# Learning with Noisy Triplet Correspondence for Composed Image Retrieval

Shuxian Li[1,*], Changhao He[1,*], Xiting Liu[2], Joey Tianyi Zhou[3], Xi Peng[1,4], Peng Hu[1,†]
[1]College of Computer Science, Sichuan University, China. [2]Georgia Insitute of Technology, USA.
[3]Centre for Frontier AI Research (CFAR) and Institute of High Performance Computing (IHPC), A*STAR, Singapore.
[4]National Key Laboratory of Fundamental Algorithms and Models for Engineering Simulation, China.

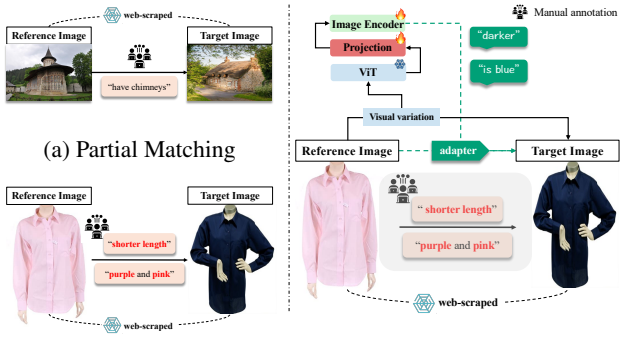
## Abstract

*Composed Image Retrieval (CIR) enables editable image search by integrating a query pair—a reference image $ref$ and a textual modification $mod$—to retrieve a target image $tar$ that reflects the intended change. While existing CIR methods have shown promising performance using well-annotated triplets $\langle ref, mod, tar \rangle$, almost all of them implicitly assume these triplets are accurately associated with each other. In practice, however, this assumption is often violated due to the limited knowledge of annotators, inevitably leading to incorrect textual modifications and resulting in a practical yet less-touched problem: noisy triplet correspondence (NTC). To tackle this challenge, we propose a **T**ask-oriented **M**odification **E**nhancement framework (TME) to learn robustly from noisy triplets, which comprises three key modules: Robust Fusion Query (RFQ), Pseudo Text Enhancement (PTE), and Task-Oriented Prompt (TOP). Specifically, to mitigate the adverse impact of noise, RFQ employs a sample selection strategy to divide the training triplets into clean and noisy sets, thus enhancing the reliability of the training data for robust learning. To further leverage the noisy data instead of discarding it, PTE unifies the triplet noise as an adapter mismatch problem, thereby adjusting $mod$ to align with $ref$ and $tar$ in the mismatched triplet. Finally, TOP replaces $ref$ in the clean set with a trainable prompt, which is then concatenated with $mod$ to form a query independent of the visual reference, aiming to mitigate visually irrelevant noise. Extensive experiments on two domain-specific datasets demonstrate the robustness and superiority of TME, particularly in noisy scenarios. Code is available at https://github.com/CharlesNeilWilliams/TME.*

## 1. Introduction

With the rise of e-commerce and the advancement of multimodal search engines, single-modality retrieval (*e.g.*, image or text) increasingly struggles to meet the diverse needs of users [5, 36]. Recently, Composed Image Retrieval (CIR) has emerged as a promising solution in this field, offering flexible image retrieval [39]. However, CIR faces challenges in bridging the cross-modal gap while unifying multimodal queries into common representations that accurately align with target images.

To address these challenges, various methods have been proposed to improve the fusion and alignment processes, primarily falling into two main categories. One straightforward approach is to project image-text pairs into common representations aligning with the target images, though



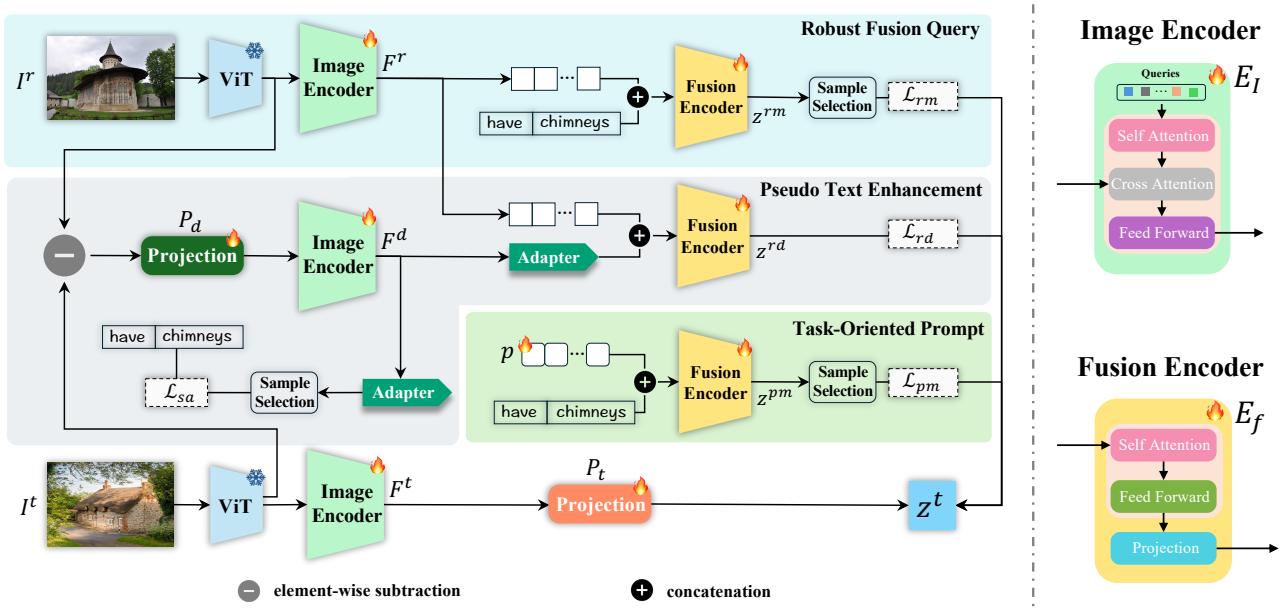

(a) Partial Matching

(b) Incorrect Matching

(c) Modification Enhancement

Figure 1. The illustration of noisy triplet correspondence (NTC) in CIR task **(left)** and our adapter enhancement strategy **(right)**. **(a)** Example from the CIRR dataset [24]: when $mod$ in a triplet only partially describes the variation from $ref$ to $mod$, a partial matching problem arises, resulting in partially matched triplets. **(b)** Example from the FashionIQ dataset [44]: where the arrows above and below represent two manually annotated modifications within each triplet. When $mod$ does not describe the features in $tar$ at all, it leads to an incorrect matching problem, resulting in completely mismatched triplets. **(c)** Our key idea is to reframe both types of noisy triplets as an adapter mismatch problem and employ explicit modeling of visual variation to capture authentic modifications, thereby facilitating the semantic alignment of the adapter at the visual level.

*The first two authors contributed equally.
†Corresponding author: Peng Hu (penghu.ml@gmail.com).




(a) Overall Pipeline (b) Encoder Architectures

Figure 2. The pipeline and encoder architectures of the proposed TME, where the encoder with the same weights are indicated by the same color. **(a) The overall pipeline of TME**: Given training triplets $\langle \boldsymbol{I}^r, \boldsymbol{m}, \boldsymbol{I}^t \rangle$, TME begins by leveraging reference-modification pairs $\langle \boldsymbol{I}^r, \boldsymbol{m} \rangle$, along with a sample selection strategy, to construct robust fusion queries. To capture real-world visual variations, we then project reference-target difference features into pseudo-text tokens (*i.e.*, adapters) along with a semantical alignment loss $\mathcal{L}_{sa}$ to align textual modifications, directly addressing noisy triplet correspondence. To alleviate the adverse influence of irrelevant references $\boldsymbol{I}^r$, we introduce a trainable prompt to replace the reference images to learn the alignment between modifications and targets. **(b) Encoder Architectures**: TME includes an image encoder $\boldsymbol{E}_I$ and a fusion encoder $\boldsymbol{E}_f$. $\boldsymbol{E}_I$ is a BLIP-2 Query Transformer used to extract visual features, while $\boldsymbol{E}_f$ comprises a transformer followed by a linear projection layer, designed to extract multimodal query features. Note that the text is tokenized and embedded via BLIP's word embedding layer $\boldsymbol{L}_t$, then concatenated after image representation $\boldsymbol{F}^r$ or prompt $\boldsymbol{p}$.

well-labeled triplets are cost-prohibitive [18, 48]. To reduce annotation requirements, recent studies introduce zero-shot methods that coverts reference images into latent pseudo-word tokens to construct multimodal pairs or triplets for self-supervised CIR training [3, 8, 16, 32, 37, 48]. However, these methods often significantly underperform compared to supervised methods using explicit triplets (*i.e.*, $< ref, mod, tar >$) [15, 17, 18, 41, 43, 45], limiting their practical value. Another compromise is to obtain economically annotated images by using web crawlers, crowdsourcing, or multimodal large language models (MLLMs) [18, 24, 44, 48], but resulting in unavoidable noise in cross-modal correspondence due to the imperfections of annotators (even humans), *i.e.*, noisy correspondence [10, 12].

Intuitively, some robust cross-modal learning methods [10, 29, 30] could be introduced to mitigate the negative effects of noisy correspondence in CIR. However, these methods mainly focus on maximizing reliable visual-textual alignment to address the noisy correspondence in image-text pairs, which lacks a robust design tailored for the noise in triplets, *i.e.*, noisy triplet correspondence (NTC), resulting in suboptimal CIR performance. Unlike noisy dual correspondence (NDC), NTC introduces two distinct challenges: 1) $mod$ only partially describes changes from $ref$ to $tar$, leading to partially matched triplets (*i.e.*, matched modification-target pairs), as shown in Figure 1(a). 2) $mod$ incorrectly describes attributes not present in $tar$, resulting in completely mismatched triplets, as illustrated in Figure 1(b). Consequently, learning with NTC is inherently more challenging and complex than learning with NDC, requiring handling both mismatched triplets as well as matched pairs.

To this end, we present a Task-oriented Modification Enhancement framework (TME) to learn with NTC for CIR as shown in Figure 2, which consists of three modules: Robust Fusion Query (RFQ), Pseudo Text Enhancement (PTE), and Task-Oriented Prompt (TOP). Specifically, our RFQ employs a sample selection strategy to divide the training triplets into clean and noisy sets while conducting a noise-robust loss on the clean set to optimize query-target alignment, thus effectively reducing the adverse impact of incorrect triplet associations. Meanwhile, PTE reframes NTC as an adapter mismatch problem, adjusting the pseudo-text tokens to align with authentic reference-target modification,

which are explicitly modeled by the visual variations between reference and target images. In brief, our PTE could comprehensively use noisy triplets to provide reliable supervision, thereby boosting effectiveness and robustness. Additionally, TOP replaces reference images with a trainable prompt to learn from matched modification-target pairs in the clean set, mitigating the negative influence of irrelevant reference images. Notably, although these three modules are designed to address NTC, they remain effective on clean triplets as visual variations are often too complex to fully capture with texts, leading to weak noisy correspondence. Our main contributions can be summarized as follows:

1. We reveal and study a novel setting in CIR—learning with noisy triplet correspondence—offering a new design perspective for existing supervised methods and potential extension to the community of learning with noisy labels.
2. To overcome the limitations of existing supervised methods in noisy scenarios, we propose a sample selection-based optimization method that strategically enhance query-target alignment among clean samples.
3. By reframing triplet noise as an adapter mismatch problem, our PTE leverages visual variation modeling to capture underlying authentic reference-target modifications, facilitating the semantic realignment at the visual level. Different from NDC methods [12, 13, 49] that only learn from the clean set, our PTE could leverage both clean and noisy sets to learn realignment, boosting the CIR performance.
4. Extensive experiments on two domain-specific datasets confirm the robustness and effectiveness of our approach in addressing noisy triplet correspondence.

## 2. Related Work

### 2.1. Composed image retrieval

Composed Image Retrieval (CIR) is an emerging task that typically leverages Vision-Language Pre-trained (VLP) models to integrate a visual reference and a textual modification into a unified query, aligning with the target image. In recent years, various methods have been proposed to optimize this supervised fine-tuning paradigm [4, 15, 18, 26, 41, 43, 45]. Among them, CLIP4Cir [4] was a pioneering approach, encoding bi-modal inputs with the CLIP model [31] and introducing a combiner network to fuse them, thereby improving the image-text matching between the unified query and target image. However, this approach does not fully exploit the latent associations within multimodal triplets. To address this, Jiang *et al*. [15] proposed a method that treats triplets as graph nodes and incorporates a hinge-based attention mechanism along with a twin-attention visual compositor to extract valuable information beyond the query-target pair. By swapping the positions of $ref$ and $tar$ and adding a learnable token, Liu *et al*. [26] and Levy *et al*. [18] further enhanced the traditional query-target paradigm with the bidirectional training scheme. While effective, these methods implicitly assume that training triplets are accurately associated. In practice, however, the existing dataset construction process (where $ref$ and $tar$ are often obtained through web scraping and $mod$ relies on manual annotations) could introduce considerable noise due to suboptimal component alignment [19, 48]. To mitigate the adverse impact of the noise, we propose a framework that enhances the traditional fine-tuning paradigm, reconstructing auxiliary adapters for noisy data to re-bridge the semantic gap at the visual level.

### 2.2. Learning with Noisy Correspondence

Unlike noisy labels [7, 20, 22, 27, 40], learning with noisy correspondence—where misalignments occur between different modalities—has received increasing attention in recent years due to its practicality and flexibility in multimodal applications [12, 28, 49]. Similar to learning from noisy labels, noisy correspondence poses significant risks of model overfitting and performance degradation. To address this issue, various methods have been proposed across different domains, such as vision-language pre-training [14], cross-modal retrieval [10, 13], person re-identification [30, 46], and clustering tasks [9, 35]. Most of these methods, however, focus only on noisy correspondences between two modalities, which is insufficient in practical applications. In CIR, noise often exists extensively in all three components due to semantic ambiguities in the modifications and errors in data grouping during the collection process. This introduces a new paradigm of noise correspondence—noisy triplet correspondence—which more closely reflects real-world scenarios, expands the scope of learning with noisy labels, and offers valuable insights into designing novel training strategies specifically for CIR tasks.

## 3. Method

In this section, we introduce Task-oriented Modification Enhancement (TME), a solution designed to address noisy triplet correspondence in CIR task, as illustrated in Figure 2. Specifically, we first elaborate on the noisy triplet correspondence setting in Section 3.1, and then provide an overview of TME in Section 3.2. Subsequently, the three key modules, namely the Robust Fusion Query, the Pseudo-Text Enhancement module, and the Task-Oriented Prompt Module are presented in Section 3.3, Section 3.4, and Section 3.5, respectively. Finally, we discuss the training and inference details of TME in Section 3.6.

### 3.1. Problem Formulation

Given a triplet set $\{\langle \boldsymbol{I}_i^r, \boldsymbol{m}_i, \boldsymbol{I}_i^t \rangle\}_{i=1}^N$, where $N$ denotes the dataset size, and $\boldsymbol{I}^r$, $\boldsymbol{m}$, and $\boldsymbol{I}^t$ represent the reference im-

age, modification, and the target image, respectively. The objective of composed image retrieval (CIR) is to retrieve the target image $\boldsymbol{I}^t$ from a large-scale image corpus $\mathcal{D}$ using a composed query $(\boldsymbol{I}^r, \boldsymbol{m})$. However, when suffering noisy triplet correspondence (NTC), misalignment in the training triplets manifests in two forms: 1) $\boldsymbol{m}$ only partially describes the intended changes from $\boldsymbol{I}^r$ to $\boldsymbol{I}^t$, leading to partially matched triplets. 2) $\boldsymbol{m}$ suggests inaccurate alterations to $\boldsymbol{I}^r$ that are not reflected in $\boldsymbol{I}^r$, resulting in completely mismatched triplets. To simulate scenarios of the above two triplets, we randomly select a subset of training triplets based on a noise ratio, $\sigma$, and split them into three equally sized, non-overlapping groups. Within each group, the components are shuffled to generate noisy triplets. Specifically, triplets with shuffled $\boldsymbol{I}^r$ form matched modification-target pairs, representing partially matched triplets, while triplets with shuffled $\boldsymbol{m}$ or $\boldsymbol{I}^t$ lead to $\boldsymbol{m}$ suggesting incorrect changes that do not align with $\boldsymbol{I}^t$, representing completely mismatched triplets.

### 3.2. Overview

To address the noisy triplet correspondence problem, existing noisy dual correspondence methods face challenges in handling the relationships between different components, making direct adaptation difficult. By reconsidering the intrinsic associations among $\boldsymbol{I}^r$, $\boldsymbol{m}$, and $\boldsymbol{I}^t$, we propose that noisy triplets correspondence can be reframed as an adapter mismatch problem. To this end, we first apply a sample selection strategy to filter out noisy samples, followed by a noise-robust loss function, $\mathcal{L}_{rm}$, to ensure reliable query-target aligning. Next, to confront the triplet noise directly, a semantic alignment loss, $\mathcal{L}_{sa}$, is used to reduce discrepancies between the adapter and the true intended changes. Meanwhile, the adapter-based query can effectively leverage noisy data, with $\mathcal{L}_{rd}$ maximizing the agreement with $\boldsymbol{I}^t$. Finally, a learnable task-oriented prompt $\boldsymbol{p}$ is introduced to replace $\boldsymbol{I}^r$, creating a reference-independent query $\boldsymbol{z}^{rm}$. Minimizing $\mathcal{L}_{pm}$ between $\boldsymbol{z}^{rm}$ and $\boldsymbol{I}^t$ achieves better modification-image alignment, mitigating overfitting caused by partially matched pairs, i.e., matched modification-target pairs. The overall loss function for TME is as follows:

$$\mathcal{L} = \mathcal{L}_{rm} + \alpha\mathcal{L}_{sa} + \beta\mathcal{L}_{rd} + \gamma\mathcal{L}_{pm}, \tag{1}$$

where $\alpha$, $\beta$, and $\gamma$ are three trade-off hyperparameters. The details of our TME will be elaborated in the following subsections.

### 3.3. Robust Fusion Query Module

To enhance the robustness of traditional approaches against the adverse impact of noisy triplet correspondences, RFQ module incorporates a sample selection strategy alongside a complementary contrastive learning framework. Inspired by [9, 12, 30], we leverage the memory effect of neural networks during early training, where losses for clean samples tend to be lower [1, 2]. Therefore, at the beginning of each epoch, RFQ calculates the infoNCE loss $\{\ell_i\}_{i=1}^N$ between each multimodal query representation $\boldsymbol{z}^{rm}$ and its target image representation $\boldsymbol{z}^t$. The distribution of these losses across all pairs is then fitted using a two-component Gaussian Mixture Model (GMM), where the component with the lower mean is assumed to correspond to clean samples. Next, we compute the posterior probabilities $\{p_i\}_{i=1}^N$ for this component to identify relatively clean samples. Following [12], we consider samples with $p_i > 0.5$ as relatively clean and use this criterion to divide the training data into a clean set $\boldsymbol{S}^c$ and a noisy set $\boldsymbol{S}^n$, enhancing data reliability and supporting robust learning for both RFQ and other modules. To further bolster robustness, the query representations $\boldsymbol{z}^{rm}$ in $\boldsymbol{S}^n$ are discarded, while the associated target images $\boldsymbol{I}^t$ are retained as negative examples for other query-target pairs. Inspired by [11], we employ a complementary contrastive learning approach between clean $\boldsymbol{z}^{rm}$ and $\boldsymbol{z}^t$:

$$\mathcal{L}_{rm} = -\frac{1}{\sum_i^B y_i} \sum_{i,j\neq i}^{B} y_i \log\left(1 - \frac{\exp(\tau(\boldsymbol{z}_i^{rm})^T \boldsymbol{z}_j^t)}{\sum_j^B \exp(\tau(\boldsymbol{z}_i^{rm})^T \boldsymbol{z}_j^t)}\right), \tag{2}$$

where $B$ denotes the batch size, $\tau$ is the temperature, and $y_i$ is a pseudo-label indicating the cleanliness of a triplet, with $y_i = 1$ for triplets in $\boldsymbol{S}^c$ and $y_i = 0$ for those in $\boldsymbol{S}^n$. Thanks to $\mathcal{L}_{rm}$, the model focuses on negative samples between clean $\boldsymbol{z}^{rm}$ and $\boldsymbol{z}^t$, pushing them apart in the shared representation space. This enhances model robustness by using numerous true negatives within each batch. Note that, given the imperfect reliability of the loss-based sample selection strategy and the relatively mild noise in partially matched triplets, $\boldsymbol{S}^c$ may still contain noisy triplets—a point further discussed in subsequent sections.

### 3.4. Pseudo Text Enhancement Module

Existing noisy dual correspondence learning methods rigidly apply query-target alignment strategies to the CIR task, which limits their ability to leverage intrinsic relationships [10, 30]. By rethinking the underlying relationships within triplets, PTE reframes noisy triplet correspondence as an adapter mismatch problem. Specifically, it models visual variation to generate pseudo text, thereby reconstructing associations and enabling learning from noisy data. To accurately reconstruct adapters, we leverage visual variation modeling to generate a pseudo text:

$$\boldsymbol{F}^d = \boldsymbol{E}_I(\boldsymbol{q}, \boldsymbol{P}_d(\boldsymbol{v}(\boldsymbol{I}^t) - \boldsymbol{v}(\boldsymbol{I}^r))), \tag{3}$$

where $\boldsymbol{v}$ is the frozen ViT model, $\boldsymbol{q}$ is the query tokens of Q-Former, and $\boldsymbol{P}_d$ is a linear projection to bridge the gap between image and image difference. To reduce the discrepancies between the pseudo text and intended changes,

we directly align the pseudo text $\boldsymbol{F}^d$ with the embeddings of tokenized modification $\boldsymbol{m}$ in $\boldsymbol{S}^c$:

$$\mathcal{L}_{sa} = \frac{1}{\sum_i^B y_i} \sum_i^B y_i ||\boldsymbol{F}_i^d - \boldsymbol{L}_t(\boldsymbol{m}_i)||^2, \tag{4}$$

where $y_i$ is the pseudo label from RFQ module, and $\boldsymbol{L}_t$ is the word embedding layer of BLIP. As $\boldsymbol{F}^d$ reflects the differences between images, we adopt complementary contrastive learning between adapter-based query representation $\boldsymbol{z}^{rd}$ and $\boldsymbol{z}^t$:

$$\mathcal{L}_{rd} = -\frac{1}{B} \sum_{i,j\neq i}^B \log \left(1 - \frac{\exp(\tau(\boldsymbol{z}_i^{rd})^T \boldsymbol{z}_j^t)}{\sum_j^B \exp(\tau(\boldsymbol{z}_i^{rd})^T \boldsymbol{z}_j^t)}\right). \tag{5}$$

For triplets in $\boldsymbol{S}^n$, $\boldsymbol{F}^d$ replaces the role of modification, enabling effective learning from noisy data. While for the partially matched triplets in $\boldsymbol{S}^c$, minimizing $\mathcal{L}_{rd}$ compensates for missing intended changes, thus enhancing performance.

### 3.5. Task-Oriented Prompt Module

To mitigate overfitting caused by partially matched triplets, TOP module replaces $\boldsymbol{I}^r$ with a learnable prompt $\boldsymbol{p}$, creating a reference-independent query. As shown in Figure 1, with limited attention spans and knowledge, human annotators may only describe the features of target images, making $\boldsymbol{m}$ a weak caption of $\boldsymbol{I}^t$, leading to a modification-target matched pair. TOP leverages this intrinsic relation by directly aligning $\boldsymbol{m}$ in $\boldsymbol{S}^c$ with $\boldsymbol{I}^t$. Specifically, TOP forms a reference-independent query $(\boldsymbol{p}, \boldsymbol{m})$ and adopts complementary contrastive learning between fused query feature $\boldsymbol{z}^{pm}$ and $\boldsymbol{z}^t$:

$$\mathcal{L}_{pm} = -\frac{1}{\sum_i^B y_i} \sum_{i,j\neq i}^B y_i \log \left(1 - \frac{\exp(\tau(\boldsymbol{z}_i^{pm})^T \boldsymbol{z}_j^t)}{\sum_j^B \exp(\tau(\boldsymbol{z}_i^{pm})^T \boldsymbol{z}_j^t)}\right). \tag{6}$$

As a task-oriented prompt, $\boldsymbol{p}$ replaces the role of $\boldsymbol{I}^r$ and acts as a blank canvas, which helps better text-image alignment in the CIR task, thereby enhancing the expressiveness of modification, mitigating overfitting cause by partially mismatched pairs in $\boldsymbol{S}^c$.

### 3.6. Training and Inference

In the training phase, to enhance stability, a warmup phase is required before sample selection, during which no data splitting occurs, and the data is treated as clean. Specifically, this phase comprises three steps: 1) Encoder warmup, to prepare $\boldsymbol{E}_I$ and $\boldsymbol{E}_f$ in RFQ module for the CIR task; 2) Projection and prompt warmup, to stabilize the training of prompts $\boldsymbol{p}$ and projection $\boldsymbol{P_d}$ that have not yet undergone BLIP pretraining; and 3) Final warmup, to integrate all modules. In the inference phase, PTE and TOP are disabled, and TME performs similarity retrieval in a large-scale image corpus $\mathcal{D}$ using composed queries $(\boldsymbol{I}^r, \boldsymbol{m})$. The complete training procedure of TME is detailed in Algorithm 1.

**Algorithm 1** TME procedure for NTC problem.

*// Training*

**Input:** Warmup epoch $w_i$, $w_p$, $w_l$, max training epoch $E$, pretrained model $M$, noisy training set $\{\langle \boldsymbol{I}_i^r, \boldsymbol{m}_i, \boldsymbol{I}_i^t \rangle\}_{i=1}^N$, loss weight $\alpha, \beta, \gamma$.
Initialize: epoch $e = 0$.
**while** $e < w_i$ **do**
  Compute loss $\mathcal{L}_{rm}$, and train the model $M$.
  $e = e + 1$.
**end while**
Freeze all parameters except projection $\boldsymbol{P}_d$ and task-oriented prompt $\boldsymbol{p}$ in $M$.
**while** $e < w_i + w_p$ **do**
  Compute loss $\mathcal{L} = \alpha\mathcal{L}_{sa} + \beta\mathcal{L}_{rd} + \gamma\mathcal{L}_{pm}$, and tuning the parameters in $\boldsymbol{P}_d$ and $\boldsymbol{p}$.
  $e = e + 1$.
**end while**
**while** $e < w_i + w_p + w_l$ **do**
  Compute loss $\mathcal{L} = \mathcal{L}_{rm} + \alpha\mathcal{L}_{sa} + \beta\mathcal{L}_{rd} + \gamma\mathcal{L}_{pm}$ and train the model $M$.
  $e = e + 1$.
**end while**
**while** $e < E$ **do**
  Compute per sample losses $\{\ell_i\}_{i=1}^N$ for each sample.
  Fit the losses into two-component Gaussian mixture distribution and assign $y_i$ for each sample.
  Compute $\mathcal{L} = \mathcal{L}_{rm} + \alpha\mathcal{L}_{sa} + \beta\mathcal{L}_{rd} + \gamma\mathcal{L}_{pm}$ and train the model $M$.
  $e = e + 1$.
**end while**

**Output:** Fine-tuned model M.

*// Inference*

**Input:** Fine-tuned model M, query pair (image $\boldsymbol{I}^r$, modification $\boldsymbol{m}$), image corpus $\boldsymbol{I}^c$.
$\boldsymbol{F}^r = \boldsymbol{E}_I(\boldsymbol{q}, \boldsymbol{I}^r)$, $\boldsymbol{z}^{rm} = \boldsymbol{E}_f([\boldsymbol{F}^r, \boldsymbol{L}_t(\boldsymbol{m})])$
$\boldsymbol{z}^c = \boldsymbol{E}_I(\boldsymbol{q}, I^c)$, $\boldsymbol{S} = (\boldsymbol{z}^{rm})^T(\boldsymbol{z}^c)$
$\{i_1, i_2, ..., i_k\} = \text{top-k}(\boldsymbol{S}_i)$

**Output:** Retrieval images $\{\boldsymbol{I}_{i_1}^c, \boldsymbol{I}_{i_2}^c, ..., \boldsymbol{I}_{i_k}^c\}$

## 4. Experiment

In this section, we conduct extensive experiments to verify the effectiveness and superiority of the proposed TME on two domain-specific datasets. Specifically, we first elaborate on the experimental setup in Section 4.1, and then compare TME with various competitive state-of-the-art methods in Section 4.2. Subsequently, we conduct ablation studies to evaluate the contribution of key components in Section 4.3. Finally, sensitivity analyses for hyperparameters are presented in Section 4.4.

Table 1. Performance of ordinary methods and robust methods (gray shading) on the **CIRR test** set. The best and second-best results are marked in **bold** and underlined, respectively.

| Noise | Methods | $R@K$ | | | | $R_{subset}@K$ | | | Avg($R@5$, $R_{subset}@1$) |
|---|---|---|---|---|---|---|---|---|---|
| | | K=1 | K=5 | K=10 | K=50 | K=1 | K=2 | K=3 | |
| 0% | TG-CIR (ACM MM'23) | 45.25 | 78.29 | 87.16 | 97.30 | 72.84 | 89.25 | 95.13 | 75.57 |
| | CASE (AAAI'24) | 48.00 | 79.11 | 87.25 | 97.57 | 75.88 | 90.58 | 96.00 | 77.50 |
| | CASE Rre-LaSCo.Ca (AAAI'24) | 49.35 | 80.02 | 88.75 | 97.47 | 76.48 | 90.37 | 95.71 | 78.25 |
| | COVR-BLIP (AAAI'24) | 49.69 | 78.60 | 86.77 | 94.31 | 75.01 | 88.12 | 93.16 | 80.81 |
| | Re-ranking (TMLR'24) | 50.55 | 81.75 | 89.78 | 97.18 | 80.04 | 91.90 | 96.58 | 80.90 |
| | SSN (AAAI'24) | 43.91 | 77.25 | 86.48 | 97.45 | 71.76 | 88.63 | 95.54 | 74.51 |
| | CALA (SIGIR'24) | 49.11 | 81.21 | 89.59 | 98.00 | 76.27 | 91.04 | 96.46 | 78.74 |
| | SPRC (ICLR'24) | 51.96 | 82.12 | 89.74 | 97.69 | 80.65 | 92.31 | 96.60 | 81.39 |
| | RCL (TPAMI'23) | 53.16 | 82.41 | 90.12 | **98.34** | 79.57 | 92.02 | 96.87 | 80.99 |
| | RDE (CVPR'24) | 51.81 | 82.02 | **90.60** | 97.93 | 78.17 | 91.90 | 96.70 | 80.10 |
| | **TME** | **53.42** | **82.99** | 90.24 | 98.15 | **81.04** | **92.58** | **96.94** | **82.01** |
| 20% | SSN (AAAI'24) | 34.02 | 65.90 | 75.78 | 91.33 | 66.92 | 85.90 | 93.45 | 66.41 |
| | CALA (SIGIR'24) | 41.33 | 72.70 | 82.84 | 94.34 | 71.66 | 88.15 | 94.94 | 72.18 |
| | SPRC (ICLR'24) | 45.90 | 75.86 | 83.52 | 93.37 | 78.10 | **91.40** | 96.05 | 76.98 |
| | RCL (TPAMI'23) | 50.43 | **81.11** | **88.82** | 96.68 | 77.52 | 90.80 | 95.71 | 79.31 |
| | RDE (CVPR'24) | 49.23 | 78.63 | 86.80 | 95.78 | 76.58 | 90.31 | 96.07 | 77.60 |
| | **TME** | **51.35** | 81.01 | 88.53 | **97.81** | **78.46** | 91.25 | **96.39** | **79.74** |
| 50% | SSN (AAAI'24) | 25.93 | 53.71 | 63.40 | 82.10 | 62.10 | 82.27 | 91.57 | 57.90 |
| | CALA (SIGIR'24) | 36.10 | 66.12 | 77.76 | 92.10 | 68.12 | 85.66 | 93.59 | 67.12 |
| | SPRC (ICLR'24) | 39.93 | 66.00 | 73.59 | 86.48 | 75.81 | 89.21 | 95.37 | 70.90 |
| | RCL (TPAMI'23) | **48.58** | 77.45 | 85.93 | 94.70 | 75.60 | 89.28 | 94.80 | 76.52 |
| | RDE (CVPR'24) | 45.98 | 75.30 | 83.73 | 94.48 | 73.98 | 88.99 | 95.13 | 74.64 |
| | **TME** | 48.48 | **78.94** | **87.28** | **96.99** | **76.48** | **90.07** | **95.83** | **77.71** |
| 80% | SSN (AAAI'24) | 20.48 | 43.98 | 54.27 | 74.80 | 56.48 | 77.20 | 89.54 | 50.23 |
| | CALA (SIGIR'24) | 31.52 | 61.49 | 72.60 | 89.86 | 64.34 | 83.52 | 92.60 | 62.92 |
| | SPRC (ICLR'24) | 29.95 | 51.25 | 58.51 | 73.86 | 70.22 | 86.05 | 93.21 | 60.74 |
| | RCL (TPAMI'23) | 44.94 | 74.43 | 82.99 | 92.31 | 71.93 | 86.84 | 92.96 | 73.18 |
| | RDE (CVPR'24) | 42.92 | 71.30 | 80.51 | 92.96 | 69.64 | 85.86 | 93.54 | 70.47 |
| | **TME** | **46.31** | **75.78** | **84.89** | **95.83** | **73.37** | **88.02** | **94.89** | **74.58** |

### 4.1. Experimental Setup

**Implementation Details** TME is implemented with Pytorch on a Tesla V100 GPU with 32 GB memory. Following the design in [45], the image encoders and fusion encoders are initialized from the BLIP-2 [21] pre-trained model with ViT-g/14 from EVA-CLIP [6]. The input image is resized to $224 \times 224$ with a padding ratio of 1.25 for uniformity [4]. We use a cosine linear learning rate decay with a peak learning rate of 1e-5 and a warmup of 1.5 epoch [33]. The hyperparameters $\alpha$ and $\gamma$ are fixed at 1.0, while $\beta$ is set to 0.2 for the CIRR dataset and 0.1 for FashionIQ.

**Datasets** We evaluate TME on two widely-used domain-specific CIR benchmarks: 1) **FashionIQ** [44], a dataset designed for fashion-conditioned image retrieval, containing $30,134$ triplets derived from a collection of $77,684$ web-scraped images. It covers three fashion categories: Dress, Shirt, and Toptee; 2) **CIRR** [23], a real-world image dataset that comprises $36,554$ triplets sourced from $21,552$ images from the popular natural language inference dataset NLVR2 [34]. CIRR captures rich object interactions, overcoming the issue of overly narrow domains. It tries to address the limitations of incomplete labeling, which causes many false negatives in datasets like FashionIQ. Additionally, CIRR includes a subset for fine-grained differentiation.

**Metric** We use R@K as the evaluation metric, which measures the percentage of queries whose target image is ranked within the top $K$ results. Consistent with prior works [4, 15, 45], we report the Recall at ranks 1, 5, 10, and 50, as well as Recall$_{subset}$ at ranks 1, 2, and 3 for CIRR. For FashionIQ, we report Recall at ranks 10 and 50 for each category. When the noise ratio $\sigma = 0$, in aligned with established publications [4, 15, 45], we report the result when the model achieves its best performance on the validation set. Early stopping based on validation performance is impractical in real-world noisy scenarios where the dataset contains noise and no clean validation set is available. Therefore, for $\sigma > 0$, we report results from the final checkpoint to evaluate model robustness and the degree of overfitting.

### 4.2. Comparison with State-of-the-arts

In this section, we evaluate the performance of our TME across varying noise ratios. For a comprehensive comparison, we compare TME with several state-of-the-art methods: 1) ordinary methods, SPRC [45], CaLa [15], SSN [47], TG-CIR [42], CASE [18], COVR-BLIP [38] and Re-ranking [25]; 2) robust methods, RCL [11] and RDE [30]. We reproduce the results of SPRC, CaLa, and SSN on two datasets with synthetic noise. Due to the robust methods designed for noisy dual correspondence, we adapt them for CIR by integrating them with our fundamental archi-

Table 2. Performance of ordinary methods and robust methods (gray shading) on the **FashionIQ validation** set. The best and second-best results are marked in **bold** and underlined, respectively.

| Noise | Methods | Dress | | Shirt | | Toptee | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | R@10 | R@50 | R@10 | R@50 | R@10 | R@50 | R@10 | R@50 | AVG. |
| 0% | TG-CIR (ACM MM'23) | 45.22 | 69.66 | 52.60 | 72.52 | 56.14 | 77.10 | 51.32 | 73.09 | 58.05 |
| | CASE (AAAI'24) | 47.44 | 69.36 | 48.48 | 70.23 | 50.18 | 72.24 | 48.79 | 70.68 | 59.74 |
| | COVR-BLIP (AAAI'24) | 44.55 | 69.03 | 48.43 | 67.42 | 52.60 | 74.31 | 48.53 | 70.25 | 59.39 |
| | Re-ranking (TMLR'24) | 48.14 | 71.43 | 50.15 | 71.25 | 55.23 | 76.80 | 51.17 | 73.13 | 62.15 |
| | SSN(AAAI'24) | 44.26 | 69.05 | 34.36 | 60.78 | 38.13 | 61.83 | 38.92 | 63.89 | 51.40 |
| | CALA (SIGIR'24) | 42.38 | 66.08 | 46.76 | 68.16 | 50.93 | 73.42 | 46.69 | 69.22 | 57.96 |
| | SPRC (ICLR'24) | 49.18 | 72.43 | 55.64 | 73.89 | **59.35** | 78.58 | 54.92 | 74.97 | 64.85 |
| | RCL(TPAMI'23) | 48.79 | **72.68** | 55.89 | 73.90 | 56.91 | 77.41 | 53.86 | 74.66 | 64.26 |
| | RDE (CVPR'24) | 47.84 | 71.89 | 54.37 | 73.55 | 56.91 | 77.21 | 53.04 | 74.22 | 63.63 |
| | **TME** | **49.73** | 71.69 | **56.43** | **74.44** | 59.31 | **78.94** | **55.15** | **75.02** | **65.09** |
| 20% | SSN (AAAI'24) | 22.61 | 45.56 | 27.87 | 48.58 | 31.82 | 55.28 | 27.43 | 49.81 | 38.62 |
| | CALA (SIGIR'24) | 29.05 | 51.36 | 35.28 | 56.23 | 36.05 | 58.24 | 33.46 | 55.28 | 44.37 |
| | SPRC (ICLR'24) | 39.81 | 62.22 | 48.58 | 66.29 | 50.48 | 70.58 | 46.29 | 66.36 | 56.33 |
| | RCL(TPAMI'23) | 47.05 | **70.65** | 53.14 | 71.74 | 55.28 | 75.62 | 51.82 | 72.67 | 62.25 |
| | RDE (CVPR'24) | 44.62 | 68.91 | 50.74 | 69.09 | 52.12 | 73.38 | 49.16 | 70.46 | 59.81 |
| | **TME** | **49.03** | 70.35 | **55.84** | **73.16** | **57.22** | **78.23** | **54.03** | **73.91** | **63.97** |
| 50% | SSN (AAAI'24) | 15.27 | 33.71 | 23.36 | 41.61 | 22.79 | 42.94 | 20.47 | 39.42 | 29.95 |
| | CALA (SIGIR'24) | 20.77 | 40.95 | 26.69 | 46.57 | 27.03 | 46.81 | 24.83 | 44.78 | 34.80 |
| | SPRC (ICLR'24) | 35.94 | 57.16 | 42.25 | 61.63 | 44.98 | 64.76 | 41.06 | 61.19 | 51.12 |
| | RCL(TPAMI'23) | 43.68 | 66.44 | 50.74 | 69.19 | 52.63 | 73.84 | 49.01 | 69.82 | 59.42 |
| | RDE (CVPR'24) | 41.30 | 64.75 | 47.06 | 66.34 | 50.13 | 70.63 | 46.16 | 67.24 | 56.70 |
| | **TME** | **46.26** | **68.27** | **53.09** | **71.88** | **55.07** | **76.59** | **51.47** | **72.25** | **61.86** |
| 80% | SSN (AAAI'24) | 11.16 | 25.24 | 16.98 | 30.72 | 17.03 | 32.64 | 15.05 | 29.53 | 22.29 |
| | CALA (SIGIR'24) | 14.28 | 30.59 | 19.73 | 35.82 | 19.48 | 36.10 | 17.83 | 34.17 | 26.00 |
| | SPRC (ICLR'24) | 28.41 | 50.77 | 36.21 | 54.37 | 35.90 | 56.96 | 33.51 | 54.03 | 43.77 |
| | RCL(TPAMI'23) | 38.82 | 60.54 | 45.44 | 64.38 | 47.42 | 68.38 | 43.89 | 64.43 | 54.16 |
| | RDE (CVPR'24) | 37.63 | 59.64 | 43.62 | 62.12 | 46.10 | 66.50 | 42.45 | 62.75 | 52.60 |
| | **TME** | **41.45** | **64.35** | **47.30** | **68.20** | **51.25** | **73.23** | **46.67** | **68.60** | **57.63** |

tecture. Table 1 and Table 2 present the evaluation results of various competitive methods on the CIRR test dataset and the FashionIQ validation dataset, respectively. These tables demonstrate that TME achieves state-of-the-art performance across all noise levels on both datasets, with the following key observations:

**Ordinary Methods vs. TME** On datasets with synthetic noise, TME shows a smaller decline and maintains high accuracy as $\sigma$ increases. Specifically, compared to the best baseline SPRC on CIRR, TME significantly improves **Avg.** as $\sigma$ increases. For example, at $\sigma = 0.2$, performance improves from 76.98 to **79.74**; at $\sigma = 0.5$, from 70.90 to **77.71**; and at $\sigma = 0.8$, from 60.74 to **74.58**. This indicates that due to their limited ability to handle overfitting, ordinary methods suffer significant performance degradation and overfitting as the noise ratio $\sigma$ increases. In contrast, TME leverages robust learning strategies and allows learning from noisy data, thereby improving performance. Note that, when $\sigma = 0$, TME still outperforms SPRC. For example, on CIRR, TME improves R@1 from 51.96 to **53.42** and R@5 from 82.12 to **82.99**, and on FashionIQ, it boots R@10 on Dress from 49.18 to **49.73**, and R@10 on Shirt from 55.64 to **56.43**. This is mainly because the CIR datasets still contain inherent noise, especially general existing partially matched triplets, thereby degrading the performance of ordinary methods.

**Robust Methods vs. TME** Compared to the robust method under noise, our approach demonstrates apparent advantages, highlighting the effectiveness and superiority of our solution for noisy triplet correspondence. As $\sigma$ increases, the performance gap between TME and the best baseline RCL widens. Specifically, TME outperforms RCL by +0.43%, +1.21%, and +1.40% at $\sigma$ of 0.2, 0.5, and 0.8, respectively on CIRR and by +1.72%, +2.44%, and +3.53% on FashionIQ. RCL only pushes the negative query target pairs apart without exploiting the intrinsic relationships in noisy triplets. In contrast, TME frames noisy triplet correspondence as an adapter-mismatch problem and leverages visual alteration modeling to reconstruct the association, thereby improving the performance.

### 4.3. Ablation Studies

In this section, we conduct ablation studies on the CIRR validation set with a noise ratio of $\sigma = 0.2$ to investigate the contribution of each component in TME, which includes the built-in enhancements in RFQ, namely GMM-filtering

Table 3. Ablation study on **CIRR validaion** dataset with $\sigma = 0.2$. The best and second-best results are marked in **bold** and underlined, respectively.

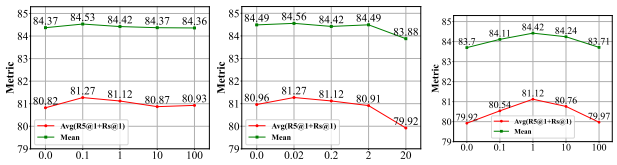
| No. | RFQ | | PTE | TOP | $R@K$ | | | | $R_{subset}@K$ | | | Avg($R@5$, $R_{subset}@1$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GMM | CCL | | | K=1 | K=5 | K=10 | K=50 | K=1 | K=2 | K=3 | |
| #1 | | | | | 50.92 | 80.58 | 87.72 | 94.96 | 76.91 | 90.50 | 95.46 | 78.75±0.23 |
| #2 | ✓ | | | | 51.61 | 81.57 | 88.92 | 97.04 | 77.82 | 91.25 | 96.18 | 79.69±0.26 |
| #3 | ✓ | ✓ | | | 53.17 | 82.64 | 89.73 | 97.49 | 78.36 | 91.49 | 96.29 | 80.50±0.26 |
| #4 | ✓ | ✓ | ✓ | | 52.99 | 83.01 | 89.91 | 97.63 | 78.37 | 91.56 | 96.35 | 80.69±0.16 |
| #5 | ✓ | ✓ | | ✓ | 53.78 | 83.02 | 90.08 | 97.60 | 79.73 | 92.25 | 96.66 | 81.38±0.23 |
| #6 | ✓ | ✓ | ✓ | ✓ | **53.84** | **83.27** | **90.18** | **97.77** | **79.92** | **92.36** | **96.68** | **81.60±0.26** |

(GMM) and complementary contrastive loss (CCL), as well as the two modules designed for modification enhancement, PTE and TOP. We evaluate various combinations of these components, as presented in Table 3. The results demonstrate that each component consistently improves performance, leading to the following observations:

- **RFQ**. Comparing #1 and # 2, the GMM-filtering significantly enhances performance, particularly improving $R@50$ from 94.96 to **97.04** and the AVG metric from 78.75 to **79.69**, demonstrating the effectiveness of GMM in identifying noisy triplets. Otherwise, the noisy triplets lead to overfitting in the baseline method (#1), and degrade overall performance. Furthermore, Comparing #2 and #3, CCL yields notable performance gains, increasing $R@1$ from 51.61 to **53.17** and the AVG metric from 79.69 to **80.50**, which highlight the robustness introduced by CCL in enhancing model performance.
- **PTE**. Comparing #3 with #4 and #5 with #6, adding PTE improves the performance of RFQ and RFQ with TOP. Specifically, adding PTE improves $R@5$ from RFQ's 82.64 to **83.01** and from 83.02 of RFQ with TOP to **83.27**. RFQ and TOP discard the query representation in the noisy set $\boldsymbol{S}^n$ entirely, leading to a suboptimal result. By incorporating both $\mathcal{L}_{sa}$ and $\mathcal{L}_{rd}$, which are a unified mechanism, TME models visual variation to pseudo-text for association reconstruction, enabling learning from triplets in $\boldsymbol{S}^n$ and improving performance.
- **TOP**. Comparing #3 and #5 and #4 with #6, adding TOP enhances performance remarkably. Specifically, it improves $R_{subset}@1$ from RFQ's 78.36 to **79.73** and from 78.37 of RFQ with PTE to **79.92**, demonstrating the effectiveness of $\mathcal{L}_{pm}$. RFQ and RFQ with PTE ignore the intrinsic relationships in partially mismatched triplets, limiting their performance. Using reference-independent query-target matching, TOP achieves better text-image alignment and mitigates overfitting caused by partially matched triplets, thereby improving performance.

### 4.4. Parametric Analysis

To study the impact of different hyperparameter settings on performance, we perform sensitivity analyses for three key hyperparameters: $\alpha$, $\beta$, and $\gamma$, which are the loss weights of $\mathcal{L}_{sa}$, $\mathcal{L}_{rd}$ and $\mathcal{L}_{pm}$. We record the mean recall rates and the AVG(R@5 + $R_{subset}@1$) for different hyperparameters. The performance on the CIRR validation dataset with $\sigma = 0.2$ is shown in Figure 3, and we can draw the following observations: 1) A relatively small $\alpha$ achieves the best performance. This suggests that large discrepancies between $\boldsymbol{F}^d$ and the intended changes exist without alignment between $\boldsymbol{F}^d$ and $\boldsymbol{m}$, while a large $\alpha$ causes $\boldsymbol{F}^d$ losing its ability to capture visual variations; 2) A relatively small $\beta$ achieves the best performance. Without $\beta$, adapter-based query-target matching is not performed, hindering learning from noisy data. Conversely, a large $\beta$ causes the model to rely on $\boldsymbol{F}^d$, which is unavailable during inference; 3) A $\gamma$ that is either too large or too small results in suboptimal performance. This indicates that a moderate $\gamma$ improves text-image alignment in reference-independent query-target matching, whereas a very large $\gamma$ causes the model to ignore the reference image, leading to performance degradation.




(a) $\alpha$ (b) $\beta$ (c) $\gamma$

Figure 3. Variation of performance with different $\alpha$, $\beta$ and $\gamma$ on the CIRR validation set with $\sigma = 0.2$.

## 5. Conclusion

In this work, we reveal and investigate a novel challenging problem of noisy triplet correspondence in CIR, which contradicts the common assumption of existing methods that triplets are accurately aligned. To address this, we propose TME, a robust method to effectively handle noisy triplet correspondence and improve performance. Extensive experiments on two domain-specific datasets comprehensively demonstrate TME's superior performance and robustness, both with and without synthetic noise.

## 6. Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2024YFB4710604; in part by NSFC under Grant 62472295, 62176171 and U21B2040; in part by Sichuan Science and Technology Planning Project under Grant 2024NSFTD0047 and 2024NSFTD0038; in part by System of Systems and Artificial Intelligence Laboratory pioneer fund grant; and in part by the Fundamental Research Funds for the Central Universities under Grant CJ202403 and CJ202303.

## References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR, 2019. 4

[2] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 4

[3] 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*, pages 15338–15347, 2023. 2

[4] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Composed image retrieval using contrastive learning and task-oriented clip-based features. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–24, 2023. 3, 6

[5] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2):1–60, 2008. 1

[6] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 6

[7] Yanglin Feng, Hongyuan Zhu, Dezhong Peng, Xi Peng, and Peng Hu. Road: Robust unsupervised domain adaptation with noisy labels. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7264–7273, 2023. 3

[8] 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*, pages 13225–13234, 2024. 2

[9] Changhao He, Hongyuan Zhu, Peng Hu, and Xi Peng. Robust variational contrastive learning for partially view-unaligned clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 4167–4176, 2024. 3, 4

[10] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9595–9610, 2023. 2, 3, 4

[11] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9595–9610, 2023. 4, 6

[12] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021. 2, 3, 4

[13] Zhenyu Huang, Peng Hu, Guocheng Niu, Xinyan Xiao, Jiancheng Lv, and Xi Peng. Learning with noisy correspondence. *International Journal of Computer Vision*, pages 1–22, 2024. 3

[14] Zhenyu Huang, Mouxing Yang, Xinyan Xiao, Peng Hu, and Xi Peng. Noise-robust vision-language pre-training with positive-negative learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–13, 2024. 3

[15] Xintong Jiang, Yaxiong Wang, Mengjian Li, Yujiao Wu, Bingwen Hu, and Xueming Qian. Cala: Complementary association learning for augmenting comoposed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2177–2187, 2024. 2, 3, 6

[16] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*, 2023. 2

[17] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1771–1779, 2021. 2

[18] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2991–2999, 2024. 2, 3, 6

[19] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2991–2999, 2024. 3

[20] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020. 3

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 6

[22] Yongxiang Li, Yang Qin, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Romo: Robust unsupervised multimodal learning with noisy pseudo labels. *IEEE Transactions on Image Processing*, 2024. 3

[23] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with

pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134, 2021. 6

[24] 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*, pages 2125–2134, 2021. 1, 2

[25] Zheyuan Liu, Weixuan Sun, Damien Teney, and Stephen Gould. Candidate set re-ranking for composed image retrieval with dual multi-modal encoder. *arXiv preprint arXiv:2305.16304*, 2023. 6

[26] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5753–5762, 2024. 3

[27] Ruitao Pu, Yuan Sun, Yang Qin, Zhenwen Ren, Xiaomin Song, Huiming Zheng, and Dezhong Peng. Robust self-paced hashing for cross-modal retrieval with noisy labels. *arXiv preprint arXiv:2501.01699*, 2025. 3

[28] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4948–4956, 2022. 3

[29] Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal active complementary learning with self-refining correspondence. *Advances in neural information processing systems*, 36:24829–24840, 2023. 2

[30] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27197–27206, 2024. 2, 3, 4, 6

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[32] 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*, pages 19305–19314, 2023. 2

[33] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019. 6

[34] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. 6

[35] Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*, 2024. 3

[36] Yuan Sun, Yang Qin, Dezhong Peng, Zhenwen Ren, Chao Yang, and Peng Hu. Dual self-paced hashing for image retrieval. *IEEE Transactions on Multimedia*, 2024. 1

[37] 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*, pages 26951–26962, 2024. 2

[38] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5270–5279, 2024. 6

[39] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448, 2019. 1

[40] Longan Wang, Yang Qin, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Robust contrastive cross-modal hashing with noisy labels. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5752–5760, 2024. 3

[41] Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. Target-guided composed image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 915–923, 2023. 2, 3

[42] Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. Target-guided composed image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 915–923, 2023. 6

[43] Haokun Wen, Xuemeng Song, Xiaolin Chen, Yinwei Wei, Liqiang Nie, and Tat-Seng Chua. Simple but effective raw-data level multimodal fusion for composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 229–239, 2024. 2, 3

[44] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021. 1, 2, 6

[45] Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, Chun-Mei Feng, et al. Sentence-level prompts benefit composed image retrieval. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 6

[46] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14308–14317, 2022. 3

[47] Xingyu Yang, Daqing Liu, Heng Zhang, Yong Luo, Chaoyue Wang, and Jing Zhang. Decompose semantic shifts for com-

posed image retrieval. *arXiv preprint arXiv:2309.09531*, 2023. 6

[48] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*, 2024. 2, 3

[49] Xu Zhang, Hao Li, and Mang Ye. Negative pre-aware for noisy cross-modal matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7341–7349, 2024. 3