Hierarchical Prompt for Task-Adaptive Composed Image Retrieval

Zeli Yan and Guosun Zeng(✉)

School of Computer Science and Technology, Tongji University, Shanghai, China
gszeng@tongji.edu.cn

Abstract. The task of composed image retrieval (CIR) is to locate target images using multimodal queries, specifically a reference image paired with modification text. The core of this task lies in fusing textual descriptions with visual content. However, existing approaches typically employ a static fusion paradigm, failing to account for the semantic heterogeneity of user queries, which encompass diverse task types (e.g., addition, replacement) and varied content. To address these limitations, we propose the Task-Adaptive Hierarchical Prompt (TAHP) framework. TAHP guides feature extraction through dynamically generated, task-specific prompts structured at three hierarchical levels: task-type, task-content, and general prompts. Furthermore, we design a Prompt Dynamic Generation Module to adaptively synthesize prompts conditioned on user queries and introduce a False Negative Correction Loss to optimize cross-modal feature fusion. We evaluated TAHP on standard FashionIQ and CIRR benchmarks and the results indicate that our method achieves superior performance, outperforming all existing CIR methods.

Keywords: Composed image retrieval, False Negative Correction, Prompt learning, Semantic alignment.

1 Introduction

With the explosive growth of internet content, users demand higher precision and flexibility in image retrieval. Traditional single-modality retrieval paradigms [1], which rely solely on text or images, often struggle to satisfy complex queries [2]. Composed Image Retrieval (CIR) overcomes this drawback by merging a reference image and a modification text into a single multimodal query. This approach empowers users to express retrieval intent in a more flexible and fine-grained manner. It thus holds significant utility in scenarios like multimodal recommendation [3] and social media content search [4].

Despite notable progress in CIR research, a core challenge remains insufficiently addressed: how to accurately comprehend the diverse task semantics implied by the modifying text. As illustrated in Figure 1, composed image retrieval encompasses various task types, including object addition, background modification, viewpoint transformation, and others. These tasks are implicitly embedded in the modification text, and a single modification text may involve multiple tasks. The semantic differences among

these distinct task types are substantial. Failing to accurately perceive and correctly handle them can harm retrieval performance. However, existing methods typically adopt a static fusion paradigm, such as feature concatenation or fixed attention mechanisms, treating all modifying instructions as homogeneous inputs and overlooking the semantic heterogeneity across modification task types and content [5]. This one-size-fits-all strategy hinders the model's discriminative capacity. As a result, the fused features fail to effectively encode fine-grained modification semantics and consequently limiting retrieval performance.

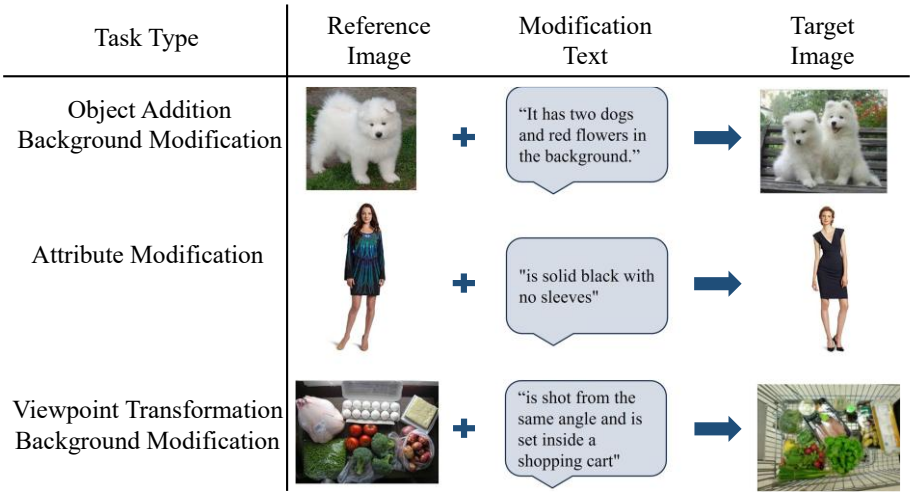

Fig. 1. The modification text implies different types of tasks. The three examples in the figure illustrate the significant semantic differences between these tasks.

To tackle these shortcomings, we propose a new architecture termed Task-Adaptive Hierarchical Prompt network (TAHP). Our core idea is to dynamically generate hierarchical prompts that match the semantics of the modification task, thereby guiding the visual encoder to adaptively extract the features most relevant to the current modification task. Specifically, we decouple prompt learning into three semantic levels: task-type prompts, which activate feature extraction pathways corresponding to the task category; task-content prompts, which encode specific modification details (e.g., color attributes, object names); and general prompts, which preserve the model's basic visual representation capability. To enable dynamic adaptation to user queries, we design a prompt generation module that adaptively synthesizes matching prompts based on the semantic characteristics of the input query. In addition, to address the label noise commonly encountered in composed retrieval, we introduce a false negative correction loss that prevents the model from erroneously pulling semantically similar false negatives closer to the query, further optimizing cross-modal feature alignment.

We conduct comprehensive experiments on multiple benchmarks including CIRR and FashionIQ to validate the performance of TAHP. Overall, our contributions are threefold:

- We identify and analyze a critical bottleneck limiting CIR performance: the lack of explicit perception and understanding of modification task types and contents.
- We propose a Task-Adaptive Hierarchical Prompt (TAHP) framework, which designs three levels of prompts (task type, task content, and general prompt) and integrates a dynamic prompt generation module to achieve adaptive feature extraction for different queries.
- We further design a false negative correction loss to refine the training objective, and thorough experiments on multiple benchmarks confirm the clear advantages of our method.

2 Related Work

2.1 Composed Image Retrieval.

Current CIR methods can be divided into cross-modal feature fusion approaches and Image-to-text mapping approaches. In typical feature fusion methods, pre-trained visual and language encoders are first used to separately derive features from the reference image and the modification text [6], after which different fusion operations are employed to generate the query representation [7]. It is also necessary to align the query and target image representations within a shared latent space. Some methods explore zero-shot CIR via Image-to-text mapping: first mapping the reference image to word embeddings, then concatenating it with the modification text to perform text-to-image retrieval, thereby converting CIR into unimodal retrieval [8, 9]. However, these approaches generally overlook the latent modification tasks inherent in CIR, lacking explicit perception and understanding of task semantics.

2.2 Prompt Learning

By training on large-scale image–caption pairs, vision-language models like CLIP [10] and ALIGN [11] learn general-purpose representations with strong generalization capabilities and are widely applied to zero-shot classification and image retrieval. Subsequently, CoOp [12] introduced prompt learning from NLP into vision by embedding continuous tokens into encoders, allowing large pre-trained models to be efficiently transferred to various downstream tasks. CoCoOp [13] further designed conditional prompts that generate prompts dynamically based on input. MaPLe [14] extended this to multimodal prompting by injecting prompts into both image and text encoders simultaneously.

Inspired by Mixture-of-Experts (MoE) architectures in large language models, recent studies explore expanding prompts into expert ensembles. pFedMoAP [15] downloads multiple prompts as non-local experts from clients and combines them with local experts to construct an MoE system, enabling more efficient personalized learning under data heterogeneity. MoPE [16] splits a unified long prompt into multiple specialized short-prompt experts and dynamically routes the most suitable expert based on multimodal prior information.

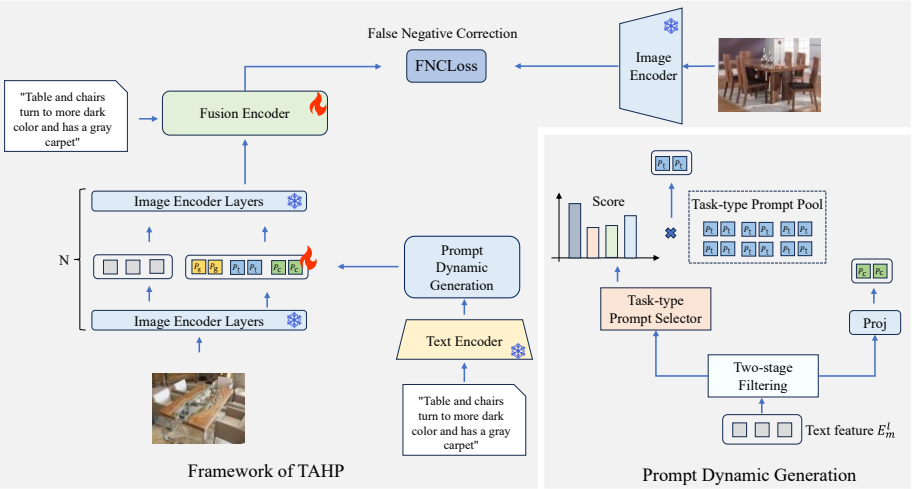

Fig. 2. The overall framework of TAHP. The prompt generation module produces task-adaptive hierarchical prompts, which are used to guide the feature extraction of the visual encoder and enhance its understanding of the specific modification task. The fusion encoder integrates image and text features and employs FNCLoss to improve robustness against noisy false negative samples.

3 Methodology

3.1 Framework Overview

Composed image retrieval leverages a multimodal query (I_r, T_m) to retrieve a target image I_t from the image collection that is most relevant to the query, where I_r denotes the reference image and T_m refers to the modification text. The objective is to train a multimodal representation network F such that the resulting query vector $Q = F(I_r, T_m)$ aligns closely with the feature of the target image I_t. Since the modification text implies diverse tasks, theoretically, a multimodal representation network should adapt its feature extraction strategy according to the specific modification task.

In this work, to address the issue that existing methods overlook the diversity of modification tasks and adopt only a static, unified feature extraction strategy, we propose the Task-Adaptive Hierarchical Prompt (TAHP) network for composed image retrieval. As illustrated in Figure 2, we first design a prompt dynamic generation module to produce task-adaptive hierarchical prompts. This module performs two-stage semantic filtering to extract features highly relevant to the task from the modification text, and subsequently generates task-type prompt and task-content prompt via weighted selection and projection mapping, respectively. These task-adaptive prompts guide how features are extracted within the image encoder, enhancing the perception and understanding of modification task semantics. Finally, we introduce a fusion encoder to integrate the extracted visual and textual features, and employ a false negative correction loss (FNCLoss) to improve robustness against noisy false negative samples.

3.2 Task-Adaptive Hierarchical Prompt Design.

The core difficulty of composed image retrieval lies in accurately perceiving the task type (e.g., attribute replacement, object addition, background transformation) and task content (e.g., "add some green plants behind") implied in the modifying text. In composed image retrieval, the reference image contains rich visual information, whereas the modifying text focuses only on a small subset of the objects. This asymmetry causes the image feature extraction process to lack focus, failing to emphasize the objects most relevant to the current modification task as specified by the modifying text.

To address this, we propose Task-Adaptive Hierarchical Prompts (TAHP), which are related to the modification task and are used to guide the feature extraction process of the visual encoder. Specifically, we decouple the prompts into three semantic levels: the task type prompt $P_{\text{type}} \in \mathbb{R}^{L_p \times D}$, which guides the visual encoder to perceive the modification task category (e.g., color modification, object addition/deletion); the task content prompt $P_{\text{content}} \in \mathbb{R}^{L_c \times D}$, which encodes the semantic representation of specific modification details (e.g., target color, added object name); and the general prompt $P_{\text{general}} \in \mathbb{R}^{L_g \times D}$, which maintains basic visual representation capability as shared prior knowledge across tasks. The general prompt is reused across all samples and encoder layers, whereas P_{type} and P_{content} are adaptively generated by the DPA module based on the specific user input. The detailed generation process of P_{type} and P_{content} will be introduced in the next subsection.

After introducing task-adaptive hierarchical prompts, the input I_i^v to the i-th layer of the visual encoder can be expressed as:

$$I_i^v = \left\{\text{CLS}, P_{\text{general}}, P_{\text{type}}, P_{\text{content}}, O_{i-1}\right\} \tag{1}$$

where [CLS] is the classification token, $O_{i-1} \in \mathbb{R}^{L_v \times D}$ is the output from the $(i-1)$-th layer of the visual encoder, L_v denotes the length of the image token sequence, and D represents the feature dimension. This design enables the encoder to focus on semantics related to the modification task at different layers, while considering both task type discrimination and task content detail understanding.

3.3 Prompt Dynamic Generation

The core objective of prompt dynamic generation is to generate prompts that guide the visual encoder based on the user-input text implying different modification tasks. Since different layers of the encoder handle semantic information at different levels, and not all content in the text is relevant to the modification task, we first perform two-stage semantic filtering on the modifying text to filter semantic information from the outputs of different layers of the text encoder, retaining only task-relevant features. Finally, in the third stage, we use the filtered features to generate dynamic prompts.

Stage 1: Layer-level feature filtering. This stage filters the output features from different layers of the text encoder. Different layers of the text encoder handle modification semantic information at different levels; both low-level and high-level information

should be considered when generating prompts. The filtering process can be expressed as:

$$f_m^i = W_f^i \odot E_m^i + E_m^i \tag{2}$$

$$E_m^* = \text{concat}(f_m^1, f_m^2, \dots, f_m^n) \cdot W_{\text{proj}} \tag{3}$$

where $E_m^i \in \mathbb{R}^{L_t \times D}$ denotes the feature output from the i-th layer of the text encoder, f_m^i is the filtered feature of the i-th layer, E_m^* is the feature obtained after layer-level filtering, L_t is the length of the text token sequence, D is the feature dimension, and $\odot$ denotes element-wise multiplication. We then concatenate all filtered features $\{f_m^i\}_{i=1}^n$ from the n layers of the text encoder along the feature dimension, and project the concatenated features back to the original space using $W_{\text{proj}} \in \mathbb{R}^{(n \cdot D) \times D}$. $\{W_f^i\}_{i=1}^n \in \mathbb{R}^{L_t \times D}$ is a set of learnable sample-agnostic filtering weights that can weight text features from different layers, ensuring that the filtered features fully account for modification semantic information at various levels.

Stage 2: Sample-level filtering. This stage filters the current specific text features, extracting the parts most relevant to the modification task. We use a multi-layer perceptron (MLP) to learn text filtering weights, extracting entities and relationships highly relevant to the specific modification content. The filtering process is:

$$W_s = \text{Softmax}\big(\text{MLP}(E_m^*)\big) \tag{4}$$

$$E_m^s = W_s^\top \cdot E_m^* \tag{5}$$

where $W_s \in \mathbb{R}^{L_t \times L_c}$ denotes the filtering weights, $E_m^s \in \mathbb{R}^{L_c \times D}$ is the final feature after sample-level filtering, and L_c denotes the length of the filtered feature sequence. This step enables the model to focus on key objects and relationships in the modification task while suppressing interference from irrelevant semantics.

Stage 3: Prompt dynamic generation. We design different dynamic generation strategies for the task type prompt and the task content prompt. For the task content prompt, we project the filtered modifying text features into the visual space via a lightweight MLP. Since the filtered modifying text features integrate semantics from different levels while retaining only the parts highly relevant to the modification task, the resulting task content prompt after projection contains sufficient details of the modification task:

$$P_{\text{content}} = \text{Proj}(E_m^s) \tag{6}$$

Furthermore, given the variety of task types implied in the modifying text (attribute modification, object addition/deletion, background replacement, etc.), we predefine K task type prompts to form a prompt pool $\mathcal{P}$, formalized as:

$$\mathcal{P} = \{P_1, P_2, \dots, P_K\} \tag{7}$$

where $P_i \in \mathbb{R}^{L_p \times D}$ is a learnable task type prompt, each representing a type of modification task.We compute the task type combination scores via an MLP:

$$s = Softmax(MLP(mean(E_m^s))) \tag{8}$$

where mean$(\cdot)$ denotes averaging over the sequence length dimension, and $s \in \mathbb{R}^K$. Finally, the task type prompt is generated via a weighted combination of the expert pool:

$$P_{\text{type}} = \sum_{i=1}^{K} s_i \cdot P_i \tag{9}$$

Since each P_i in the prompt pool represents a distinct modification task type, to enhance the representation capability of the task type prompts, we employ an orthogonal loss to encourage them to learn distinct task type representations. After average pooling and normalization, P_i yields $\bar{P}_i \in \mathbb{R}^D$, and the K vectors $\bar{P}_i$ form a matrix $\bar{P} \in \mathbb{R}^{K \times D}$. The orthogonal loss is:

$$\mathcal{L}_{\text{ortho}} = \| \bar{P}\bar{P}^{\top} - I \|_F^2 \tag{10}$$

where $I \in \mathbb{R}^{K \times K}$ is the identity matrix, and $\|\cdot\|_F^2$ denotes the squared Frobenius norm.

3.4 Composed Retrieval with False Negative Correction

After generating the task-adaptive hierarchical prompts, we leverage them to guide the visual encoder in extracting features from the reference image. The extracted visual features already encode task-aware representations:

$$E_r = F_v(\text{CLS}, P_{\text{general}}, P_{\text{type}}, P_{\text{content}}, I_r) \tag{11}$$

where F_v denotes the visual encoder and E_r is the reference image feature. To combine the image features and text features to generate the query vector, we introduce Q-Former as a fusion encoder, fusing the visual features E_r with the text features E_m to produce the query features.

Although task hierarchical prompts can incorporate task-relevant information into the feature extraction process, enhancing the model's perception and understanding of modification tasks, challenges remain. In composed image retrieval, it is typically assumed that only annotated positive pairs are matched, while all negative samples are mismatched. However, in reality, many negative samples are highly relevant to the query. Forcing the model to repel such false negatives can harm its representation capability. To this end, we improve upon the contrastive loss and propose a false negative correction loss that, while pulling query features closer to target image features, prevents erroneously increasing the distance between the query and false negatives. The false negative correction loss is formulated as:

$$\mathcal{L}_{FNC} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log P_{ii}^{q} - \lambda \cdot \frac{1}{|\mathcal{B}|^2} \sum_{i=1}^{|\mathcal{B}|} \sum_{\substack{j=1 \\ j\neq i}}^{|\mathcal{B}|} \max\left(P_{ij}^{t} - \delta, 0\right) \cdot \log P_{ij}^{q} \tag{12}$$

$$P_{ij}^{t} = \frac{\exp\{s(E_{t_i}, E_{t_j})/\tau\}}{\sum_{k=1}^{|\mathcal{B}|} \exp\{s(E_{t_i}, E_{t_k})/\tau\}} \tag{13}$$

$$P_{ij}^{q} = \frac{\exp\{s(E_{q_i}, E_{t_j})/\tau\}}{\sum_{k=1}^{|\mathcal{B}|} \exp\{s(E_{q_i}, E_{t_k})/\tau\}} \tag{14}$$

Where B is the batch size, τ is the temperature factor, $s(\cdot)$ refers to the cosine similarity function. The first term of this loss is the standard contrastive loss, and the second term is the false negative correction term. Here, P_{ij}^{q} denotes the similarity weight between the i-th query and the j-th candidate image within a batch, and P_{ij}^{t} stands for the similarity weight between the i-th and j-th candidate images. We use P_{ij}^{t} as a supervisory signal for P_{ij}^{q}. When P_{ij}^{t} exceeds the threshold δ, it means that the i-th and j-th candidate images are highly similar. In this case, the j-th image can be considered a false negative sample, and we should not increase its semantic distance from the query.

4 Experiment

4.1 Experimental Setup

Datasets and Metrics. We assess our approach on two publicly available composed image retrieval datasets: FashionIQ and CIRR. As the main evaluation metric, we use Recall@K, which quantifies the fraction of queries for which the correct target image is ranked within the top K retrieved results. For FashionIQ, we compute Recall@K (K=10, 50) for each category (Dress, Shirt, Toptee), as well as their average. For CIRR, we report Recall@K (K=1, 5, 10, 50), Recall_subset@K (K=1, 2, 3), which measures retrieval performance on specified subsets, and their averages.

Implementation Details. Our proposed method employs the pre-trained visual and text encoders of BLIP-2 as the backbone network. Training is performed with the AdamW optimizer and a weight decay of 0.05. Input images are first padded with a ratio of 1.25 and then uniformly reshaped to 224×224 pixels. We apply a cosine annealing schedule to decay the learning rate. The initial learning rate is set to 2e-6 for the FashionIQ dataset and 1e-5 for the CIRR dataset. Every learnable prompt is composed of 8 tokens, and the overall prompt pool contains 32 such prompts.

4.2 Performance Comparision

We evaluated our proposed method on the Fashion-IQ and CIRR datasets, conducting comprehensive comparisons with recent top-performing approaches. Table 1 presents a detailed summary of the Fashion-IQ dataset for all methods compared. As evidenced by the results, our proposed method (TAHP) outperforms every competitor on recall rates metrics across the three evaluated categories.

Table 1. Results on the FashionIQ dataset. The Recall@K (R@K) metric is used to measure all methods, where Avg. refers to the average results over the three fashion categories. The best and second-best results are marked in bold and underlined, respectively.

	Dress		Shirt		Topee		Average		
Method	R@10	R@10	R@10	R@50	R@10	R@50	R@10	R@50	Avg.
CIRPLANT [18]	17.45	40.41	17.53	38.81	61.64	45.38	18.87	41.53	30.20
JVSM [17]	10.70	25.90	12.00	27.10	13.00	26.90	11.90	26.60	19.26
CoSMo [20]	25.64	50.30	24.90	49.18	29.21	57.46	26.58	52.31	39.45
TRACE [19]	22.70	44.91	20.80	40.80	24.22	49.80	22.57	46.19	34.00
ARTEMIS [21]	27.16	52.40	21.78	43.64	29.20	53.83	26.05	50.29	38.04
DCNet [22]	28.95	56.07	23.95	47.30	30.44	58.29	27.78	53.89	40.84
CLIP4CIR [25]	33.81	59.40	39.99	60.45	41.41	65.37	38.32	61.74	50.03
FashionVLP [23]	32.42	60.29	31.89	58.44	38.51	68.79	34.27	62.51	48.39
LF-BLIP [24]	25.31	44.05	25.39	43.57	26.54	44.48	25.75	43.98	34.88
BLIP4CIR+Bi [26]	42.09	67.33	41.76	64.28	46.61	70.32	43.49	67.31	55.04
TG-CIR [29]	45.22	67.38	52.60	72.52	56.14	77.10	51.32	73.09	58.05
DRA [27]	33.98	60.67	40.74	61.93	42.09	66.97	38.93	63.19	51.06
Re-ranking [28]	48.14	71.43	50.15	71.25	55.23	76.80	51.17	73.13	62.15
SPRC [30]	<u>49.18</u>	<u>72.43</u>	<u>55.64</u>	<u>73.89</u>	**59.35**	<u>78.58</u>	**54.92**	<u>74.97</u>	<u>64.85</u>
Ours	**50.03**	**73.35**	**55.86**	**75.00**	<u>58.60</u>	**80.02**	<u>54.69</u>	**76.04**	**65.36**

Compared to the second-best performing method, our approach boosts performance on the R@50 metric from 74.97% to 76.04% and increases the average recall from 64.85% to 65.36%. Furthermore, our method demonstrates consistent improvements across the "Dress" and "Shirt" categories.

We further evaluated our method on the CIRR test set. Characterized by a more diverse image domain and higher-quality modification text annotations, this dataset offers a more comprehensive assessment of retrieval performance. Compared to other state-of-the-art methods, our approach yields significant improvements, specifically raising the average recall to 81.69%.

These experimental results substantiate that our proposed composed image retrieval method effectively captures the implicit modification task information within user inputs. By leveraging multi-level prompt guidance to efficiently extract query features, our approach significantly enhances overall composed image retrieval performance.

4.3 Ablation Studies

To evaluate the effectiveness of our proposed components, we conducted ablation experiments on the FashionIQ dataset. We removed the TAHP and FNCLoss modules, retaining only the frozen image and text encoders along with the fusion module, and used this configuration as the baseline.

Table 2. Results on CIRR dataset. all methods are evaluated using the Recall@K and Recalls@K metrics, where Avg. represents the average of Recall@5 and Recalls@1. The best and second-best results are marked in bold and underlined, respectively.

	Recall@K				Recalls@K			
Method	K=1	K=5	K=10	K=50	K=1	K=2	K=3	Avg.
MAAF [32]	10.31	33.03	48.30	80.06	21.05	41.81	61.60	27.04
TIRG [31]	14.61	48.37	64.08	90.03	22.67	44.97	65.14	35.52
ARTEMIS [21]	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
CIRPLANT [18]	19.55	52.55	68.39	92.38	39.20	63.03	79.49	45.88
LF-BLIP [24]	20.89	48.07	61.16	83.71	50.22	73.16	86.82	60.58
CLIP4CIR [25]	38.53	69.98	81.86	95.93	68.19	85.64	94.17	69.09
DRA [27]	39.93	72.07	83.83	96.43	71.04	87.74	94.72	71.55
BLIP4CIR+Bi [26]	40.15	73.08	83.88	96.27	72.10	88.27	95.93	72.59
CASE [33]	48.00	79.11	87.25	97.57	75.88	90.58	96.00	77.50
TG-CIR [29]	45.25	78.29	87.16	97.30	72.84	89.25	95.13	75.57
Re-ranking [28]	50.55	81.75	<u>89.78</u>	97.18	80.04	91.90	96.58	80.90
CaLa [34]	49.11	81.21	89.59	98.00	76.27	91.04	96.46	78.74
SPRC [30]	<u>51.96</u>	<u>82.12</u>	89.74	<u>97.69</u>	**80.65**	<u>92.31</u>	<u>96.60</u>	<u>81.39</u>
Ours	**52.82**	**82.87**	**90.84**	**98.04**	<u>80.51</u>	**92.36**	**97.01**	**81.69**

Effectiveness of TAHP. The TAHP module is designed to generate task-adaptive hierarchical prompts that guide the visual encoder to perceive the semantics of modification tasks. After adding the TAHP module, the average R@10 and R@50 retrieval performance on the Fashion-IQ dataset both improved. These strong results demonstrate that the TAHP module can effectively enhance the model's ability to perceive diverse modification tasks and improve the discriminability of query features.

Table 3. Ablation studies on FashionIQ dataset.

	Module			Fashion-IQ		
Index	Baseline	TAHP	FNCLoss	R@10	R@50	avg
1	√			54.31	75.57	64.94
2	√	√		54.44	75.98	65.21
3	√		√	54.49	75.81	65.15
4	√	√	√	**54.69**	**76.04**	**65.37**

Effectiveness of FNCLoss. We design the FNCLoss to correct false negatives, preventing the model from erroneously increasing the distance between queries and false negative samples. After introducing the FNCLoss, the average R@10 and R@50 on the Fashion-IQ dataset increased to 54.49% and 75.81%, respectively. Experimental results

show that our proposed FNCLoss can effectively improve the model's robustness to noisy data and enhance retrieval ranking quality.

Furthermore, when TAHP and FNCLoss are used together, the average R@10 on the Fashion-IQ dataset rises to 54.69%, and the average R@50 rises to 76.04%, indicating that TAHP and FNCLoss can mutually reinforce each other and jointly improve retrieval performance.

5 Conclusion

We present a task-adaptive hierarchical prompting framework for composed image retrieval, aiming to address two key challenges: accurately perceiving the task type and task content implied in the modification text, and handling false negative noise, i.e., samples that are highly relevant to the query but annotated as negatives. Our proposed method consists of two main modules: the Task-Adaptive Hierarchical Prompt module (TAHP) and the False Negative Correction Loss (FNCLoss). TAHP dynamically generates task-type and task-content prompts to guide the visual encoder to focus on image regions relevant to the modification task, thereby enhancing the model's semantic perception of diverse modification tasks and improving the discriminability of query features and retrieval accuracy. Furthermore, FNCLoss uses the similarity among candidate images as a supervisory signal to prevent erroneously increasing the distance between queries and false negatives, thereby correcting the optimization direction for false negatives and enhancing the model's robustness to noisy data. Experiments on the CIRR and FashionIQ datasets demonstrate that our method achieves competitive results, indicating that the proposed task-adaptive hierarchical prompts and false negative correction loss can work synergistically to significantly improve composed image retrieval performance.

References

1. Qu, L., Liu, M., Cao, D., Nie, L., Tian, Q.: Context-aware multi-view summarization network for image-text matching. In: Proceedings of the International ACM Conference on Multimedia, pp. 1047–1055. ACM (2020)
2. Ma, X., Yang, M., Li, Y., Hu, P., Lv, J., Peng, X.: Cross-modal retrieval with noisy correspondence via consistency refining and mining. IEEE Transactions on Image Processing 33, 2587–2598 (2024)
3. Li, Z., Liu, F., Wei, Y., Cheng, Z., Nie, L., Kankanhalli, M.: Attribute-driven disentangled representation learning for multimodal recommendation. In: Proceedings of the ACM International Conference on Multimedia, pp. 9660–9669. ACM (2024)
4. Chen, Y., Zheng, Z., Ji, W., Qu, L., Chua, T.-S.: Composed image retrieval with text feedback via multi-grained uncertainty regularization. In: International Conference on Learning Representations, vol. 2024, pp. 54915–54927. (2024)
5. Wen, H., Song, X., Yin, J., Wu, J., Guan, W., Nie, L.: Self-training boosted multi-factor matching network for composed image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 46(5), 3665–3678 (2023)

6. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2125–2134. IEEE/CVF (2021)
7. Wen, H., Zhang, X., Song, X., Wei, Y., Nie, L.: Target-guided composed image retrieval. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 915–923. ACM (2023)
8. Saito, K., Sohn, K., Zhang, X., Li, C.-L., Lee, C.-Y., Saenko, K., Pfister, T.: Pic2word: mapping pictures to words for zero-shot composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19305–19314. IEEE/CVF (2023)
9. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-shot composed image retrieval with textual inversion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15338–15347. IEEE/CVF (2023)
10. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML, pp. 8748–8763. PMLR (2021)
11. Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning, pp. 4904–4916. PMLR (2021)
12. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision 130(9), 2337–2348 (2022)
13. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16816–16825. IEEE/CVF (2022)
14. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19113–19122. IEEE/CVF (2023)
15. Luo, J., Chen, C., Wu, S.: Mixture of experts made personalized: federated prompt learning for vision-language models. arXiv preprint arXiv:2410.10114 (2024)
16. Jiang, R., Liu, L., Chen, C.: MoPE: mixture of prompt experts for parameter-efficient and scalable multimodal fusion. arXiv preprint arXiv:2403.10568 (2024)
17. Chen, Y., Bazzani, L.: Learning joint visual semantic matching embeddings for language-guided retrieval. In: European Conference on Computer Vision. Springer, Cham (2020)
18. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2125–2134. IEEE/CVF (2021)
19. Jandial, S., Chopra, A., Badjatiya, P., Chawla, P., Sarkar, M., Krishnamurthy, B.: Trace: Transform aggregate and compose visiolinguistic representations for image search with text feedback. arXiv preprint arXiv:2009.01485 (2020)
20. Lee, S., Kim, D., Han, B.: Cosmo: Content-style modulation for image retrieval with text feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 802–812. IEEE/CVF (2021)
21. Delmas, G., et al.: Artemis: attention-based retrieval with text-explicit matching and implicit similarity. arXiv preprint arXiv:2203.08101 (2022)
22. Kim, J., Yu, Y., Kim, H., Kim, G.: Dual compositional learning in interactive image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 1771–1779. AAAI (2021)

23. Goenka, S., Zheng, Z., Jaiswal, A., Chada, R., Wu, Y., Hedau, V., Natarajan, P.: Fashionvlp: Vision language transformer for fashion retrieval with feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14105–14115. IEEE/CVF (2022)
24. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Effective conditioned and composed image retrieval combining clip-based features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21466–21474. IEEE/CVF (2022)
25. Baldrati, A., Bertini, M., Uricchio, T., et al.: 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, pp. 4959–4968. IEEE/CVF (2022)
26. Liu, Z., Sun, W., Hong, Y., et al.: Bi-directional training for composed image retrieval via text prompt learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5753–5762. IEEE/CVF (2024)
27. Jiang, X., Wang, Y., Wu, Y., et al.: Dual relation alignment for composed image retrieval. arXiv preprint arXiv:2309.02169 (2023)
28. Liu, Z., Sun, W., Teney, D., Gould, S.: Candidate set re-ranking for composed image retrieval with dual multi-modal encoder. arXiv preprint arXiv:2305.16304 (2023)
29. Wen, H., Zhang, X., Song, X., et al.: Target-guided composed image retrieval. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 915–923. ACM (2023).
30. Bai, Y., Xu, X., Liu, Y., et al.: Sentence-level prompts benefit composed image retrieval. arXiv preprint arXiv:2310.05473 (2023)
31. Vo, N., Jiang, L., Sun, C., et al.: Composing text and image for image retrieval - an empirical odyssey. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6439–6448. IEEE/CVF (2019)
32. Dodds, E., Culpepper, J., Herdade, S., et al.: Modality-agnostic attention fusion for visual search with text feedback. arXiv preprint arXiv:2007.00145 (2020)
33. Levy, M., Ben-Ari, R., Darshan, N., et al.: Data roaming and quality assessment for composed image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38(4), pp. 2991–2999. AAAI (2024)
34. Jiang, X., Wang, Y., Li, M., et al.: Cala: complementary association learning for augmenting composed image retrieval. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2177–2187. ACM (2024)