. 4 with the corresponding caption text embeddings for cross-modal constraints. This results in the following **naive** cross-modal loss by replacing images with their corresponding captions:

$$\mathcal{L}_{\text{naive}} = \mathcal{L}_{\text{sim}} + \frac{1}{B}\sum_{i=1}^{B}\left[1 - \frac{1}{2}\big(\frac{|v_n^{ri}| + |m_n^i|}{2} \cdot |c_n^{ti}| + \frac{|c_n^{ri}| + |m_n^i|}{2} \cdot |v_n^{ti}|\big)\right] \tag{5}$$

Although $\mathcal{L}_{\text{naive}}$ aligns visual increments with textual modifications, bridging both modality and semantic gaps in the meantime, the separate optimization in modality and semantic space may affect each other in the training process. To make further efforts, we propose that the reference and target images should undergo **pre**-fusion with caption text to achieve an intermediate state that is closer in modality and semantics to the modified text. The paradigm of this process is represented as follows:

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{sim}} + \frac{1}{B}\sum_{i=1}^{B}\big(1 - \frac{(|v_n^{ri}| + |m_n^i|) + (|c_n^{ri}| + |m_n^i|)}{4} \cdot \frac{|v_n^{ti}| + |c_n^{ti}|}{2}\big) \tag{6}$$

Furthermore, we give theoretical justifications for the effectiveness of the proposed pre-fusion loss:

Proposition 1: *Let $\mathcal{O}(GError(\mathcal{L}_{\text{pre}}))$ and $\mathcal{O}(GError(\mathcal{L}_{\text{naive}}))$ be the upper bound of generalization error of the above two losses. Then for any hypothesis $\mathcal{L}_{\text{pre}}, \mathcal{L}_{\text{naive}}$ in $\mathcal{H} : \mathcal{V} \times \mathcal{C} \times \mathcal{M} \to [0, 1]$ and $1 > \delta > 0$, it holds that:*

$$\mathcal{O}(\text{GError}(\mathcal{L}_{\text{pre}})) \leq \mathcal{O}(\text{GError}(\mathcal{L}_{\text{naive}})) \tag{7}$$

with probability at least $1 - \delta$, given that the visual and textual encoders are reliable for generating positively correlated embeddings in the n-th turn and clustering embeddings with the same modality.

4.4 OPTIMIZATION AND INFERENCE

Training. Given the guiding role of modified text in retrieval Chen et al. (2024a), we design the **Cyclic Combination Loss (CCL)** to align semantically similar fused modality with text modality, thereby preserving the critical semantics within the textual modality. Specifically, we employ a batch-based classification loss commonly used in CIR and MTCIR tasks Pal et al. (2023); Wen et al. (2024); Chen et al. (2024a), which is defined as:

$$\mathcal{L}_{\text{B}}(r_q, r_t) = \frac{1}{B}\sum_{i=1}^{B} -\log \frac{\exp \kappa(r_q^i, r_t^i)}{\sum_{j=1}^{B} \exp \kappa(r_q^i, r_t^j)} \tag{8}$$

where B represents the batch size, the kernel $\kappa()$ is the inner product resulting in cosine similarity. r_q denotes the reference-side representation, and r_t signifies the target-side representation.

Inspired by $\mathcal{L}_{\text{pre}}$'s paradigm in Section 4.3 and 7.1, our loss function incorporates three constraints on embeddings after Bimodal Semantic Aggregation pre-fusion, along with an additional constraint on the text modality. For the n-th turn, the cyclic constraints involve the following four sets of

embeddings: learned tokens t_n^r from MSA, encompassing semantics of the reference image, caption, and modified text; learned tokens t_n^{tg} from BSA, containing semantics of the target image and its caption; the modified text embedding m_n and the caption text feature of the target image c_n^{tg}.

In line with previous works in the MTCIR task, our overall Cyclic Combination Loss $\mathcal{L}_{\mathrm{CCL}}$ for N turns is composed of the cumulative losses from each turn:

$$\mathcal{L}_{\mathrm{CCL}} = \sum_{n=1}^{N} \mathcal{L}_{\mathrm{B}}(t_n^r, t_n^{tg}) + \mathcal{L}_{\mathrm{B}}(t_n^{tg}, m_n) + \mathcal{L}_{\mathrm{B}}(m_n, c_n^{tg}) + \mathcal{L}_{\mathrm{B}}(c_n^{tg}, t_n^r) \tag{9}$$

Inference. At the conclusive N-th turn, $t_{N-1}^{r,\mathrm{MIO}}$ encompasses key multimodal semantics from prior turns. Upon interacting with the modified text through MSA, we derive the reference-side embedding t_N^r. Meanwhile, on the gallery side, the bimodal embedding of the image and its caption, t_N^{tg}, is computed. Retrieval matching ensues by evaluating the similarity between t_N^r and t_N^{tg}.

5 Experiment

5.1 Experimental Setting

Implementation Details. We adopt BLIP-2 Li et al. (2023) with the Flan-t5-xxl language model Chung et al. (2024) for image captioning and Xwin-13B-V0.2 Ni et al. (2024) as the LLM. Optimization is performed using AdamW Loshchilov & Hutter (2019) with a batch size of 16, an initial learning rate of 1e-5, and cosine annealing. Training runs for 50 epochs, while inference uses a batch size of 2048. All model training and inference are conducted on 8 V100 GPUs. The number of learned tokens is fixed at 32, and 32 tokens are retained each turn through the MIO. Q-Former parameters are initialized with $\mathrm{blip2\_pretrain\_vitL}$, consistent with SPRC Bai et al. (2024).

Representative Methods. We select representative methods from five different categories for comprehensive performance comparison: **(i) MTCIR** methods including FashionNTM Pal et al. (2023) and CFIR Yuan & Lam (2021). **(ii) STFIR+NTM**: DQU-CIR Wen et al. (2024), single-turn methods FashionERN Chen et al. (2024a), and SPRC Bai et al. (2024), integrated with the multi-turn method FashionNTM. **(iii) ZS-CIR**: Pic2word Saito et al. (2023), Context-I2W Tang et al. (2024), and Image2Sentence Du et al. (2024), also integrated with FashionNTM. Additionally, since LLMs inherently support multi-turn interactions, we select several MLLMs as stronger baselines for comparison and fine-tune them on FashionMT using their original training methods. The selected baselines include: **(iv) Retrieval-capable MLLMs:** Fromage Koh et al. (2023), GILL Koh et al. (2024). We use the [IMG] and [RET] tags provided by these methods for retrieval. **(v) Interleaved MLLM:** MLLMs designed for interleaved multiple images and text, including MMICL Zhao et al. (2024a) and Flamingo Alayrac et al. (2022)-9B. For these methods, we perform retrieval by encoding the target's description text with the final-round LLM. All methods use ViT-L Radford et al. (2021) as the visual backbone for fair comparison.

Evaluation Metrics. Consistent with existing multimodal retrieval tasks Pal et al. (2023); Wen et al. (2024), we use the standard top-K recall metric to evaluate models' performance, denoted as R@K. Specifically, we adopt R@1, R@5, R@10, R@20 and their mean as the evaluation metrics.

5.2 Results

Quantitative Analysis. Experimental results on FashionMT are shown in Table 2. Benefiting from the strong multimodal fusion capability of the BLIP-2 architecture, methods such SPRC Bai et al. (2024) demonstrate performance advantages. Building upon this, our TSA and CCL incorporate captions as a transition, leveraging their semantic alignment with references and consistency with modified text. Furthermore, the proposed MIO effectively retains key semantics across multiple turns. Consequently, MAI significantly outperforms existing methods, achieving a remarkable 8.63 improvement in the Mean metric over the SOTA method SPRC.

Qualitative Analyses. In Figure 4, we compare MAI with two representative methods, FashionNTM and SPRC. MAI effectively handles fine-grained demands by leveraging TSA and CCL for

Table 2: Results on our proposed FashionMT dataset.

Method	Combination				Rollback				Mean
	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20	
CFIR (SIGIR'21)	11.70	23.09	30.89	40.14	8.25	22.63	31.04	41.79	26.19
Pic2word (CVPR'23)	13.35	27.12	35.42	45.40	8.69	23.98	33.15	44.41	28.94
GILL (NeurIPS'23)	19.54	38.17	47.63	56.14	9.12	24.56	33.62	42.77	33.95
Fromage (ICML'23)	19.45	39.12	49.00	59.65	10.12	26.54	34.97	45.61	35.56
FashionNTM (ICCV'23)	18.98	38.51	48.35	58.30	10.73	27.71	37.66	49.85	36.26
FashionERN (AAAI'24)	20.36	41.37	50.18	60.51	11.42	29.67	41.02	52.98	38.44
Flamingo (NeurIPS'22)	21.38	44.17	55.16	63.09	11.55	28.18	37.81	48.76	38.76
DQU-CIR (SIGIR'24)	20.57	42.32	52.33	62.03	12.59	31.69	42.79	54.68	39.88
Context-I2W (AAAI'24)	30.62	51.84	62.50	71.75	12.63	32.98	45.48	59.30	45.89
Image2Sentence (ICLR'24)	32.44	53.71	65.16	74.52	15.79	36.87	50.17	64.56	49.15
MMICL (ICLR'24)	39.17	60.89	70.28	79.89	18.46	43.53	57.05	69.66	54.87
SPRC (ICLR'24)	39.28	62.42	72.11	80.23	23.31	49.79	62.11	74.82	58.01
MAI (ours)	**51.51**	**74.67**	**80.66**	**86.52**	**28.94**	**58.89**	**70.42**	**81.50**	**66.64**

efficient aggregation of image-caption semantics, making it responsive to domain-specific terms like "crepe fabric" and "vintage design." Furthermore, MAI addresses retrospective-based needs by utilizing the MIO component to retain multi-turn historical key information, enabling precise interpretation of vague expressions such as "strap design."

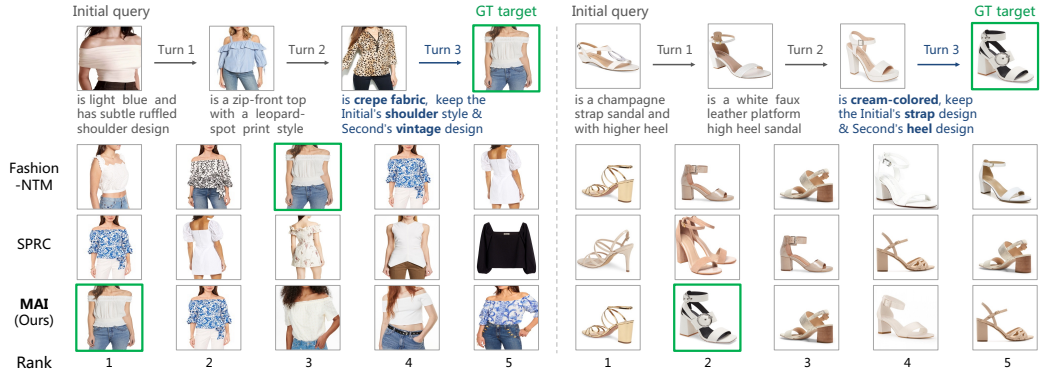

Figure 4: Qualitative results for the last turn in the FashionMT dataset. The top 5 retrieval results of MAI compared with two representative methods are shown.

Table 3: Ablation study on different components of the MAI model. Mean-Combination and Mean-Rollback denote the mean recall under the combination and rollback settings.

Settings	Mean-Combination	Mean-Rollback	Mean
Base	58.69	41.49	50.04
Base + TSA	69.22	55.74	62.48
Base + MIO	64.17	47.03	55.60
Base + TSA + CCL	72.31	58.83	65.57
Base + TSA + MIO	71.19	58.19	64.69
Base + TSA + CCL + MIO (**MAI**)	**73.34**	**59.94**	**66.64**

5.3 Ablation Studies

Effects of Different Components. Our baseline method employs Q-Former from BLIP-2 Li et al. (2023) for reference and modified text semantic fusion and adopts the multi-turn information aggregation model from FashionNTM for task adaptation. We gradually add the TSA, CCL and MIO, comparing their performance in both combination and rollback settings. Table 3 demonstrates the positive contribution of each component to performance improvement in both settings.

Table 4: Effects of TSA and CCL. Mean-C and Mean-S denote using caption adaptation and single-turn results.

Method	Mean-C	Mean-S
FashionERN 2024a	44.47	50.29
Image2Sentence 2024	47.61	51.64
FashionNTM 2023	48.03	45.75
MMICL 2024a	58.67	47.28
SPRC 2024	62.90	52.66
TSA + CCL	**65.57**	**53.73**

Table 5: The comparison between MIO and other methods on memory cost and average retrieval metrics. "N" represents # turns.

Method	Memory Cost (MB)	Mean
None	0	50.04
Concat	$64\times N$	50.56
LSTM 2012	$957 + 64\times N$	52.08
GRU 2017	$858 + 64\times N$	51.80
NTM 2023	$1270 + 64\times N$	53.19
MIO	64	**55.60**

Table 6: Effects of each loss in CCL. For simplicity, we denote $\mathcal{L}_{\mathrm{B}}(x, y)$ as $\mathcal{L}(x, y)$.

Settings	w/ CCL (total)	w/o CCL	w/o $\mathcal{L}(t_n^r, t_n^{tg})$	w/o $\mathcal{L}(t_n^{tg}, m_n)$	w/o $\mathcal{L}(m_n, c_n^{tg})$	w/o $\mathcal{L}(c_n^{tg}, t_n^r)$
Recall	**65.57**	62.48	63.54	64.39	64.51	64.90
Δ	-	-3.09	-2.03	-1.18	-1.06	-0.67

Effects of TSA and CCL. We further conduct two sets of experiments, as shown in Table 4 with Mean-C and Mean-S. (i) Caption adaptation. We adapt several representative methods to a two-stage fusion process, allowing the reference image to interact with both the caption text and the modified text. Specifically, FashionERN, Image2Sentence, and FashionNTM utilize the Combiner Baldrati et al. (2022) widely employed in this field, for interaction with caption embeddings. (ii) Single-turn retrieval. We evaluate performance using the first turn from FashionMT. The results in Table 4 indicate that the two-stage fusion significantly improves the performance of the methods. Additionally, the combination of TSA and CCL effectively integrates the critical semantics from the modified text. Consequently, it achieves superior retrieval performance compared to existing methods in both settings, shown in Table 4. We also conduct ablation experiments for each loss in CCL. Since CCL computes losses based on the outputs from TSA, it requires TSA to be present, resulting in 6 settings. Results in Table 6 show that each loss contributes to performance gains.

Effects of MIO. Due to the extensive storage of historical tokens in multi-turn retrieval, Table 5 presents the memory cost and mean retrieval performance of various methods. *None* denotes randomly initialized learned tokens t_N, *Concat* is concatenating all $t_n^{\mathrm{MIO}}, n \in [1, N-1]$, and *LSTM* and *GRU* respectively indicate using LSTM Graves & Graves (2012) and GRU Dey & Salem (2017) to aggregate all t_n^{MIO}. To adapt baselines to multi-step settings, we incorporate the multi-turn aggregation module *NTM* from the SOTA multi-turn method FashionNTM Pal et al. (2023). This module outperforms alternatives like LSTM or GRU. Since our parameter-free MIO can adaptively retain key semantics from historical turns and iterate over a set of learned tokens, it significantly reduces memory usage by **converting linear memory cost into constants**, while enhancing performance.

For further ablation studies on performance in existing datasets, reducing modality gap, and rollback setting, please refer to Section 7.1.

6 Conclusion and Discussion

In this paper, we have constructed the first dataset specifically designed for Multi-turn Composed Image Retrieval, named FashionMT. We also propose MAI model, a multi-turn key information-aware approach that uses paired captions as a transition for better semantic consistency and modality alignment while adaptively filtering and preserving significant attributes to reduce spatial occupancy. We have conducted extensive experiments on FashionMT and observed that MAI achieves state-of-the-art performance, demonstrating its usefulness and effectiveness.

Limitations. As the first dedicated MTCIR dataset, we standardize the number of turns to 3 for method comparison. However, real-world scenarios may involve more diverse transactions and cover more general contexts beyond e-commerce, aligning with our ongoing development efforts. Furthermore, we aim to upgrade our model with integrated dialogue and retrieval capabilities.

REFERENCES

Bijaya Adhikari, Parikshit Sondhi, Wenke Zhang, Mohit Sharma, and B Aditya Prakash. Mining e-commerce query relations using customer interaction networks. In *Proceedings of the 2018 World Wide Web Conference*, pp. 1805–1814, 2018.

Lorenzo Agnolucci, Alberto Baldrati, Marco Bertini, and Alberto Del Bimbo. Zero-shot image retrieval with human feedback. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9417–9419, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Shahbaz Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval. In *ICLR*. OpenReview.net, 2024.

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*, pp. 21466–21474, 2022.

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

Junyang Chen and Hanjiang Lai. Pretrain like you inference: Masked tuning improves zero-shot composed image retrieval. *arXiv preprint arXiv:2311.07622*, 2023.

Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3001–3011, 2020.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. *arXiv preprint arXiv:2311.10081*, 2023a.

Yanzhe Chen, Huasong Zhong, Xiangteng He, Yuxin Peng, and Lele Cheng. Real20m: A large-scale e-commerce dataset for cross-domain retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4939–4948, 2023b.

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*, pp. 1228–1236, 2024a.

Yanzhe Chen, Jiahuan Zhou, and Yuxin Peng. Spirit: Style-guided patch interaction for fashion image retrieval with text feedback. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024b.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Charles Corbiere, Hedi Ben-Younes, Alexandre Ramé, and Charles Ollion. Leveraging weakly annotated data for fashion image retrieval and label prediction. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 2268–2274, 2017.

Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pp. 1597–1600. IEEE, 2017.

Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Michael C Kampffmeyer, Xiaoyong Wei, Minlong Lu, Yaowei Wang, and Xiaodan Liang. M5product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21252–21262, 2022.

Mingjing Du, Shifei Ding, and Hongjie Jia. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99:135–145, 2016.

Yongchao Du, Min Wang, Wengang Zhou, Shuping Hui, and Houqiang Li. Image2sentence based asymmetrical zero-shot composed image retrieval. In *ICLR*. OpenReview.net, 2024.

Zhangchi Feng, Richong Zhang, and Zhijie Nie. Improving composed image retrieval via contrastive learning with scaling positives and negatives. *arXiv preprint arXiv:2404.11317*, 2024.

Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5337–5345, 2019.

Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14105–14115, 2022.

Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval. *arXiv preprint arXiv:2312.01998*, 2023.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13225–13234, 2024.

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.

Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pp. 1463–1471, 2017.

Yunpeng Han, Lisai Zhang, Qingcai Chen, Zhijian Chen, Zhonghua Li, Jianxin Yang, and Zhao Cao. Fashionsap: Symbols and attributes prompt for fine-grained fashion vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15028–15038, 2023.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2256–2264, 2024.

Jingjia Huang, Yinan Li, Jiashi Feng, Xinglong Wu, Xiaoshuai Sun, and Rongrong Ji. Clover: Towards a unified video-language alignment and fusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14856–14866, 2023.

Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. Learning instance-level representation for large-scale multi-modal pretraining in e-commerce. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11060–11069, 2023.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *ICLR*. OpenReview.net, 2024.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*, pp. 17283–17300. PMLR, 2023.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. *Advances in Neural Information Processing Systems*, 36, 2024a.

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*, pp. 2991–2999, 2024b.

Huayang Li, Siheng Li, Deng Cai, Longyue Wang, Lemao Liu, Taro Watanabe, Yujiu Yang, and Shuming Shi. Textbind: Multi-turn interleaved multimodal instruction-following in the wild. In *ACL (Findings)*, pp. 9053–9076. Association for Computational Linguistics, 2024a.

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*, pp. 19730–19742. PMLR, 2023.

Wenhao Li, Mengyuan Liu, Hong Liu, Pichao Wang, Jialun Cai, and Nicu Sebe. Hourglass tokenizer for efficient transformer-based 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 604–613, 2024b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

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*, pp. 2125–2134, 2021.

Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5753–5762, 2024b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR (Poster)*. OpenReview.net, 2019.

Mehryar Mohri. Foundations of machine learning, 2018.

Bolin Ni, JingCheng Hu, Yixuan Wei, Houwen Peng, Zheng Zhang, Gaofeng Meng, and Han Hu. Xwin-lm: Strong and scalable alignment practice for llms. *arXiv preprint arXiv:2405.20335*, 2024.

Anwesan Pal, Sahil Wadhwa, Ayush Jaiswal, Xu Zhang, Yue Wu, Rakesh Chada, Pradeep Natarajan, and Henrik I Christensen. Fashionntm: Multi-turn fashion image retrieval via cascaded memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11323–11334, 2023.

Sanghyuk Park, Minchul Shin, Sungho Ham, Seungkwon Choe, and Yoohoon Kang. Study on fashion image retrieval methods for efficient fashion visual search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

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*, pp. 8748–8763. PMLR, 2021.

Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018.

Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2765–2775, 2023.

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

Teresa Schubert. People choose options that leave future options open. *Nature Reviews Psychology*, 2(3):135–135, 2023.

Amin Muhammad Shoib, Jabeen Summaira, Changbo Wang, and Abdul Jabbar. Methods and advancement of content-based fashion image retrieval: A review. *arXiv preprint arXiv:2303.17371*, 2023.

Eunsuk Sung, Won Young Chung, and Daeho Lee. Factors that affect consumer trust in product quality: a focus on online reviews and shopping platforms. *Humanities and Social Sciences Communications*, 10(1):1–10, 2023.

Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26951–26962, 2024.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5180–5188, 2024.

Yuxin Tian, Shawn Newsam, and Kofi Boakye. Fashion image retrieval with text feedback by additive attention compositional learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1011–1021, 2023.

Peter M Todd and Gerd Gigerenzer. Précis of simple heuristics that make us smart. *Behavioral and brain sciences*, 23(5):727–741, 2000.

Noël Vouitsis, Zhaoyan Liu, Satya Krishna Gorti, Valentin Villecroze, Jesse C Cresswell, Guangwei Yu, Gabriel Loaiza-Ganem, and Maksims Volkovs. Data-efficient multimodal fusion on a single gpu. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27239–27251, 2024.

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*, pp. 915–923, 2023.

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 *SIGIR*, pp. 229–239. ACM, 2024.

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*, pp. 11307–11317, 2021.

Junda Wu, Rui Wang, Handong Zhao, Ruiyi Zhang, Chaochao Lu, Shuai Li, and Ricardo Henao. Few-shot composition learning for image retrieval with prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4729–4737, 2023.

Qian Xu and S Shyam Sundar. Lights, camera, music, interaction! interactive persuasion in e-commerce. *Communication Research*, 41(2):282–308, 2014.

Onesun Steve Yoo and Rakesh Sarin. Consumer choice and market outcomes under ambiguity in product quality. *Marketing Science*, 37(3):445–468, 2018.

Tong Yu, Yilin Shen, and Hongxia Jin. A visual dialog augmented interactive recommender system. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 157–165, 2019.

Yifei Yuan and Wai Lam. Conversational fashion image retrieval via multiturn natural language feedback. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 839–848, 2021.

Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11782–11791, 2021.

Gangjian Zhang, Shikun Li, Shikui Wei, Shiming Ge, Na Cai, and Yao Zhao. Multimodal composition example mining for composed query image retrieval. *IEEE Transactions on Image Processing*, 2024.

Ruiyi Zhang, Tong Yu, Yilin Shen, Hongxia Jin, and Changyou Chen. Text-based interactive recommendation via constraint-augmented reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

Ruiyi Zhang, Tong Yu, Yilin Shen, and Hongxia Jin. Text-based interactive recommendation via offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11694–11702, 2022.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. MMICL: empowering vision-language model with multi-modal in-context learning. In *ICLR*. OpenReview.net, 2024a.

Xiangyu Zhao, Yuehan Zhang, Wenlong Zhang, and Xiao-Ming Wu. Unifashion: A unified vision-language model for multimodal fashion retrieval and generation. *arXiv preprint arXiv:2408.11305*, 2024b.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. Mmchat: Multi-modal chat dataset on social media. In *LREC*, pp. 5778–5786. European Language Resources Association, 2022.

Hongyi Zhu, Jia-Hong Huang, Stevan Rudinac, and Evangelos Kanoulas. Enhancing interactive image retrieval with query rewriting using large language models and vision language models. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 978–987, 2024.

Husein Zolkepli, Aisyah Razak, Kamarul Adha, and Ariff Nazhan. Mmmmodal–multi-images multi-audio multi-turn multi-modal. *arXiv preprint arXiv:2402.11297*, 2024.

7 APPENDIX

7.1 PROOF OF PROPOSITION 1

For simplicity, given two representations x, y, we denote their similarity score as $s(x, y) = x \cdot y$. We first compare the i-th similarity score terms in L_{sim} and L_{pre} respectively:

$$\begin{aligned} S^i_{\text{naive}} =& \frac{1}{2}\left(\frac{|v^{ri}_n| + |m^{ri}_n|}{2} \cdot |c^{ti}_n| + \frac{|c^{ri}_n| + |m^{ri}_n|}{2} \cdot |v^{ti}_n|\right) \\ =& \frac{1}{4}[s(|v^{ri}_n| + |m^{ri}_n|, |c^{ti}_n|) + s(|c^{ri}_n| + |m^{ri}_n|, |v^{ti}_n|)], \\ S^i_{\text{pre}} - S^i_{\text{naive}} =& \frac{1}{8}[s(|v^{ri}_n|, |v^{ti}_n|) + s(|c^{ri}_n|, |c^{ti}_n|) - s(|v^{ri}_n|, |c^{ti}_n|) - s(|c^{ri}_n|, |v^{ti}_n|)]. \end{aligned}$$

The modality gaps between visual images and textual captions indicate that representations within the same modality are closer to each other, leading to higher similarity scores, i.e., $s(|v^{ri}_n|, |v^{ti}_n|) > s(|v^{ri}_n|, |c^{ti}_n|)$, and $s(|c^{ri}_n|, |c^{ti}_n|) > s(|c^{ri}_n|, |v^{ti}_n|)$. Therefore, $S^i_{\text{pre}} - S^i_{\text{naive}} > 0$. Notice that:

$$L_{\text{naive}} = L_{\text{sim}} + \frac{1}{B}\sum_{i=1}^{B}(1 - S^i_{\text{naive}}),$$
$$L_{\text{pre}} = L_{\text{sim}} + \frac{1}{B}\sum_{i=1}^{B}(1 - S^i_{\text{pre}}),$$

where $L_{\text{naive}} > L_{\text{pre}}$. Furthermore, based on the Rademacher Complexity Theory Mohri (2018), the upper bound of generalization errors is estimated as follows, with probability at least $1 - \delta$:

$$\begin{aligned} E[L_{naive}] \leq & E[L_{\text{sim}}] + \frac{1}{B}\sum_{i=1}^{B}(1 - S^i_{\text{naive}}) + 2R_B(G) + \sqrt{\frac{\log\frac{1}{\delta}}{2B}} \\ := & O(\text{GError}(\text{L}_{\text{naive}})), \\ E[L_{\text{pre}}] \leq & E[L_{\text{sim}}] + \frac{1}{B}\sum_{i=1}^{B}(1 - S^i_{\text{pre}}) + 2R_B(G) + \sqrt{\frac{\log\frac{1}{\delta}}{2B}} \\ := & O(\text{GError}(\text{L}_{\text{pre}})), \end{aligned}$$

where $R_B(G)$ is the Rademacher Complexity of the family of all possible loss functions, independent of our design for loss functions. From the above analysis, we have $O(\text{GError}(L_{\text{pre}})) < O(\text{GError}(L_{\text{naive}}))$, which indicates the superiority of our pre loss L_{pre} to the original naive cross-modal loss L_{naive}. □

7.2 More Ablation Studies

Validation on Existing Datasets. Despite the limitations of existing datasets, we further validate the effectiveness and generalization of our approach by adding performance comparisons on the real datasets MT FashionIQ Yuan & Lam (2021) and MT Shoes Pal et al. (2023). The results from Table 7 shows that our proposed approach, MAI, achieves the best performance in all settings due to its fine-grained semantic capture and efficient modality alignment.

Table 7: Validation on existing datasets MT FashionIQ and MT Shoes.

Method	Train on FashionMT Test on MT FashionIQ	Train on FashionMT Test on MT Shoes	Train on FashionMT Test on FashionMT	Means
FashionNTM Pal et al.	42.3	25.6	36.26	34.72
Image2Sentence Du et al.	43.8	28.2	49.15	40.38
MMICL Zhao et al.	46.9	29.8	54.87	43.86
SPRC Bai et al.	48.0	28.2	58.01	44.74
MAI (ours)	**50.6**	**33.8**	**66.64**	**50.35**

Modality Gap. We observe that recent works enhance feature-level representations to reduce modality gaps. We compare our approach with these methods, shown in the Table 8. The results show that noise addition methods are effective for large modality gaps, but once our approach reduces the gap through aligning modalities and semantics, the gains are limited.

Table 8: Comparison of various methods for reducing modality gaps.

Settings	Recall	Settings	Recall
Base	50.04	Base	50.04
+ Mixing Vouitsis et al.	49.87	+ **ours**	65.57
+ $N(0,1)$	52.17	+ ours + $N(0,1)$	65.23
+ $U(-1,1)$	52.09	+ ours + $U(-1,1)$	65.30
+ $N(0,1) \times U(-1,1)$ Gu et al.	54.88	+ ours + $N(0,1) \times U(-1,1)$	65.66

Ablation Study on Rollback. In the current setup for handling Rollback operations, the reference image for the current turn is replaced with the specified rollback image. We conduct comparative experiments under various Rollback settings: (1) Replace: the current setting. (2) Ignore: no replacement is performed. (3) Random: selecting randomly from previous turns. (4) Blend: using the PIL library's Image.blend() to merge two images into one. The results from Table 9 indicate that since the Rollback operation approximates redefining the current local optimal point, the setting Replace achieves the best performance.

Table 9: Ablation study on Rollback setting

Settings	Replace	Ignore	Random	Blend
Recall	**59.94**	49.59	53.62	57.80

7.3 Further clarification on the FashionMT dataset

We further clarify the setting, utility, quality control, and benefits of our FashionMT dataset below.

Explanation of Modified Text in Multi-turn. Our current approach for constructing modified text is based on two main reasons:

- **User Target Ambiguity.** Humans often make decisions heuristically, so selecting while browsing aligns with human intuition Todd & Gigerenzer (2000); Schubert (2023). In the e-commerce domain, our analysis of multi-turn interaction data from a well-known platform shows that users frequently experience "**target ambiguity**" during online shopping. Initially, users are unsure of the exact target and its details, and they need to select and refine attributes throughout the multi-turn interaction process. This behavior is also supported by psychological studies Yoo & Sarin (2018); Sung et al. (2023). To simulate this, we use combination and rollback settings to better mirror real-world scenarios.
- **Benchmark for Multi-turn.** Initially, we explored constructing the dataset by describing the difference between the current and target images. However, this "clear goal" setting resulted in models achieving precise retrieval within 1-2 turns. This results in the multi-turn retrieval task **degrading into a single-turn** retrieval task, thus failing to serve as a benchmark requiring algorithms to integrate information from multiple historical interactions.

Utility. Although the FashionMT is synthetically constructed, we conduct an in-depth analysis of user behaviors during multi-turn purchases on a famous e-commerce platform. We categorize these behaviors into two representative scenarios: "combination" and "rollback", aiming to replicate the real-world process where users refine their choices through iterative comparisons. Compared to existing datasets that concatenate single-turn retrieval data, FashionMT more accurately reflects real-world scenarios.

Table 10: Quality assessments among multi-turn datasets.

Datasets	Acc	HA	Gra	Con	Cov	Mean
MT FashionIQ Yuan & Lam	93.3	43.1	67.2	86.4	70.2	72.0
MT Shoes Pal et al.	95.1	65.6	76.3	91.3	80.9	81.8
MAI (ours)	96.2	98.7	91.3	90.5	93.2	94.0

Quality Control. Additionally, to validate its utility, we conduct a quality assessment of FashionMT and existing datasets, scoring them on a scale from 1 to 5 in the following aspects:

- Accuracy (**Acc**): Whether the modified text reflects the actual differences between images, with 1 being very inaccurate and 5 being very accurate.
- Historical Awareness (**HA**): Whether the modified text involves attributes from previous turns, with 1 being not involved and 5 being fully involved.

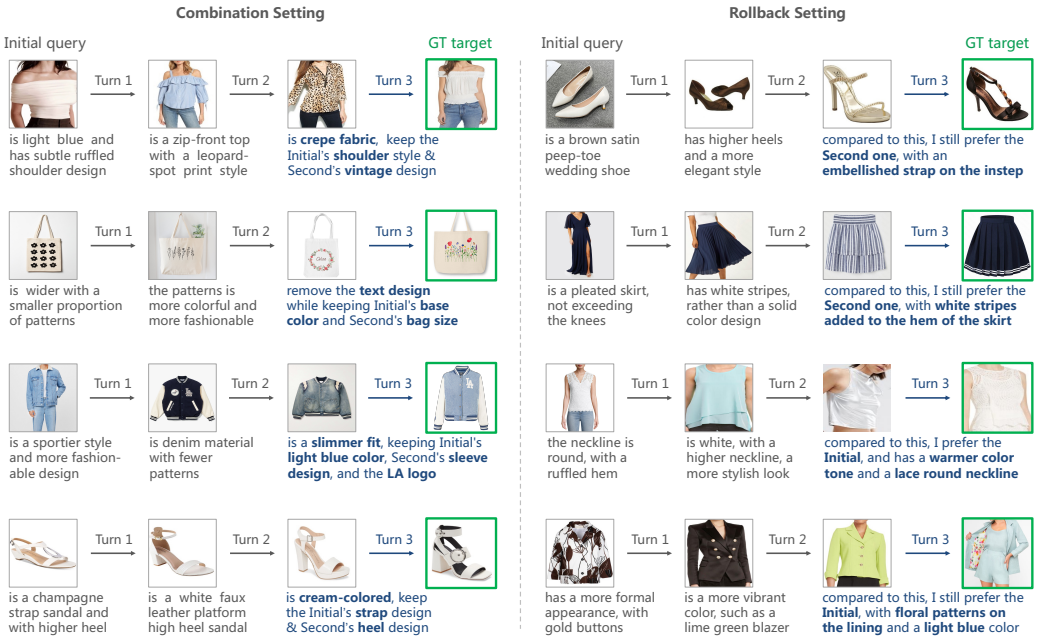

Figure 5: Examples of image sequences across multiple turns in the Combination and Rollback settings from our proposed FashionMT dataset.

- Granularity (**Gra**): Whether the text provides enough detail to cover subtle differences between images, with 1 being lacking detail and 5 being very detailed.
- Consistency (**Con**): Whether the differences between items in multi-turn retrieval are realistic, with 1 being unrealistic and 5 being very realistic.
- Coverage (**Cov**): Whether the description covers all major differences between items, with 1 being minimal coverage and 5 being comprehensive coverage.

We provided explanations of the rating requirements to 20 e-commerce platform staff members and calculated the average scores independently. The scores were then converted into percentages, as shown below. Specifically, we randomly selected 20% of the data from each dataset for manual scoring without informing the evaluators of the data source. Quality assessments among multi-turn datasets are shown in Table 10.

7.4 More Visualization

Dataset Examples. To facilitate a better understanding of our newly proposed dataset, FashionMT, we present data samples under the Combination and Rollback settings in Figure 5. In each transaction, the Ground Truth is highlighted with a green bounding box, and the retrospective-based modified text is marked in dark blue.

Dataset Statistics Visualizations. Figure 6 provides visualizations of various statistics in the FashionMT dataset. (1) The proportions of Combination and Rollback settings and their respective main categories. (2) The average length of modified text, along with separate averages for Combination and Rollback settings, and the average caption length. (3) A scale comparison between FashionMT and existing multi-turn datasets, MT FashionIQ and MT Shoes.

Modality gap. As shown in Figure 7, we visualize the modality gap between the query and target sides in the final round using t-SNE on FashionMT and existing datasets. Leveraging captions as a bridge between visual and textual modalities, our proposed MAI approach effectively reduces the modality gap between the query and target sides.

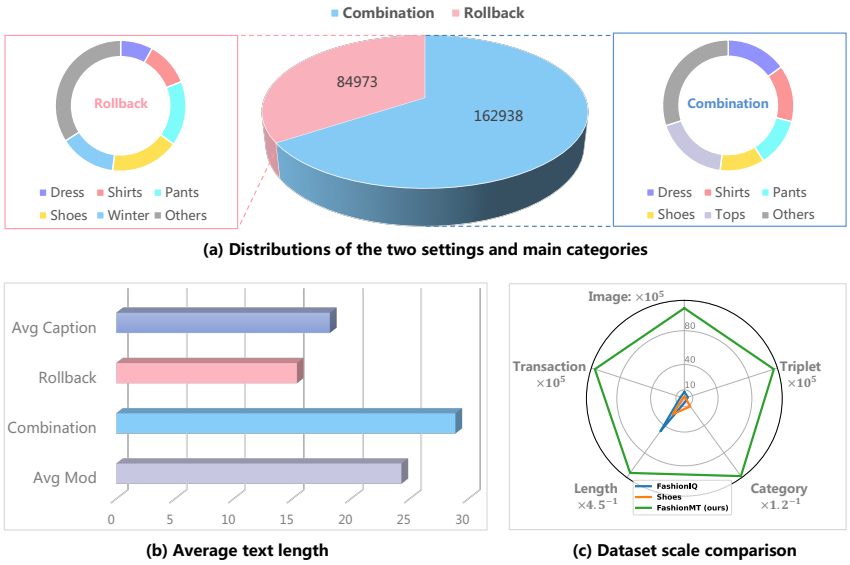

Figure 6: (1) Proportions of Combination and Rollback settings and main category distributions; (2) Average lengths of Modified text and captions; (3) Scale comparison of FashionMT with existing multi-turn datasets.

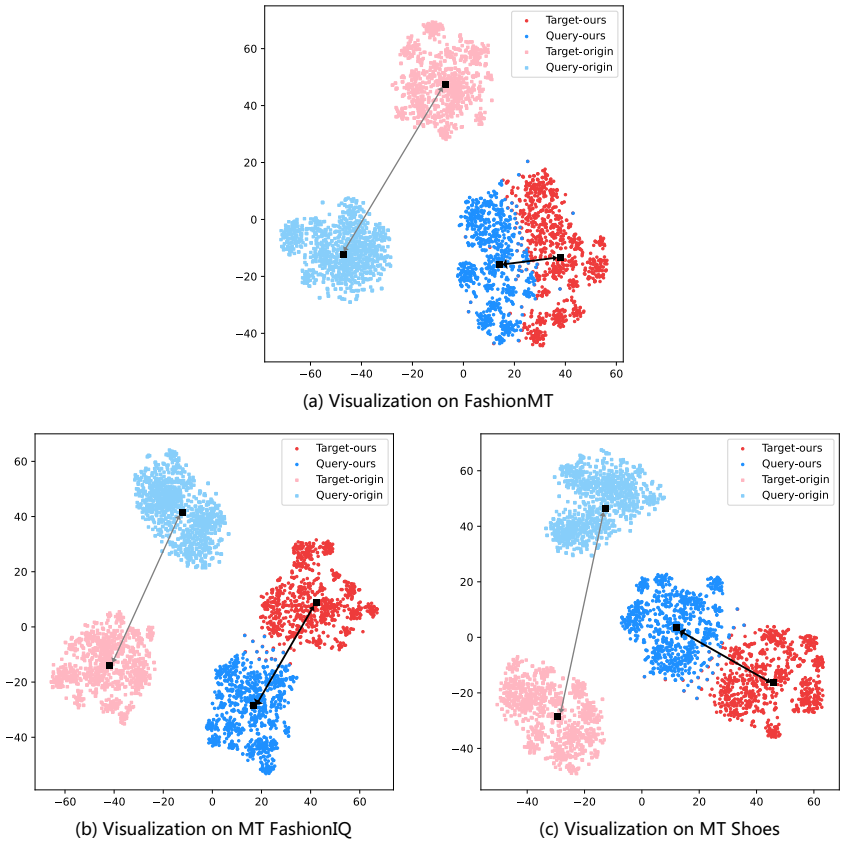

Figure 7: Visualization of modality gaps in FashionMT and existing datasets. Our approach significantly reduces the gap between the query and target sides.