

CCIN: Compositional Conflict Identification and Neutralization for Composed Image Retrieval

Likai Tian^{1,*} Jian Zhao^{2,3,*} Zechao Hu¹ Zhengwei Yang¹ Hao Li¹ Lei Jin⁴
 Zheng Wang^{1,†} Xuelong Li²

¹National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University ²Institute of Artificial Intelligence (TeleAI), China Telecom

³School of Artificial Intelligence, Optics and Electronics (iOPEN),

Northwestern Polytechnical University ⁴Beijing University of Posts and Telecommunications

{tianlikai, 00034599, yzw_aim, sc.lihao}@whu.edu.cn, zhaojian90@u.nus.edu,
 jinlei@bupt.edu.cn, wangzwhu@whu.edu.cn, li@nwpu.edu.cn

Abstract

Composed Image Retrieval (CIR) is a multi-modal task that seeks to retrieve target images by harmonizing a reference image with a modified instruction. A key challenge in CIR lies in compositional conflicts between the reference image (e.g., blue, long sleeve) and the modified instruction (e.g., grey, short sleeve). Previous works attempt to mitigate such conflicts through feature-level manipulation, commonly employing learnable masks to obscure conflicting features within the reference image. However, the inherent complexity of feature spaces poses significant challenges in precise conflict neutralization, thereby leading to uncontrollable results. To this end, this paper proposes the **Compositional Conflict Identification and Neutralization (CCIN)** framework, which sequentially identifies and neutralizes compositional conflicts for effective CIR. Specifically, CCIN comprises two core modules: 1) *Compositional Conflict Identification module*, which utilizes LLM-based analysis to identify specific conflicting attributes, and 2) *Compositional Conflict Neutralization module*, which first generates a kept instruction to preserve non-conflicting attributes, then neutralizes conflicts under collaborative guidance of both the kept and modified instructions. Extensive experiments demonstrate the superiority of CCIN over the state-of-the-arts. Code repository: <https://github.com/LikaiTian/CCIN>.

1. Introduction

Composed Image Retrieval (CIR) offers a holistic approach by integrating both the reference image and the

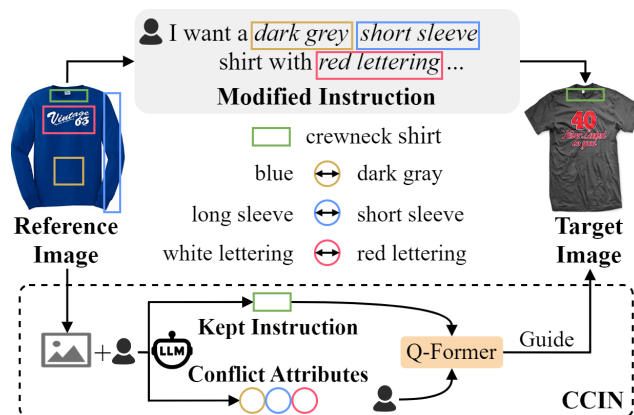


Figure 1. An example of the multi-modal query that contains compositional conflicts (e.g., shirt color, sleeve length, and letter color), which may misrepresent the user intention and result in ambiguous retrieval results. To address this issue, we propose the CCIN framework to sequentially identify and neutralize such conflicts, thereby contributing to positive retrieval.

modified instruction into a composed query, which differs from traditional image retrieval tasks [26, 35, 43, 53, 57] that proceed with single-modal queries. By capitalizing on the strengths of both modalities, CIR significantly enhances search flexibility in practical applications [49]. Several studies [2–4, 10, 25, 28] have investigated CIR by employing various multi-modal fusion strategies to enhance query representations. However, these fusion approaches often struggle to address inherent discrepancies in CIR scenarios, where the visual modality serves as the foundational reference while the textual modality conveys specific modifications. In line with TG-CIR [45], we refer to this discrepancy as compositional conflict (as shown in the middle area of Figure 1). Integrating the multi-modal

*Equal contribution

†Corresponding author

query without resolving these conflicts poses a significant challenge in accurately capturing the user intention, thereby diminishing the retrieval effectiveness.

Previous works [45, 51] have sought to mitigate compositional conflicts through feature-level manipulation, commonly employing learnable masks to obscure conflicting features within the reference image. However, these methods lack explicit awareness of the specific conflicting attributes, relying instead on the coarse-grained suppression strategy that broadly attenuates potentially conflicting features. Additionally, the inherent complexity [13, 31] of feature spaces poses substantial challenges for achieving controllable conflict resolution. Given these limitations, effectively resolving compositional conflicts requires a shift from broad suppression to a more targeted approach, focusing on two key aspects: *conflict identification* (i.e., precisely determining the specific attributes in conflict) and *conflict neutralization* (i.e., developing a refined mechanism to resolve these conflicts in a controllable manner).

In this work, we propose the **Compositional Conflict Identification and Neutralization (CCIN)** framework, which incorporates two main components: (1) a **Compositional Conflict Identification (CCI)** module that adaptively analyzes underlying associations within queries to accurately identify compositional conflicts, and (2) a dual-instruction **Compositional Conflict Neutralization (CCN)** module, designed to neutralize such conflicts in a more deterministic and controllable manner. Specifically, the CCI module initiates conflict resolution by converting the reference image into textual captions. These captions are then integrated with the modified instruction to enable conflict identification through Large Language Model (LLM) analysis. In the subsequent CCN module, we first employ the LLM to neutralize compositional conflicts by selectively filtering out conflicting attributes within the reference image, yielding a refined kept instruction. This kept instruction, in conjunction with the modified instruction, further supports the conflict neutralization process: the kept instruction guides the non-conflicting visual feature extraction, while the modified instruction adjusts these features to adhere to the specified modifications. This dual-instruction mechanism ensures a more coherent alignment with the user intention.

The CCIN is delicately designed to identify and neutralize compositional conflicts within multi-modal queries for effective retrieval. Through comprehensive analysis of conflict sources and finer-grained control over conflict neutralization, this approach functions as a more informative tool and is less susceptible to misrepresenting the user intention. CCIN adeptly converts compositional conflict identification and neutralization into improved effectiveness in CIR, thereby enhancing the retrieval performance. We hope this work could bring fundamental insights into related fields.

The contributions of this paper are threefold:

- We highlight the presence of compositional conflicts in the Composed Image Retrieval (CIR) task and revisit CIR from a fresh perspective aimed at identifying and neutralizing such conflicts in a controllable manner.
- We propose a **Compositional Conflict Identification and Neutralization (CCIN)** framework, which sequentially identifies and neutralizes compositional conflicts for effective retrieval.
- Extensive experiments on three standard CIR datasets, FashionIQ [48], CIRR [27] and Shoes [16], demonstrate the superiority of the proposed CCIN.

2. Related Work

2.1. Composed Image Retrieval

Various methodologies have been developed for Composed Image Retrieval (CIR). Late-fusion approaches [3, 4, 10, 25, 28] utilize a range of feature fusion mechanisms [33, 37, 40] to integrate multi-modal features extracted separately from the reference image and the modified instruction. In contrast, early-fusion approaches [46] emphasize the integration of raw data, seeking to harness the complementary capabilities in the original inputs. Additionally, visual prompting approaches endeavor to convert the reference image into structured representations (e.g., pseudo-word [32] or sentence-level prompt [2]), to improve multi-modal comprehension. Despite the complexity and sophistication of their fusion strategies, they are inevitably limited by compositional conflicts within queries, which may mislead the user intention and result in ambiguous predictions when straightforwardly merging the visual and textual modalities. To cope with the above challenge in CIR, we propose the Compositional Conflict Identification and Neutralization (CCIN) framework. This approach aims to effectively identify and neutralize such conflicts in a controllable manner, thereby enhancing the retrieval effectiveness.

2.2. Vision-Language Pre-training

Vision-Language Pre-training (VLP) models [22, 23, 30] have emerged as powerful tools in multi-modal learning, demonstrating remarkable performance across various downstream tasks [42, 50]. In CIR domain, VLP models act as robust encoders [2, 3, 19, 28], facilitating the extraction of comprehensive multi-modal representations. Notably, the advanced InstructBLIP [11], built upon the lightweight Q-Former [24], proposes a novel mechanism for extracting instruction-aware visual features from the image encoder by incorporating text instructions as additional input to the Q-Former. Expanding on this single-instruction approach, we propose a dual-instruction Q-Former-based Compositional Conflict Neutralization (CCN) module, designed to extract informative visual features guided by dual instructions (the kept instruction and the modified instruction). This

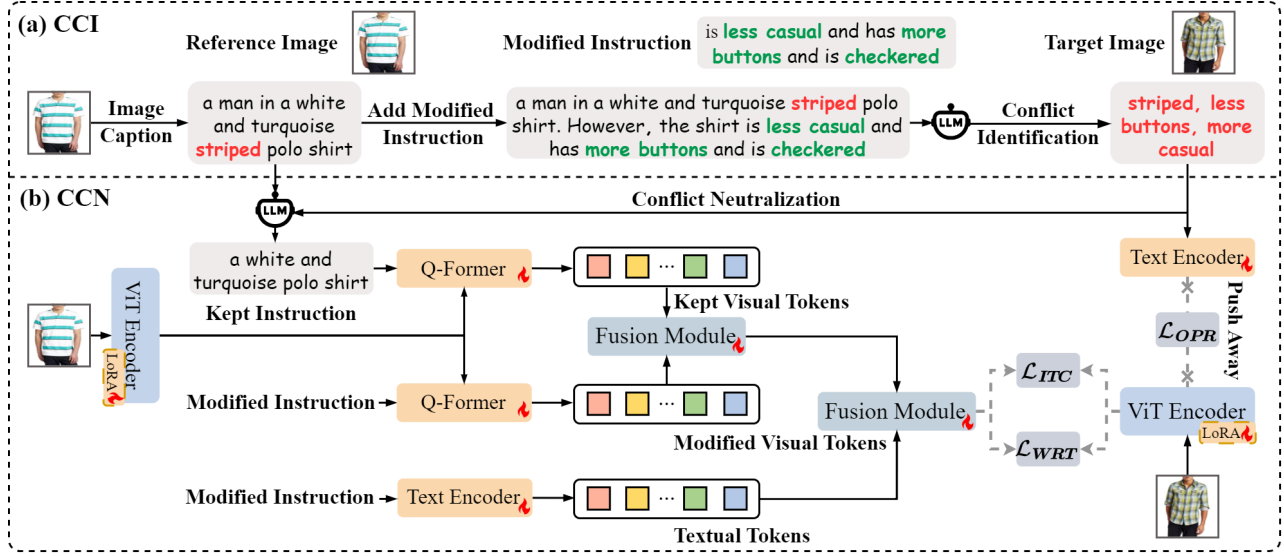


Figure 2. Structure overview of the proposed method. It consists of two main components, a Compositional Conflict Identification (CCI) module and a dual-instruction Compositional Conflict Neutralization (CCN) module. Such structure is to identify and neutralize conflicting information in a controllable manner, thereby enhancing the effectiveness of CIR process.

dual-instruction module enables finer-grained control over conflict neutralization, thereby ensuring a more seamless alignment with the user search intention.

2.3. Conflict Neutralization

Multi-modal scenarios often contain inconsistent or conflicting information, which poses challenges in learning high-quality multi-modal representations. As a solution, multi-view learning task [47, 50, 55] has spurred the development of various conflict resolution strategies (e.g., conflictive instance elimination and conflictive instance decision) aimed at filtering out irrelevant information to improve the fidelity of multi-view relationship modeling. In the context of the CIR task, previous studies [45, 51] have also explored this issue through employing learnable masks to obscure conflicting features within the reference image. However, these approaches generally operate in the feature space, which restricts their ability to precisely identify and neutralize specific conflicting attributes. Moreover, the intrinsic complexity of feature spaces poses significant challenges in achieving controllable results. To address these limitations, we propose the CCIN framework, which leverages advanced LLM to precisely identify conflicting attributes, coupled with a dual-instruction module to facilitate more controllable conflict neutralization.

3. Proposed Method

3.1. Overview

Problem Definition. The task of CIR aims to retrieve target image I_t from an image corpus through a multi-modal

Prompt:

You are an intelligent chatbot designed to **extract conflicting attributes** from the provided text, which describes the feature changes from a reference image to a target image. Your task is to identify and list the attributes in the reference image that conflict with those in the target image

Input:

The input text for this task is: <combined_text>.

Prompt:

You are an intelligent chatbot designed to extract non-conflicting attributes within the reference image caption while **excluding those specified in the conflicting keywords**

Input:

The input text <ref_caption> is the reference image caption. The keywords <conflicting_attributes> are the conflicting attributes we don't want.

Figure 3. Illustration of the customized prompts, respectively.

query $\{I_r, T_{\text{mod}}\}$, where I_r donates the reference image and T_{mod} represents the modified instruction.

Overall Architecture. In this work, we concentrate on compositional conflicts stemming from the juxtaposition of the reference image I_r and the modified instruction T_{mod} . Such conflicts inevitably lead to ambiguities when coarsely integrating multi-modal features, thereby diminishing the CIR effectiveness. To tackle the above challenge, we propose a novel method for CIR, termed CCIN, as illustrated in Figure 2. At the outset, we devise a Compositional Conflict Identification (CCI) module, which merges the reference image caption T_{ref} with the modified instruction T_{mod} , enabling LLM to identify specific conflicting attributes. By neutralizing the conflicting attributes from

T_{ref} , we generate a refined kept instruction T_{kep} that retains only the non-conflicting attributes in the reference image. This operation further facilitates the development of a dual-instruction Q-Former-based Compositional Conflict Neutralization (CCN) module, which leverages the modified instruction T_{mod} in parallel with the kept instruction T_{kep} to extract instruction-aware visual features that precisely align with the user intention. Subsequently, an adaptive fusion module is introduced to integrate the conflict-neutralized visual features with textual features.

3.2. Compositional Conflict Identification

To enable precise awareness of conflicting attributes, we propose the Compositional Conflict Identification (CCI) module, as depicted in the upper dashed box of Figure 2. The CCI begins by applying an image captioning model to generate detailed captions for the reference image I_r . Subsequently, we amalgamate the reference image caption T_{ref} with the modified instruction T_{mod} , utilizing the conjunction ‘‘However’’ to form a combined textual summary. Guided by the customized prompt P_{con} (as depicted at the top of Figure 3), the LLM meticulously analyzes the combined text to scrutinize the underlying conflicts T_{con} :

$$T_{\text{con}} = P_{\text{con}} \rightarrow \text{LLM}(T_{\text{ref}} + \text{However} + T_{\text{mod}}). \quad (1)$$

3.3. Compositional Conflict Neutralization

To enable fine-grained control over conflict resolution, we propose the dual-instruction Q-Former-based Compositional Conflict Neutralization (CCN) module. Inspired by the effective InstructBLIP [11], the CCN module is designed to leverage dual instructions for precisely extracting instruction-aware visual features in alignment with the user intention, as depicted in the bottom dashed box of Figure 2. Building upon the conflict attributes identified through CCI, we initially devise an auxiliary prompt P_{kep} (as depicted at the bottom of Figure 3) to exclude conflicting attributes T_{con} from the reference image caption T_{ref} and formulate a kept instruction T_{kep} :

$$T_{\text{kep}} = P_{\text{kep}} \rightarrow \text{LLM}(T_{\text{ref}}, T_{\text{con}}). \quad (2)$$

Subsequently, the kept instruction T_{kep} , in conjunction with the modified instruction T_{mod} , leverages the lightweight Q-Former $\mathcal{F}_Q$ to interact with the reference image I_r , adeptly facilitating this dual-guided process:

$$\mathbf{f}_{\text{kep}} = T_{\text{kep}} \rightarrow \mathcal{F}_Q(I_r), \mathbf{f}_{\text{mod}} = T_{\text{mod}} \rightarrow \mathcal{F}_Q(I_r), \quad (3)$$

where $\mathbf{f}_{\text{kep}}$ is extracted under the guidance of T_{kep} , capturing the non-conflicting visual feature within the reference image. In parallel, $\mathbf{f}_{\text{mod}}$ is extracted under the guidance of T_{mod} , adjusting the visual features to align with the specified modifications. Additionally, to bolster the proficiency in handling high-level semantics, we apply LoRA fine-tuning to pivotal layers of the ViT backbone.

3.4. Adaptive Fusion Module

To effectively integrate the kept visual tokens T_{kep} and the modified visual tokens T_{mod} , we employ an adaptive fusion module [3] which leverages the Multi-layer Perceptron (MLP) network in combination with the Sigmoid function to predict weight coefficients for each feature type:

$$\theta_1 = \text{Sigmoid}(\text{MLP}(\mathbf{f}_{\text{kep}}, \mathbf{f}_{\text{mod}})), \quad (4)$$

$$\mathbf{f}_{\text{neu}} = \theta_1 \cdot \mathbf{f}_{\text{kep}} + (1 - \theta_1) \cdot \mathbf{f}_{\text{mod}}, \quad (5)$$

where $\mathbf{f}_{\text{neu}}$ denotes the conflict-neutralized visual feature derived from the kept visual tokens $\mathbf{f}_{\text{kep}}$ and modified visual tokens $\mathbf{f}_{\text{mod}}$. Subsequently, we again apply adaptive modulation on $\mathbf{f}_{\text{neu}}$ and the text tokens $\mathbf{f}_t$ to obtain the final query representation $\mathbf{f}_{\text{query}}$:

$$\theta_2 = \text{Sigmoid}(\text{MLP}(\mathbf{f}_{\text{neu}}, \mathbf{f}_t)), \quad (6)$$

$$\mathbf{f}_{\text{query}} = \theta_2 \cdot \mathbf{f}_{\text{neu}} + (1 - \theta_2) \cdot \mathbf{f}_t. \quad (7)$$

3.5. Objective Function

Training Objectives. During training, we employ three loss functions to enhance CIR performance. First, we utilize the image-text contrastive loss $\mathcal{L}_{\text{ITC}}$ [30], a prevalent choice in existing CIR methodologies. Then, we integrate the weighted regularization triplet loss $\mathcal{L}_{\text{WRT}}$ [54], leveraging relative distances to accurately discern between positive and negative pairs of queries and target images. To further distinguish conflict information from the user intention, we introduce the orthogonal projection regularization loss $\mathcal{L}_{\text{OPR}}$ to expand the divergence between conflict information and the target image. This divergence is refined through the application of dual projection layers, wherein orthogonality is rigorously maintained during training. The total loss $\mathcal{L}_{\text{total}}$ is calculated as the summation:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ITC}} + \mathcal{L}_{\text{OPR}} + \lambda \mathcal{L}_{\text{WRT}}. \quad (8)$$

Inference. The losses $\mathcal{L}_{\text{WRT}}$ and $\mathcal{L}_{\text{OPR}}$ serve as auxiliary objectives, without imposing additional computational or parameter overhead during inference. In the inference phase, ranking scores are derived from the similarity between the multi-modal representation $\mathbf{f}_{\text{query}}$ and the target image feature $\mathbf{f}_{\text{tar}}$ within the query-target matching process.

4. Experiments

4.1. Datasets and Metrics

FashionIQ [48] dataset comprises 77,684 fashion images, carefully organized into 30,134 triplets. Each triplet consists of a reference image, a target image, and modified instructions delineating the specific modifications from the reference image to the target image. These images are curated from the web and categorized into three distinct classes: Dress, Shirt, and Tootie. Specifically, there are 18K triplets allocated for training and 6K for test.

Method	FashionIQ							CIRR				
	Dress		Shirt		Toptee		Mean	R@1	R@5	R@10	R@50	Mean
	R@10	R@50	R@10	R@50	R@10	R@50						
TIRG [39]	14.13	34.61	13.10	30.91	14.79	34.37	23.66	14.61	48.37	64.08	90.03	54.27
ARTEMIS [12]	25.68	51.05	21.57	44.13	28.59	55.06	37.68	16.96	46.10	61.31	87.73	53.03
CLIP4CIR [3]	33.81	59.40	39.99	60.45	41.41	65.37	50.03	38.53	69.98	81.86	95.93	71.58
CompoDiff [15]	40.65	57.14	36.87	57.39	43.93	61.17	49.53	22.35	54.36	73.41	91.77	60.47
SPIRIT [10]	39.86	64.30	44.11	65.60	47.68	71.70	55.54	40.23	75.10	84.16	96.88	74.09
BLIP4CIR [28]	42.09	67.33	41.76	64.28	46.61	70.32	55.04	40.15	73.08	83.88	96.27	73.35
FAME-ViL [17]	42.19	67.38	47.64	68.79	50.69	73.07	58.29	-	-	-	-	-
CaLa [19]	42.38	66.08	46.76	68.16	50.93	73.42	60.08	49.11	81.21	89.59	98.00	79.47
CoVR-BLIP [38]	44.55	69.03	48.43	67.42	52.60	74.31	59.39	49.69	78.60	86.77	94.31	77.34
SSN [51]	34.36	60.78	38.13	61.83	44.26	69.05	51.40	43.91	77.25	86.48	97.45	76.27
TG-CIR [45]	45.22	69.66	52.60	72.52	56.14	77.10	58.05	45.25	78.29	87.16	97.30	77.00
CASE [21]	47.44	69.36	48.48	70.23	50.18	72.24	59.74	48.00	79.11	87.25	97.57	77.98
Re-ranking [29]	48.14	71.43	50.15	71.25	55.23	76.80	62.15	50.55	81.75	89.78	97.18	79.82
SPRC [2]	48.83	72.09	53.83	74.14	58.13	78.58	64.27	51.96	82.12	89.74	97.69	80.37
Ours	49.38	72.58	55.93	74.14	57.93	77.56	64.59	53.41	84.05	91.17	98.00	81.66

Table 1. Evaluation of CIR performance in comparison with other state-of-the-art methods on FashionIQ [48] and CIRR [27] datasets. Specifically, traditional model-based methods (top) and VLP model-based methods (middle) are compared. For FashionIQ, “Mean” represents the average results across all metrics in the three classes. For CIRR, “Mean” denotes the average results across all metrics in the four different settings. The best results are highlighted in bold, respectively.

CIRR [27] dataset comprises 21,552 real-world images derived from the popular natural language inference dataset NLVR2 [34]. It contains 36,554 triplets randomly assigned in 80% for training, 10% for validation, and 10% for test. This dataset includes rich object interactions, effectively addressing the limitations of overly narrow domains and the high incidence of false negatives in FashionIQ [48], thereby allowing for a comprehensive evaluation.

Shoes [16] dataset is initially proposed for the attribute discovery study [5] and then developed with modified instruction annotations for the CIR task. Following the previous methods [7, 10, 44], the dataset is divided into 10K triplets for training and 4.6K for test.

Evaluation Metric. According to previous studies [2, 3, 10, 19, 28, 45], we choose the standard top-K recall metric R@K to evaluate the CIR performance. For FashionIQ [48] dataset, the evaluation focuses on R@10, R@50, and the mean of them. For CIRR [27] dataset, we specifically compare R@1, R@5, R@10, R@50, and the mean of them. For Shoes [16] dataset, we report the results concerning R@1, R@10, R@50, and the mean of them.

4.2. Implementation Details

Our implementation utilizes PyTorch on a Linux server equipped with a single NVIDIA RTX 4090 GPU. For the CCI module, we employ BLIP-2 [24] to generate captions for the reference image. We then utilize GPT-4 [1] to analyze the merged text and identify conflicting attributes

Method	R@1	R@10	R@50	Mean
TIRG [39]	12.60	45.45	69.39	42.48
Uncertainty [8]	18.41	53.63	79.84	
ARTEMIS [12]	18.72	53.11	79.31	50.38
AMC [58]	19.99	56.89	79.27	52.05
CMAP [25]	21.48	56.18	81.14	52.93
FashionVLP [14]	-	49.08	77.32	-
CF-TCIR [52]	18.95	53.17	79.59	50.57
FashionERN [9]	-	55.59	81.71	-
SPIRIT[10]	-	56.90	81.49	-
Prog. Lrn.[56]	22.88	58.83	84.16	55.29
CAFF [41]	25.21	60.17	80.79	55.39
TG-CIR [45]	25.89	63.20	85.07	58.05
Ours	25.95	65.76	86.54	59.42

Table 2. Evaluation of CIR performance in comparison with other state-of-the-art methods on Shoes [16] dataset. The best results are highlighted in bold, respectively.

against the user intention. Subsequently, we leverage GPT-4 to isolate the conflicting attributes from the reference image caption, thereby producing a refined kept instruction. For the CCN module, we leverage the pre-trained CLIP aligned with SPRC [2] for feature extraction and employ Adam [20] optimizer with a weight decay of 0.05. The training process is conducted for 30 epochs on FashionIQ [48], 50 epochs on CIRR [27], and 30 epochs on Shoes [16]. The learning rates are initialized at $7e-6$, $4e-6$, and $1e-5$, following a cosine schedule for three datasets. Additionally, we perform LoRA [18] fine-tuning on ViT backbone at the 38th and 39th layers. For the hyperparameter λ , we adopt the value of (1, 0.1, 1) across the three datasets.

B/L	CCIN			FashionIQ		
	Architecture	$\mathcal{L}_{WRT}$	$\mathcal{L}_{OPR}$	R@10	R@50	Mean
●	○	○	○	52.57	73.66	63.11
●	●	○	○	53.30	73.90	63.60
●	●	●	○	53.84	74.46	64.15
●	●	●	●	53.86	74.75	64.31

Table 3. Ablation studies of the proposed CCIN framework on FashionIQ [48] dataset, respectively.

CCN		Textual	FashionIQ			CIRR
Kept	Modified		R@10	R@50	Mean	Mean
●	○	○	11.34	26.52	18.93	39.30
●	●	○	53.93	74.27	64.10	81.33
○	○	●	37.14	58.83	47.99	71.87
●	●	●	54.42	74.76	64.59	81.66

Table 4. Ablation studies of the CCN module on FashionIQ [48] and CIRR [27] datasets, respectively.

4.3. Comparison with State-of-the-art Methods

We evaluate the effectiveness of the proposed method by comparing it with other state-of-the-art approaches on FashionIQ [48] and CIRR [27] datasets, as detailed in Table 1. Unsurprisingly, traditional model-based methods, such as TIRG [39] and ARTEMIS [12], which rely on conventional architectures like ResNet and LSTM for feature extraction, perform poorly across all metrics when compared to advanced VLP model-based methods. This trend underscores the superior multi-modal encoding capabilities inherent to VLP models. In comparison to the second-best method SPRC [2], our method delivers substantial improvements in the “Shirt” and “Dress” categories on FashionIQ, while also demonstrating superior performance on CIRR. Such observation indicates the effectiveness of conflict identification and neutralization, as opposed to merely converting query images into structured visual prompts. Additionally, compared to TG-CIR [45], which leverages learnable masks to obscure conflicting features within the reference image, our method also achieves better performance across two settings. This result highlights the enhanced efficacy of our CCIN framework, which surpasses the limitations of heuristic mask prediction by providing a more precise and controllable approach for conflict identification and neutralization. It is worth noting that our method yields an average performance improvement of 2.44% on FashionIQ and 1.91% on CIRR compared to the two-stage Re-ranking mechanism [29], though the latter requires longer training and inference time. Overall, the results clearly demonstrate that our method outperforms other comparable methods across both the FashionIQ and CIRR datasets.

Furthermore, we conduct additional experiments on Shoes [16] dataset, where the proposed method consistently

Dataset	Conflict Attribute Quantity			
	1 conflict	2 conflicts	3 conflicts	4 conflicts
FashionIQ	8741	2984	708	103
CIRR	11119	4177	926	140
Shoes	3063	1190	175	29
Sum	22923	8351	1809	272

Table 5. Statistical experiments of conflict attribute quantities identified by the CCI module, respectively.

Layer 39		Layer 38		FashionIQ		
MLP	Attn	MLP	Attn	R@10	R@50	Mean
●	○	○	○	53.87	74.81	64.34
●	●	○	○	54.03	75.00	64.52
●	●	●	○	53.63	74.26	63.95
●	●	●	●	54.42	74.76	64.59

Table 6. Ablation studies of the ViT layers involved in LoRA fine-tuning on FashionIQ [48] dataset, respectively.

demonstrates significant improvements over other comparable methods, as depicted in Table 2.

4.4. Ablation Studies

Basic Component Analysis. The impact of each component within the proposed method is investigated in Table 3. Specifically, “B/L” denotes the baseline, which does not consider conflict issues and relies exclusively on ITC loss to optimize the model. “Architecture” refers to the basic architecture of CCIN, while “ $\mathcal{L}_{WRT}$ ” and “ $\mathcal{L}_{OPR}$ ” represent the weighted regularization triplet loss and orthogonal projection regularization loss. The first two rows of Table 3 clearly demonstrate the beneficial impact of CCIN. By facilitating comprehensive identification and neutralization of compositional conflicts, our method achieves better alignment with user intention, contributing to enhanced CIR performance, particularly at R@10. Upon comparing rows 2 to 4, we observe that the gradual introduction of the triplet loss and orthogonal loss leads to incremental improvements in CIR performance. This progressive enhancement substantiates that adding these two losses makes the model more sensitive to conflict information and expands the distinctiveness among different query-target matches.

CCN Module Analysis. To examine the influence of three types of tokens in CCN module, we conduct ablation experiments with various combinations, as illustrated in Table 4. “Kept” and “Modified” denotes the kept instruction and the modified instruction, and “Textual” represents the textual features. Comparing the first two rows, we observe that utilizing dual instructions to guide visual feature extraction proves more effective than relying solely on the kept instruction. This is attributed to the fact that the kept instruc-



Figure 4. Qualitative results comparing SPRC [2] with the proposed method, with respect to Recall@1 metric on FashionIQ [48] and CIRRR [27] datasets. The reference image and modified instruction are displayed on the left, while the retrieval rank is presented on the right. Correct matches are highlighted within green rectangles, respectively.

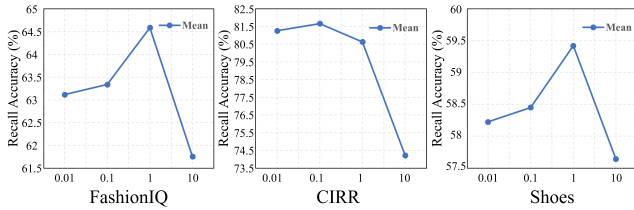


Figure 5. Mean recall accuracy across various hyper-parameter values of λ on three datasets [16, 27, 48], respectively.

tion is restricted to extracting only non-conflicting visual features within the reference image, thus capturing a small aspect of user intention. Conversely, the modified instruction conveys the major portion of the user search intention. The comparison between rows 1 and 3 further substantiates this observation. Moreover, comparing the last two rows reveals that relying exclusively on textual features without considering visual features, results in a noticeable decline in performance, as text alone is insufficient to convey the nuanced visual details essential for the CIR process.

LoRA Fine-tuning Analysis. To explore the optimal layer configuration for LoRA fine-tuning, we conduct a series of experimental trials, and the results are detailed in Table 6. Specifically, we examine the efficacy of fine-tuning operations at the 38th and 39th layers of the ViT backbone. Comparing the first two rows, we find that fine-tuning isolated modules within a layer is less effective than fine-tuning all modules simultaneously. Additionally, comparing rows 2 and 4 reveals that exclusively fine-tuning the 39th layer results in the best R@50 performance, while fine-tuning both the 39th and 38th layers leads to the highest R@10 performance. Such observation suggests that the larger parameter adjustment allows the model to better integrate complex queries, resulting in improved R@10 per-

formance. Nonetheless, excessive fine-tuning can lead to overfitting and compromise the generalization ability, as evidenced by a marginal decline in the R@50 metric.

Hyper-parameter λ Analysis. To determine the optimal value for the hyper-parameter λ , we perform an ablation study exploring a range of values: 0.01, 0.1, 1, and 10. As illustrated in Figure 5, increasing λ initially improves the CIR performance across all three datasets, but beyond a certain point, we observe a notable decline. These results indicate that λ serves as a crucial proxy for controlling the intensity of the relative distance between positive and negative pairs. Finally, we leverage the value of (1, 0.1, 1) for FashionIQ, CIRRR, and Shoes datasets, respectively.

4.5. Qualitative Results

Analysis of CIR Results. As mentioned, compositional conflicts inevitably lead to misalignment when straightforwardly integrating multi-modal features, which poses a significant challenge for existing CIR methods. To address this, the proposed method is meticulously designed to identify and neutralize such conflicts for effective CIR. To evaluate the efficacy of CCIN, we conduct a comparative analysis against SPRC [2], which does not incorporate mechanisms to address compositional conflicts. At the top of Figure 4, we illustrate a case from the FashionIQ dataset, where the query contains conflicting attributes (e.g., sleeve length and midriff pattern). Our approach effectively neutralizes these conflicting attributes, generates a kept instruction (“a woman wearing a white shirt”), and integrates it with the modified instruction to achieve a closer alignment with the user intention. In contrast, SPRC lacks this disambiguation capability, erroneously interpreting the midriff color as the shirt color, leading to inaccurate predictions. A similar example from the CIRRR dataset, focusing on the posture and



Figure 6. Qualitative results of CCI on FashionIQ [48], CIRR [27], and Shoes [16] datasets. Specifically, T_{ref} and T_{mod} denote the reference image caption and the modified instruction. The conflict identification results are highlighted in purple, respectively.

orientation of the dog, is shown at the bottom of Figure 4.

Analysis of CCI Module. To showcase the effectiveness of CCI, we present the conflict identification results from three datasets. The statistical experiments in Table 5 highlight the capabilities of the conflict identification process. Additionally, the cases presented in Figure 6 further illustrate that, whether in the fashion or natural domains, CCI consistently and accurately identifies contradictions. Comparison experiments in Table 1 and Table 2 further demonstrate the validity of overall conflict resolution process.

Analysis of Different LLMs. Moreover, to assess the performance of various LLMs in identifying compositional conflicts within queries, we visualize results across three datasets, focusing specifically on LLama2-70B [36], GPT-3.5-Turbo [6], and GPT-4 [1], as shown in Figure 7. In the first row, it is evident that GPT-3.5-Turbo fails to detect an existing conflict regarding sleeve length and instead misidentifies the shirt color as a conflicting attribute, while LLama2-70B goes even further by misinterpreting the user intention as conflicting elements. Likewise, in the second and third rows, LLama2-70B and GPT-3.5-Turbo either overlook specific conflicts or erroneously flag non-conflicting attributes as conflicting. In contrast, GPT-4 consistently achieves accurate conflict identification. These findings suggest that GPT-4 surpasses LLama2-70B and GPT-3.5-Turbo in discerning compositional conflicts, un-

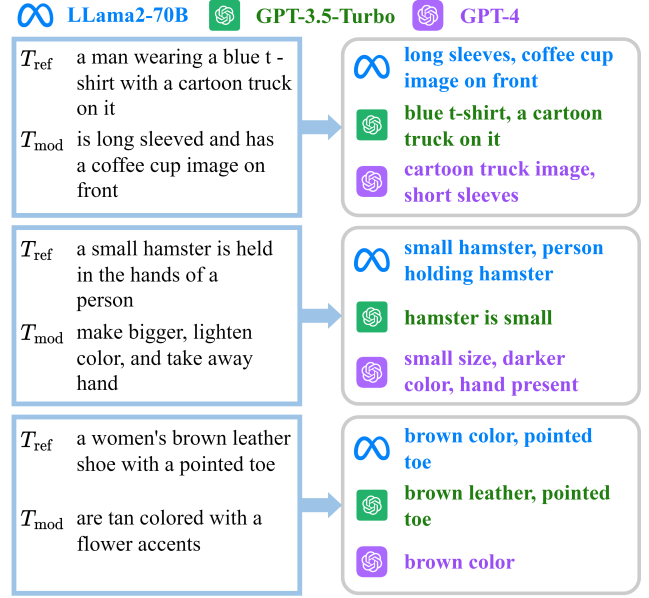


Figure 7. Qualitative results of different LLMs in identifying compositional conflicts on FashionIQ [48], CIRR [27], and Shoes [16] datasets. Specifically, T_{ref} and T_{mod} denote the reference image caption and the modified instruction. LLama2-70B [36] and GPT-3.5-Turbo [6] results are highlighted in blue and green, while GPT-4 [1] results are in purple, respectively.

derscoring its advanced capabilities in the CIR task.

5. Conclusion

In this paper, we concentrate on compositional conflicts arising from the juxtaposition of the reference image and the modified instruction within the Composed Image Retrieval (CIR) task. To address this issue, we propose the Compositional Conflict Identification and Neutralization (CCIN) framework to facilitate the CIR task. The CCIN is uniquely designed to accurately identify specific conflicting attributes and neutralize them in a controllable manner. The proposed method significantly improves the efficiency of conflict resolution, thereby boosting the overall effectiveness of retrieval. Extensive experiments demonstrate that CCIN surpasses the state-of-the-art methods. In future work, we plan to explore employing Multi-modal Large Language Models (e.g., GPT-4o) for robust conflict neutralization. Additionally, incorporating more input modalities (e.g., sketch) could further broaden the task scope.

Acknowledgements

This research is supported by Hubei Key R&D Project (2022BAA033) and National Natural Science Foundation of China (62171325, 62476224). The numerical calculations have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5, 8
- [2] Yang bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval. In *ICLR*, 2024. 1, 2, 5, 6, 7
- [3] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *CVPR*, pages 4959–4968, 2022. 2, 4, 5
- [4] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *CVPR*, pages 21466–21474, 2022. 1, 2
- [5] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, pages 663–676, 2010. 5
- [6] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 8
- [7] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *CVPR*, pages 3001–3011, 2020. 5
- [8] Yiyang Chen, Zhedong Zheng, Wei Ji, Leigang Qu, and Tat-Seng Chua. Composed image retrieval with text feedback via multi-grained uncertainty regularization. In *ICLR*, 2024. 5
- [9] Yanzhe Chen, Huasong Zhong, Xiangteng He, Yuxin Peng, Jiahuan Zhou, and Lele Cheng. Fashionern: Enhance-and-refine network for composed fashion image retrieval. In *AAAI*, pages 1228–1236, 2024. 5
- [10] Yanzhe Chen, Jiahuan Zhou, and Yuxin Peng. Spirit: Style-guided patch interaction for fashion image retrieval with text feedback. *TOMM*, 20(6):1–17, 2024. 1, 2, 5
- [11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 2, 4
- [12] Ginger Delmas, Rafael S Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *ICLR*, 2022. 5, 6
- [13] Zhenpeng Feng, Hongbing Ji, Miloš Daković, Xiyang Cui, Mingzhe Zhu, and Ljubiša Stanković. Cluster-cam: Cluster-weighted visual interpretation of cnns’ decision in image classification. *Neural Networks*, page 106473, 2024. 2
- [14] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *CVPR*, pages 14105–14115, 2022. 5
- [15] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *TMLR*, 2024. 5
- [16] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. *NeurIPS*, 31, 2018. 2, 5, 6, 7, 8
- [17] Xiao Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. In *CVPR*, pages 2669–2680, 2023. 5
- [18] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 5
- [19] Xintong Jiang, Yaxiong Wang, Mengjian Li, Yujiao Wu, Bingwen Hu, and Xueming Qian. Cala: Complementary association learning for augmenting composed image retrieval. In *SIGIR*, pages 2177–2187, 2024. 2, 5
- [20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [21] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval. In *AAAI*, pages 2991–2999, 2024. 5
- [22] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 34:9694–9705, 2021. 2
- [23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 2
- [24] 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 *ICML*, pages 19730–19742, 2023. 2, 5
- [25] Shenshen Li, Xing Xu, Xun Jiang, Fumin Shen, Zhe Sun, and Andrzej Cichocki. Cross-modal attention preservation with self-contrastive learning for composed query-based image retrieval. *TOMM*, 20(6):1–22, 2024. 1, 2, 5
- [26] Kejun Lin, Zhixiang Wang, Zheng Wang, Yinqiang Zheng, and Shin’ichi Satoh. Beyond domain gap: exploiting subjectivity in sketch-based person retrieval. In *ACM MM*, pages 2078–2089, 2023. 1
- [27] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, pages 2125–2134, 2021. 2, 5, 6, 7, 8
- [28] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning. In *WACV*, pages 5753–5762, 2024. 1, 2, 5
- [29] Zheyuan Liu, Weixuan Sun, Damien Teney, and Stephen Gould. Candidate set re-ranking for composed image retrieval with dual multi-modal encoder. *TMLR*, 2024. 5, 6
- [30] 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 *ICML*, pages 8748–8763, 2021. 2, 4
- [31] Jie Ren, Mingjie Li, Zexu Liu, and Quanshi Zhang. Interpreting and disentangling feature components of various complexity from dnns. In *ICML*, pages 8971–8981, 2021. 2
- [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 *CVPR*, pages 19305–19314, 2023. 2
- [33] Minchul Shin, Yoonjae Cho, Byungsoo Ko, and Geonmo Gu. Rtic: Residual learning for text and image composition using graph convolutional network. *arXiv preprint arXiv:2104.03015*, 2021. 2
- [34] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, pages 6418–6428, 2019. 5
- [35] Likai Tian, Zhengwei Yang, Zechao Hu, Hao Li, Yifang Yin, and Zheng Wang. Expressiveness is effectiveness: Self-supervised fashion-aware clip for video-to-shop retrieval. In *IJCAI*, pages 1335–1343, 2024. 1
- [36] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 8
- [37] A Vaswani. Attention is all you need. *NeurIPS*, 30, 2017. 2
- [38] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. In *AAAI*, pages 5270–5279, 2024. 5
- [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 *CVPR*, pages 6439–6448, 2019. 5, 6
- [40] 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 *CVPR*, pages 6439–6448, 2019. 2
- [41] Yongquan Wan, Wenhai Wang, Guobing Zou, and Bofeng Zhang. Cross-modal feature alignment and fusion for composed image retrieval. In *CVPR*, pages 8384–8388, 2024. 5
- [42] Jinpeng Wang, Pan Zhou, Mike Zheng Shou, and Shuicheng Yan. Position-guided text prompt for vision-language pre-training. In *CVPR*, pages 23242–23251, 2023. 2
- [43] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin’ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*, pages 618–626, 2019. 1
- [44] Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie. Comprehensive linguistic-visual composition network for image retrieval. In *SIGIR*, pages 1369–1378, 2021. 5
- [45] Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. Target-guided composed image retrieval. In *ACM MM*, pages 915–923, 2023. 1, 2, 3, 5, 6
- [46] Haokun Wen, Xuemeng Song, Xiaolin Chen, Yinwei Wei, Liqiang Nie, and Tat-Seng Chua. Simple but effective raw-data level multimodal fusion for composed image retrieval. In *SIGIR*, pages 229–239, 2024. 2
- [47] Jie Wen, Chengliang Liu, Gehui Xu, Zhihao Wu, Chao Huang, Lunke Fei, and Yong Xu. Highly confident local structure based consensus graph learning for incomplete multi-view clustering. In *CVPR*, pages 15712–15721, 2023. 3
- [48] 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 *CVPR*, pages 11307–11317, 2021. 2, 4, 5, 6, 7, 8
- [49] Xiaohui Xie, Jiaxin Mao, Yiqun Liu, and Maarten de Rijke. Modeling user behavior for vertical search: images, apps and products. In *SIGIR*, pages 2440–2443, 2020. 1
- [50] Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning. In *AAAI*, pages 16129–16137, 2024. 2, 3
- [51] Xingyu Yang, Daqing Liu, Heng Zhang, Yong Luo, Chaoyue Wang, and Jing Zhang. Decomposing semantic shifts for composed image retrieval. In *AAAI*, pages 6576–6584, 2024. 2, 3, 5
- [52] Yuchen Yang, Yu Wang, and Yanfeng Wang. Cf-tcir: A compositor-free framework for hierarchical text-conditioned image retrieval. In *ACL*, pages 16315–16325, 2024. 5
- [53] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *CVPR*, pages 1472–1481, 2023. 1
- [54] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *ICCV*, pages 13567–13576, 2021. 4
- [55] Xianchao Zhang, Mengyan Chen, Jie Mu, and Linlin Zong. Adaptive view-aligned and feature augmentation network for partially view-aligned clustering. In *PAKDD*, pages 223–235, 2023. 3
- [56] Yida Zhao, Yuqing Song, and Qin Jin. Progressive learning for image retrieval with hybrid-modality queries. In *SIGIR*, pages 1012–1021, 2022. 5
- [57] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. Group-aware label transfer for domain adaptive person re-identification. In *CVPR*, pages 5310–5319, 2021. 1
- [58] Hongguang Zhu, Yunchao Wei, Yao Zhao, Chunjie Zhang, and Shujuan Huang. Amc: Adaptive multi-expert collaborative network for text-guided image retrieval. *TOMM*, 19(6), 2023. 5