TMCIR: Token Merge Benefits Composed Image Retrieval

Anonymous ACL submission

Abstract

Composed Image Retrieval (CIR) retrieves target images using a multi-modal query that combines a reference image with text describing desired modifications. The primary challenge is effectively fusing this visual and textual information. Current cross-modal feature fusion approaches for CIR exhibit an inherent bias in intention interpretation. These methods tend to disproportionately emphasize either the reference image features (visual-dominant fusion) or the textual modification intent (text-dominant fusion through image-to-text conversion). Such an imbalanced representation often fails to accurately capture and reflect the actual search intent of the user in the retrieval results. To address this challenge, we propose TMCIR, a novel framework that advances composed image retrieval through two key innovations: 1) Intent-Aware Cross-Modal Alignment. We first fine-tune CLIP encoders contrastively using intent-reflecting pseudo-target images, synthesized from reference images and textual descriptions via a diffusion model. This step enhances the encoder ability of text to capture nuanced intents in textual descriptions. 2) Adaptive Token Fusion. We further fine-tune all encoders contrastively by comparing adaptive token-fusion features with the target image. This mechanism dynamically balances visual and textual representations within the contrastive learning pipeline, optimizing the composed feature for retrieval. Extensive experiments on Fashion-IQ and CIRR datasets demonstrate that TMCIR significantly outperforms state-of-the-art methods, particularly in capturing nuanced user intent.

1 Introduction

Retrieving images based on a combination of a reference image and textual modification instructions defines the task of Composed ∼Image Retrieval (CIR) (Lee, 2005; Vo et al., 2019; Baldrati et al., 2022b). Specifically, the goal of CIR is to retrieve a target image from a candidate set that maintains overall visual similarity to the reference image while fulfilling the localized modification requirements specified in the textual description. CIR enables precise, interactive retrieval, making it valuable for applications like e-commerce and personalized web search.

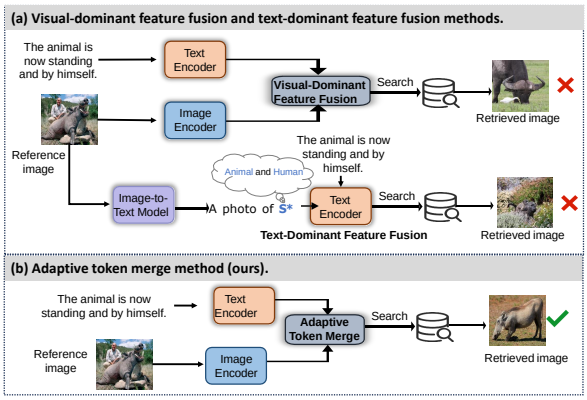

Figure 1: Workflows of existing CIR methods and our proposed TMCIR

However, composed queries from two distinct modalities introduces unique challenges. Unlike text-to-image or image-to-image retrieval which rely on a single query type, CIR must interpret *relative changes* described textually and apply them to the *specific visual content* of the reference image. The core difficulty lies in effectively integrating these cross-modal signals into a unified representation for similarity comparison with candidate images. To achieve this integration, most current approaches predominantly employ visual-dominant feature fusion mechanisms (Anwaar et al., 2021; Chen and Bazzani, 2020; Liu et al., 2021; Dodds et al., 2020). which extract image and text features separately and then combining them. However, these methods exhibit two critical limitations: 1) They often fail to preserve essential visual details from the reference image; 2) They tend to inadvertently incorporate irrelevant background information (i.e., regions unrelated to the textual modifi-

cations) into the final query representation. These shortcomings become particularly pronounced in scenarios requiring fine-grained image modifications, such as precise color variations or localized texture alterations, where maintaining both visual fidelity and modification accuracy is paramount.

Other recent approaches have adopted text-dominant fusion mechanisms that leverage CLIP-based image-to-text conversion (Gal et al., 2022; Baldrati et al., 2022b), where reference images are mapped to pseudo-word embeddings for integration with textual descriptions. While this paradigm benefits from established cross-modal alignment, it faces fundamental limitations: 1) The generated pseudo-word tokens primarily capture global image semantics while losing fine-grained visual details; 2) The constrained length of the word tokens restricts comprehensive visual representation. These factors lead to granular-level discrepancies in the cross-modal representations, ultimately compromising retrieval accuracy.

As shown in Fig. 1(a), both the visual-dominant fusion and the text-dominant fusion methods are fail to accurately capture and reflect the actual search intent of the user in the retrieval results. To address these challenges, we propose a TMCIR framework. Our framework is carefully designed to preserve the critical visual information present in the reference image while accurately conveying the modification intent of the user as specified by textual description. The TMCIR comprises two key steps: 1) **Intent-Aware Cross-Modal Alignment:** This step is designed to precisely capture textual intent from descriptions, addressing a critical limitation in CIR. Conventional target images often contain extraneous variations (e.g., lighting conditions, irrelevant background objects, or stylistic inconsistencies) that deviate from the specified intent of text, making them suboptimal for fine-tuning. To overcome this, we introduce a pseudo-target generation module that leverages a diffusion model conditioned on both the reference image and the relative textual description. The generated pseudo-target image eliminates noise, serving as a cleaner supervisory signal that faithfully reflects the intended modifications. Using an image-text paired dataset constructed from these pseudo-targets and their corresponding descriptions, we perform contrastive fine-tuning of pre-trained visual and textual encoders. This approach ensures precise cross-modal token alignment in a shared embedding space, with a focused emphasis on

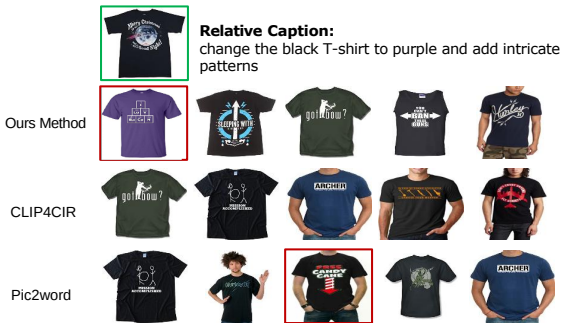

Figure 2: Retrieval examples using the proposed TMCIR, CLIP4CIR (Baldrati et al., 2022a) (visual-dominant feature fusion), and Pic2word (Saito et al., 2023) (text-dominant fusion) methods, respectively.

text-intent preservation from the description. 2) **Adaptive Token Fusion**: Following alignment fine-tuning, we introduce an adaptive fusion strategy that computes token-wise cosine similarity between visual and textual encoder outputs, enhanced with positional encoding for weighted feature fusion (Fig. 1(b). This design serves two key purposes: 1) The positional cues establish explicit correspondences between textual concepts and their spatial counterparts in the image. 2) The similarity-weighted fusion preserves critical visual details while precisely encoding the nuanced modification intents specified in relative descriptions. The fused representations then drive a final contrastive fine-tuning stage, where we optimize all encoders by comparing the adaptive token-fusion features against target images. This dynamic balancing mechanism simultaneously refines both modalities within a unified contrastive framework, ultimately producing composite features that are optimally discriminative for retrieval tasks.

Experiments on Fashion-IQ and CIRR show that our method achieves state-of-the-art performance. As illustrated in Fig. 2, visual-dominant fusion (CLIP4CIR) may introduce background noise, while text-dominant fusion (Pic2word) can suppress critical visual details, both leading to incorrect retrieval. In contrast, our adaptive token fusion selectively aligns visual regions with key textual concepts, emphasizing relevant object tokens (e.g., the T-shirt) while attenuating background information, resulting in more accurate modeling of fine-grained modifications.

In summary, our contributions are as follows:

- We propose a novel CIR approach that integrates intent-aware cross-modal alignment (IACMA) and adaptive token fusion (ATF) to better capture user intent. The IACMA lever-

ages a diffusion model to generate pseudo-target images that more accurately reflect user modification intent compared to potentially noisy real target images, providing a purer supervisory signal for encoder fine-tuning.

- The ATF adaptively fuse visual and textual tokens through weighted integration and positional encoding. This strategy ensures comprehensive preservation of key visual details while accurately capturing subtle user modification intent.

- Experimental results on the Fashion-IQ and CIRR datasets indicate that our proposed method outperforms current state-of-the-art CIR approaches in both retrieval accuracy and robustness.

2 Related work

Composed Image Retrieval. Existing CIR methods mainly fall into two categories. The first fuses reference image features with relative captions and matches the fused representation against candidate images, leveraging feature fusion and attention mechanisms for effective retrieval (Anwaar et al., 2021; Chen et al., 2020; Dodds et al., 2020; Liu et al., 2021; Vo et al., 2019). With the advent of pre-trained models, recent approaches further combine independently trained visual and textual encoders to improve performance (Baldrati et al., 2022a; Goenka et al., 2022; Ray et al., 2023). The second category transforms the reference image into a pseudo-word embedding and performs text-to-image retrieval (Saito et al., 2023; Baldrati et al., 2023). However, such embeddings often capture only coarse semantics and lack fine-grained visual details, limiting representation capacity. Prompt-based variants (Liu et al., 2023b) also struggle to modify intrinsic caption semantics, constraining retrieval accuracy. Overall, existing methods suffer from information loss and insufficient alignment between visual details and user intent. To address these issues, we propose TMCIR, which leverages pseudo-target image generation, task-specific encoder fine-tuning, and similarity-based token fusion, achieving improved retrieval accuracy and robustness.

Token Merge for Modal Fusion. Multi-modal fusion requires efficient representation learning to reduce redundancy and computational cost. Token Merge was initially proposed for vision Transformers to merge redundant tokens and has since been extended to multimodal tasks, improving efficiency and mitigating cross-modal conflicts in image–text and video understanding (Bolya et al., 2022; Chen et al., 2024; Luo et al., 2023; Liu et al., 2023a; Shen et al., 2023). In composed image retrieval, Token Merge helps address granularity mismatches in cross-modal alignment, with prior work exploring entity-level alignment and contrastive strategies (Wang et al., 2023, 2024). Compared to attention-heavy interaction mechanisms, Token Merge provides a flexible and scalable fusion pathway. Building on this, we apply Token Merge to visual–textual fusion in composed image retrieval, adaptively merging tokens to preserve fine-grained visual details and textual semantics, leading to more accurate cross-modal alignment and improved retrieval performance.

3 Method

3.1 Preliminary

Assume that a composed image retrieval (CIR) dataset contains N annotated triplet samples, where the ith triplet sample x_i is represented as:

$$x_i = (r_i, m_i, t_i), \quad r_i, t_i \in \Omega, \quad m_i \in \mathcal{T}. \quad (1)$$

Here, r_i, m_i, and t_i denote the reference image, the *relative* description, and the target image of the ith triplet sample, respectively. The term *relative* emphasizes that m_i specifies the modifications to be applied to r_i to obtain the target image, capturing how the target image differs from the reference image. Ω represents the candidate image set that contains all reference and target images, and $\mathcal{T}$ denotes the text set containing all relative descriptions.

In the CIR task, the query q_i, which is composed of the reference image r_i and the relative description m_i, is used to retrieve the target image t_i from the candidate set Ω. In the classical CIR training paradigm, multiple annotated triplet samples are first grouped into a mini-batch. Within the same batch, the reference images and relative descriptions are encoded into query representations by a query encoder $F(\cdot)$, while the target images are encoded by an image encoder $G(\cdot)$ to obtain their embeddings. For brevity, we denote the representations of the triplet (r_i, m_i, t_i) as

$$q_i = F(r_i, m_i) \quad \text{and} \quad v_i = G(t_i),$$

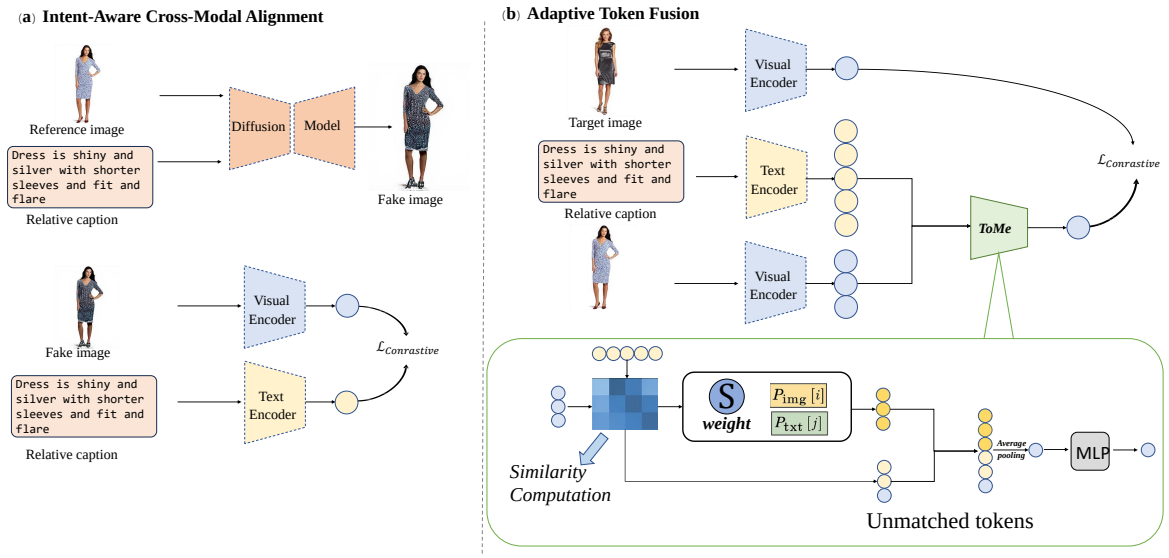

Figure 3: An Overview of the TMCIR Framework.It consists of two modules: the "Intent-Aware Cross-Modal Alignment" module and the "Adaptive Token Fusion" module. First, we input the reference image and the relative description into a diffusion model to generate a pseudo-target image. Through contrastive learning, we guide the visual and textual encoders to achieve cross-modal token distribution alignment. Then, the reference image and the relative description are fused using an adaptive token fusion strategy based on positional encoding and semantic similarity, generating a joint representation that captures both the user intent and the key visual information from the reference image.

where v_i denotes the embedding of the target image. The cosine similarity function $f(\cdot, \cdot)$ computes the similarity between the query representation and the target image embedding. Most current methods adopt a contrastive learning paradigm, which pulls together the query and target image representations of positive pairs (i.e., the query paired with its matching target image) while pushing apart those of negative pairs (i.e., a query paired with a target image from a different triplet). The corresponding loss function is formulated as:

$$L_{cl} = \frac{1}{B} \sum_{i=1}^{B} -\log \left(\frac{\exp(f(q_i, v_i)/\tau)}{\sum_{j=1}^{B} \exp(f(q_i, v_j)/\tau)} \right), \tag{2}$$

where B is the batch size and τ is the temperature hyperparameter, which controls the sharpness of the similarity distribution and thus regulates the strength of the contrastive signal.

Our method follows the same contrastive learning paradigm.

3.2 Overview

As depicted in Figure 1, our proposed TMCIR pipeline comprises two steps:

(1) **Intent-Aware Cross-Modal Alignment.** This step contains **Pseudo Target Image Generation** (PTIG) and **Encoder Fine-Tuning for Token Alignment** (EFTTA) modules. In the PTIG module, a diffusion model, specifically Stable Diffusion 3, is utilized to generate a pseudo target image p_i by conditioning on the reference image r_i and the relative description m_i. This pseudo target image accurately reflects the modification requirements specified in m_i, ensuring a high level of controllability and reproducibility. In the EFTTA module, we construct an image-text pair dataset from the relative description and the pseudo target image, then fine-tune the visual and text encoders of the CLIP model. This process promotes more consistent cross-modal token distributions in the shared embedding space. Here, *token alignment* refers to the process of harmonizing tokens from different modalities in the embedding space so that their semantic representations and attention patterns become more correlated and comparable.

(2) **Adaptive Token Fusion.** After obtaining visual and text tokens from the fine-tuned bimodal encoder, we design an adaptive token fusion strategy. In our approach, token merging is performed on a token-by-token basis by computing the cosine similarity between individual tokens and incorporating positional encoding via weighted averaging. This fusion strategy constructs a unified and seman-

tically rich cross-modal representation.

In the following subsections, we provide details for these two steps.

3.3 Intent-Aware Cross-Modal Alignment

The Intent-Aware Cross-Modal Alignment step aims to enhance the encoder ability of text to capture nuanced intents in textual descriptions, which includes pseudo target image generation and encoder fine-tuning for token alignment modules.

Pseudo Target Image Generation Large-scale vision–language models such as CLIP show strong generalization and have been widely applied to CIR. However, directly using pre-trained visual and text encoders often yields inconsistent token distributions across modalities in CIR settings, leading to suboptimal token fusion and degraded retrieval performance.

To address this issue, we first select image-text pairs from existing CIR datasets—typically sampling all available pairs or a fixed number per batch in our experiments—and perform task-specific fine-tuning of the visual and text encoders to achieve more consistent token representations from the reference image and the relative description. Considering that the manually collected triplet samples in current CIR datasets contain target images sourced from diverse origins (which may include background interference or noise not aligned with the modification description), we generate a pseudo target image p_i by conditioning a diffusion model D on the reference image r_i and the relative description m_i:

$$p_i = D(r_i, m_i). \tag{3}$$

The pseudo target image p_i accurately embodies the modification requirements stipulated in m_i, while excluding irrelevant background noise. This provides a purer and more precise supervisory signal for subsequent encoder fine-tuning. The pseudo target image p_i and the relative description m_i are then combined to form an image-text pair dataset $\mathcal{D}$:

$$\mathcal{D} = \{(m_i, p_i) \mid i = 1, 2, \ldots, N\}. \tag{4}$$

We utilize $\mathcal{D}$ as a dedicated dataset in a distinct training stage for fine-tuning the encoders, separate from the original CIR training set.

Encoder Fine-Tuning for Token Alignment After constructing the image-text pair dataset $\mathcal{D} = \{(m_i, p_i)\}_{i=1}^{N}$ as described in Section 3.3, we adopt a contrastive learning strategy to fine-tune the CLIP-pretrained visual encoder and text encoder, thereby further enhancing their cross-modal representation and alignment abilities for the CIR task.

Specifically, for each image-text pair (m_i, p_i), the relative description m_i is input into the text encoder E_T to obtain the text feature vector:

$$t_i = E_T(m_i) \tag{5}$$

while the pseudo target image p_i is input into the visual encoder E_V to obtain the image feature vector:

$$v_i = E_V(p_i) \tag{6}$$

We then normalize these feature representations:

$$\hat{t}_i = \frac{t_i}{\|t_i\|_2}, \quad \hat{v}_i = \frac{v_i}{\|v_i\|_2} \tag{7}$$

to facilitate subsequent cosine similarity calculations.

Next, we optimize the model parameters using the InfoNCE loss function:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\hat{v}_i, \hat{t}_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\hat{v}_i, \hat{t}_j)/\tau)} \tag{8}$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity function and τ is a learnable temperature parameter that controls the smoothness of the similarity distribution.

Following fine-tuning, the token representations produced by the visual and text encoders in the shared embedding space exhibit improved distribution consistency and cross-modal alignment, thereby providing a robust foundation for the subsequent token merging module.

3.4 Adaptive Token Fusion

Following the fine-tuning of the visual and text encoders, the resulting visual tokens and text tokens are well-aligned in the shared embedding space. To effectively integrate the dual-modal information, we design a token merging module based on similarity computation.

Specifically, the reference image r_i and the relative description m_i are input into the fine-tuned visual encoder E_V and text encoder E_T, respectively, to obtain the corresponding sets of tokens:

$$V = \{v_1, v_2, \ldots, v_L\} = E_V(r_i), \tag{9}$$

$$T = \{t_1, t_2, \ldots, t_M\} = E_T(m_i). \tag{10}$$

Table 1: Quantitative comparison across competing methods on the CIRR test set, where Avg. indicates the average results across all the metrics in the three different settings. The best results are marked in bold

Methods	**Recall@K**				**$\mathbf{R}_{\text{subset}}$@K**			**Avg.**
	K=1	K=5	K=10	K=50	K=1	K=2	K=3	
LF-BLIP (Baldrati et al., 2022b)	20.89	48.07	61.16	83.71	50.22	69.39	86.82	60.58
LF-CLIP (Baldrati et al., 2022b)	33.59	65.35	77.35	95.21	62.39	79.74	92.02	72.53
CLIP4CIR (Baldrati et al., 2022a)	38.53	69.98	81.86	95.93	68.19	86.31	94.17	69.09
BLIP4CIR+Bi (Liu et al., 2024)	40.15	73.08	83.88	96.27	72.10	90.16	95.93	72.59
CompoDiff (Gu et al., 2023)	22.35	54.36	73.41	91.77	35.84	58.21	76.60	29.10
CASE (Levy et al., 2024)	48.00	79.11	87.25	97.57	75.88	94.67	96.00	77.50
CASE Pre-LaSCo.CaT (Levy et al., 2024)	49.35	80.02	88.75	97.47	76.48	95.03	95.71	78.25
TG-CIR (Wen et al., 2023)	45.25	78.29	87.16	97.30	72.84	89.25	95.13	75.57
DRA (Jiang et al., 2023)	39.93	72.07	83.83	96.43	71.04	91.43	94.72	71.55
CoVR-BLIP (Ventura et al., 2024)	49.69	78.60	86.77	94.31	75.01	91.07	93.16	80.81
Re-ranking (Liu et al., 2023b)	50.55	81.75	89.78	97.18	80.04	94.29	96.80	80.90
SPRC (Bai et al., 2023)	51.96	82.12	89.74	97.69	80.65	92.31	96.60	81.39
CIR-LVLM (Sun et al., 2024)	53.64	83.76	90.60	97.93	79.12	92.33	96.67	81.44
TMCIR (Ours)	**54.12**	**84.27**	**91.06**	**98.43**	**82.64**	**92.45**	**96.77**	**82.66**

We then compute the similarity matrix $\mathbf{S} \in \mathbb{R}^{L \times M}$ between the image token set V and the text token set T:

$$\mathbf{S}_{ij} = \frac{\mathbf{v}_i \cdot \mathbf{t}_j}{\|\mathbf{v}_i\| \cdot \|\mathbf{t}_j\|}. \tag{11}$$

For each text token, we iterate over each visual token, considering a pair $(\mathbf{v}_i, \mathbf{t}_j)$ as a valid matching token pair if their similarity exceeds a preset threshold τ.

Fusion Strategy: For each matching token pair $(\mathbf{v}_i, \mathbf{t}_j)$, we compute a weighted average using their similarity coefficient λ as the weight and integrate their positional encodings ($\mathbf{P}_{\text{img}}[i]$ and $\mathbf{P}_{\text{txt}}[j]$) to preserve spatial information. The resulting fused representation $\mathbf{f}_{i,j}$ is calculated as:

$$\mathbf{f}_{i,j} = \frac{\mathbf{S}_{ij} \cdot \mathbf{v}_i + \mathbf{S}_{ij} \cdot \mathbf{t}_j}{2\mathbf{S}_{ij} + \epsilon} + 0.5 \cdot \left(\mathbf{P}_{\text{img}}[i] + \mathbf{P}_{\text{txt}}[j]\right), \tag{12}$$

where ϵ is a small constant to prevent division by zero.

For visual tokens and text tokens that do not find a matching counterpart, we retain them by directly adding their positional residuals:

$$\tilde{\mathbf{v}}_i = \mathbf{v}_i + 0.5 \cdot \mathbf{P}_{\text{img}}[i], \quad \text{if } \mathbf{v}_i \notin \text{Matched}, \tag{13}$$

$$\tilde{\mathbf{t}}_j = \mathbf{t}_j + 0.5 \cdot \mathbf{P}_{\text{txt}}[j], \quad \text{if } \mathbf{t}_j \notin \text{Matched}. \tag{14}$$

Finally, we concatenate all the fused tokens and the remaining unmatched tokens to form the cross-modal token sequence $\mathbf{Z}$:

$$\mathbf{Z} = \{\mathbf{f}_{ij} \mid (i,j) \in \mathcal{M}\} \cup \{\tilde{\mathbf{v}}_i \mid i \notin \mathcal{M}_I\} \cup \{\tilde{\mathbf{t}}_j \mid j \notin \mathcal{M}_T\} \tag{15}$$

where $\mathcal{M}$ denotes the set of matching pairs, while $\mathcal{M}_I$ and $\mathcal{M}_T$ represent the matching indices for visual and text tokens, respectively. We then apply average pooling to $\mathbf{Z}$ to obtain a single token representation $\mathbf{z}$:

$$\mathbf{z} = \frac{1}{N_Z} \sum_{n=1}^{N_Z} \mathbf{Z}_n, \tag{16}$$

and pass it through a fully connected layer F to obtain the final cross-modal embedding vector V_Q:

$$V_Q = \mathrm{F}(\mathbf{z}). \tag{17}$$

Learning Objective: Our training objective for the composed image retrieval (CIR) task is to align the joint feature representation V_Q of the mixed-modal query (r, m) with the feature representation V_T of the target image t. In each training iteration, we process a mini-batch of samples:

$$\{(V_Q^{(i)}, V_T^{(i)})\}_{i=1}^{N_B}, \tag{18}$$

where $(V_Q^{(i)}, V_T^{(i)})$ denotes the feature representations of the ith (mixed-modal query, target image)

Table 2: Quantitative comparison across competing methods on the Fashion-IQ validation set, where Average indicates the average results across all the metrics in the three different classes. The best results are marked in bold

Methods	Dress		Shirt		Top&Tee		Average		
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	Rmean
DCNet (Kim et al., 2021)	28.95	56.07	23.95	47.30	30.44	58.29	27.78	53.89	40.84
SAC w/BERT (Jandial et al., 2022)	26.52	51.01	28.02	51.86	32.70	61.23	29.08	54.70	41.89
FashionVLP (Goenka et al., 2022)	32.42	60.29	31.89	58.44	38.51	68.79	34.27	62.51	48.39
LF-CLIP(Combiner) (Baldrati et al., 2022b)	31.63	56.67	36.36	58.00	38.19	62.42	35.39	59.03	47.21
LF-BLIP (Levy et al., 2024)	25.31	44.05	25.39	43.57	26.54	44.48	25.75	43.98	34.88
CASE (Zhu et al., 2023)	47.44	69.36	48.48	70.23	50.18	72.24	48.79	70.68	59.74
AMC (Ventura et al., 2024)	31.73	59.25	30.67	59.08	36.21	66.06	32.87	61.64	47.25
CoVR-BLIP (Baldrati et al., 2022a)	44.55	69.03	48.43	67.42	52.60	74.31	48.53	70.25	59.39
CLIP4CIR (Liu et al., 2024)	33.81	59.40	39.99	60.45	41.41	65.37	38.32	61.74	50.03
BLIP4CIR+Bi (Han et al., 2023)	42 09	67.33	41.76	64.28	46.61	70.32	43.49	67.31	55.04
FAME-ViL (Wen et al., 2023)	42.19	67.38	47.64	68.79	50.69	73.07	46.84	69.75	58.29
TG-CIR (Jiang et al., 2023)	45.22	69.66	52.60	72.52	56.14	77.10	51.32	73.09	58.05
Re-ranking (Liu et al., 2023b)	48.14	71.43	50.15	71.25	55.23	76.80	51.17	73.13	62.15
CompoDiff (Gu et al., 2023)	40.65	57.14	36.87	57.39	43.93	61.17	40.48	58.57	49.53
SPRC (Bai et al., 2023)	49.18	72.43	55.64	73.89	59.35	78.58	54.92	74.97	64.85
TMCIR (Ours)	**50.67**	**73.86**	**59.12**	**76.34**	**59.93**	**79.46**	**56.57**	**76.55**	**66.56**

pair, and N_B is the mini-batch size. The batch-based classification loss function is defined as:

$$L = \frac{1}{N_B} \sum_{i=1}^{N_B} -\log \frac{\exp(\lambda \cdot \text{Sim}(V_Q^{(i)}, V_T^{(i)}))}{\sum_{j=1}^{NB} \exp(\lambda \cdot \text{Sim}(V_Q^{(j)}, V_T^{(j)}))}, \tag{19}$$

where $\text{Sim}(\cdot)$ denotes the cosine similarity function and λ is a temperature parameter.

4 Experiment

4.1 Experimental Setup

Implementation Details. Our method is implemented in PyTorch and runs on an NVIDIA RTX A100 GPU with 80GB of memory. We adhere to the design principles of CLIP, initializing both the visual and text encoders from a CLIP pre-trained model based on the ViT-L architecture. The AdamW optimizer (Loshchilov and Hutter, 2017) is employed with a weight decay coefficient set to 0.05. Input images are resized to 224×224 pixels, and a padding ratio of 1.25 is applied for uniform processing (Baldrati et al., 2022b). The initial learning rates are set to 1e-5 and 2e-5 for the CIRR and Fashion-IQ datasets, respectively, and a cosine learning rate scheduling strategy is adopted. The similarity threshold τ is set to 0.7.In the Pseudo-Target Generation stage, the diffusion model FLUX.1-dev-edit-v0 is utilized.

4.2 Comparative study

Results on Fashion-IQ Table 2 reports the results on the Fashion-IQ dataset. TMCIR achieves the best recall across all eight metrics and three categories. Compared to the second-best method SPRC, TMCIR improves R@10 from 54.92 to 56.57, with particularly notable gains on the Shirt category (R@10: 55.64 → 59.12; R@50: 73.89 → 76.34). This advantage stems from TMCIR's direct fusion of visual and textual tokens via token merging, which better preserves fine-grained visual details than text-only retrieval paradigms, leading to more accurate matching.

Results on CIRR Table 1 reports results on the CIRR dataset. Compared to the text prompt-based method SPRC, TMCIR consistently improves recall across all metrics (e.g., R@1: 51.96 → 54.12, R@10: 89.74 → 91.06), demonstrating stronger generalization. These gains indicate that prompt-based approaches struggle to capture fine-grained visual details, while TMCIR better preserves visual information through direct multimodal fusion. TMCIR also achieves superior performance on Recall-subset@K metrics and outperforms the strongest CIR-LVLM baseline, highlighting the effectiveness of similarity-based token fusion and task-specific encoder fine-tuning for composed image retrieval.

Table 3: Ablation studies with regard to the impact of using pseudo versus real target images on retrieval performance.

Method	FashionIQ		CIRR		
	R@10	R@50	R@1	R@5	R_{subset}@1
Real	54.85	75.43	53.62	83.62	79.05
Pseudo	56.57	76.55	54.12	84.27	82.64

Table 4: Ablation studies with regard to the contribution of the token merging module to retrieval performance.

Method	FashionIQ		CIRR		
	R@10	R@50	R@1	R@5	R_{subset}@1
w/o token-merge	29.68	54.85	20.88	48.24	50.33
token-merge	56.57	76.55	54.12	84.27	82.64

Table 5: Ablation studies with regard to the performance differences between pre-trained and fine-tuned models.

Method	FashionIQ		CIRR		
	R@10	R@50	R@1	R@5	R_{subset}@1
Pre-trained	54.42	75.67	53.22	83.47	79.16
fine-tuning	56.57	76.55	54.12	84.27	82.64

4.3 Ablation Study

Pseudo-target Images vs. Real Target Images We compare encoder fine-tuning using real target images versus diffusion-generated pseudo-target images. As shown in Table 3, pseudo-target images provide cleaner supervision by better capturing the intended modifications, leading to improved alignment and robustness in token fusion, whereas real targets may introduce background noise and hinder fine-grained semantic modeling.

Pre-trained Models vs. Task-specific Fine-tuning To evaluate task-specific fine-tuning, we compare a frozen CLIP encoder with a fine-tuned variant using pseudo-target image–text pairs and contrastive learning. As shown in Table 5, fine-tuning yields consistent improvements across all metrics (e.g., +0.89% Recall@1), indicating better cross-modal alignment and more robust token representations. These results highlight the importance of task-specific contrastive fine-tuning for improving CIR performance.

Contribution of the Token Merging Module To assess the token merging module, we compare a variant with token merging against one that directly pools visual and textual tokens without fusion. As shown in Table 4, removing token merging leads to clear performance drops in cross-modal retrieval, especially on Recall@K metrics. These results confirm that token merging effectively suppresses noise, enhances cross-modal alignment, and produces more consistent fused representations, thereby improving retrieval accuracy.

5 Conclusion

Addressing the challenge of biased feature fusion in Composed Image Retrieval (CIR), we introduced TMCIR. Our framework leverages **Intent-Aware Cross-Modal Alignment (IACMA)**, using diffusion-generated pseudo-target images for cleaner encoder fine-tuning, and **Adaptive Token Fusion (ATF)**, which merges tokens based on similarity and position to balance modalities. Extensive experiments demonstrate that TMCIR significantly outperforms state-of-the-art methods on the Fashion-IQ and CIRR benchmarks. By effectively preserving visual details while accurately capturing textual modification intent, TMCIR offers a more robust and precise solution for CIR.

Limitations

Despite the strong performance of TMCIR, several limitations remain. First, the intent-aware cross-modal alignment stage relies on diffusion-generated pseudo-target images, which introduces additional computational overhead and depends on the quality and controllability of the generative model. For complex or abstract modification descriptions, pseudo targets may not always perfectly reflect user intent, potentially affecting encoder fine-tuning. Second, the adaptive token fusion mechanism requires token-wise similarity computation between visual and textual representations. While effective for preserving fine-grained details, this design may incur scalability issues when handling high-resolution images or longer text inputs, limiting efficiency in large-scale or real-time retrieval settings. Finally, our evaluation is restricted to Fashion-IQ and CIRR, which primarily focus on object-centric attribute modifications. The generalization of TMCIR to more complex scenes, diverse domains, or interactive retrieval scenarios with iterative user feedback remains an open question and a promising direction for future work.

References

Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. 2021. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, pages 1140–1149.

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

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

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022a. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4959–4968.

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

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2022. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*.

Kaibing Chen, Dong Shen, Hanwen Zhong, Huasong Zhong, Kui Xia, Di Xu, Wei Yuan, Yifei Hu, Bin Wen, Tianke Zhang, and 1 others. 2024. Evlm: An efficient vision-language model for visual understanding. *arXiv preprint arXiv:2407.14177*.

Yanbei Chen and Loris Bazzani. 2020. Learning joint visual semantic matching embeddings for language-guided retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 136–152. Springer.

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

Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. 2020. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

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

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

Xiao Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, and Tao Xiang. 2023. Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2669–2680.

Surgan Jandial, Pinkesh Badjatiya, Pranit Chawla, Ayush Chopra, Mausoom Sarkar, and Balaji Krishnamurthy. 2022. Sac: Semantic attention composition for text-conditioned image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4021–4030.

Xintong Jiang, Yaxiong Wang, Yujiao Wu, Meng Wang, and Xueming Qian. 2023. Dual relation alignment for composed image retrieval. *arXiv preprint arXiv:2309.02169*.

Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. 2021. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1771–1779.

Newton Lee. 2005. Interview with bill kinder: January 13, 2005. *Comput. Entertain.*, 3(1):4.

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.

Yuqi Liu, Luhui Xu, Pengfei Xiong, and Qin Jin. 2023a. Token mixing: parameter-efficient transfer learning from image-language to video-language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1781–1789.

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

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

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023. Cheap and quick: Efficient vision-language instruction tuning for large language models. *Advances in Neural Information Processing Systems*, 36:29615–29627.

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

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

Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. 2023. Scaling vision-language models with sparse mixture of experts. *arXiv preprint arXiv:2303.07226*.

Zelong Sun, Dong Jing, Guoxing Yang, Nanyi Fei, and Zhiwu Lu. 2024. Leveraging large vision-language model as user intent-aware encoder for composed image retrieval. *arXiv preprint arXiv:2412.11087*.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. 2024. Covr-2: Automatic data construction for composed video retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448.

Xiaodan Wang, Lei Li, Zhixu Li, Xuwu Wang, Xiangru Zhu, Chengyu Wang, Jun Huang, and Yanghua Xiao. 2023. Agree: Aligning cross-modal entities for image-text retrieval upon vision-language pretrained models. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 456–464.

Yifan Wang, Liyuan Liu, Chun Yuan, Minbo Li, and Jing Liu. 2024. Negative-sensitive framework with semantic enhancement for composed image retrieval. *IEEE Transactions on Multimedia*.

Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. 2023. Target-guided composed image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 915–923.

Hongguang Zhu, Yunchao Wei, Yao Zhao, Chunjie Zhang, and Shujuan Huang. 2023. Amc: Adaptive multi-expert collaborative network for text-guided image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6):1–22.