Leveraging Large Vision-Language Model as User Intent-Aware Encoder for Composed Image Retrieval

Zelong Sun[1], **Dong Jing**[1], **Guoxing Yang**[1, 2], **Nanyi Fei**[2], **Zhiwu Lu**[1,*]

[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]MetaBrain AGI Lab, Shanghai, China
zelongsun@ruc.edu.com, luzhiwu@ruc.edu.com

Abstract

Composed Image Retrieval (CIR) aims to retrieve target images from candidate set using a hybrid-modality query consisting of a reference image and a relative caption that describes the user intent. Recent studies attempt to utilize Vision-Language Pre-training Models (VLPMs) with various fusion strategies for addressing the task. However, these methods typically fail to simultaneously meet two key requirements of CIR: comprehensively extracting visual information and faithfully following the user intent. In this work, we propose CIR-LVLM, a novel framework that leverages the large vision-language model (LVLM) as the powerful user intent-aware encoder to better meet these requirements. Our motivation is to explore the advanced reasoning and instruction-following capabilities of LVLM for accurately understanding and responding the user intent. Furthermore, we design a novel hybrid intent instruction module to provide explicit intent guidance at two levels: (1) The task prompt clarifies the task requirement and assists the model in discerning user intent at the task level. (2) The instance-specific soft prompt, which is adaptively selected from the learnable prompt pool, enables the model to better comprehend the user intent at the instance level compared to a universal prompt for all instances. CIR-LVLM achieves state-of-the-art performance across three prominent benchmarks with acceptable inference efficiency. We believe this study provides fundamental insights into CIR-related fields.

Introduction

Compared with conventional image retrieval (Gordo et al. 2016; Liu et al. 2016; Frome et al. 2013; Gao et al. 2020), which involves only a single-modality query, Composed Image Retrieval (CIR) faces greater challenges as its query comprises both visual (i.e., a reference image) and textual (i.e., a relative caption) modalities. Specifically, CIR aims to retrieve the target image according to the user intent described in the relative caption and the visual information contained in the reference image. To achieve the goal, two fundamental challenges are presented: (1) How to comprehensively extract visual information contained in the reference image. (2) How to accurately capture and understand the user intent embedded in the composed query.

*Corresponding Author

Existing methods (Baldrati et al. 2022; Zhao, Song, and Jin 2022; Levy et al. 2023; Liu et al. 2023b) attempt to utilize Vision-Language Pre-training Models (VLPMs) (Radford et al. 2021; Li et al. 2022) with various fusion strategies. Particularly, some methods (Levy et al. 2023; Liu et al. 2023b) propose to integrate image embedding with text embedding by an unimodal encoder with intermediate cross-attention layers, known as the "early-fusion". However, due to the limitation of cross attention mechanism, these methods fail to retain the desired visual information, which is contained in the reference image but not mentioned in the relative caption. For example, as shown in Fig.1 (a), such methods tend to miss the "species of Corgi". More recent approaches (Saito et al. 2023; Bai et al. 2023b) introduce a textual inversion module to generate a textual prompt from the reference image and concatenate it with the relative caption. These approaches allow a more flexible interaction for hybrid-modality queries, enabling the model to perceive more comprehensive information from the reference image. However, the limitations of the text encoder may still result in missed user intent, especially when complex relative captions are provided. As shown in Fig.1 (b), these approaches tend to ignore the user intent to "move the dog outdoors".

Recently, several generative retrieval methods (Karthik et al. 2023; Sun, Ye, and Gong 2023) have emerged to leverage the advanced reasoning capability of large language models (LLMs) to better perceive the user intent by inferring target image captions with LLMs from composed queries. However, the user intent captured in the inferred captions often remains at a coarse granularity level, resulting in suboptimal retrieval accuracy. Notably, these methods encounter substantial efficiency challenges due to their reliance on multi-pass decoding processes.

To overcome the limitations of the current state-of-the-art methods, we propose **CIR-LVLM**, a single-pass encoding framework designed to adopt a large vision-language model (LVLM) as a user intent-aware encoder to derive both query and target embeddings. As shown in Fig.1 (c), we employ the Connector module to generate a sentence-level prompt that enables the model to perceive comprehensive information from the image. Instead of using a text encoder, we leverage advanced LVLMs, which exhibit superior reasoning and instruction-following abilities, to accurately discern the user intent and obtain the desired information.

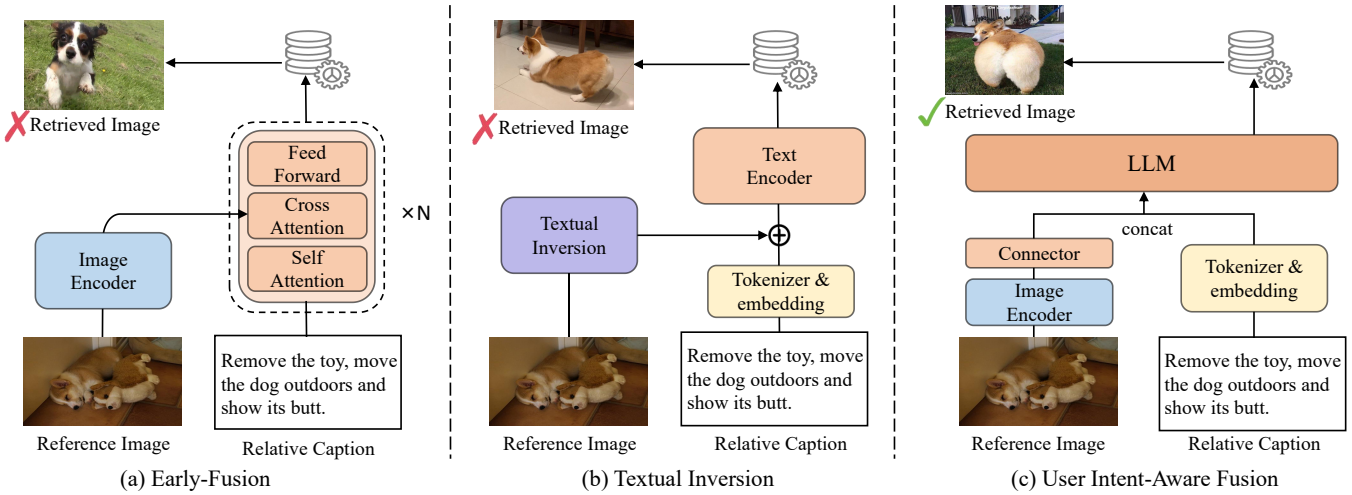


Figure 1: Workflows of existing CIR methods and our proposed CIR-LVLM: (a) Early-fusion, (b) Textual inversion, and (c) Our proposed CIR-LVLM. It can be seen that the first two fusion strategies fail to discern the user intent conveyed by the relative caption: (a) fails to retain *the species of Corgi*, and (b) fails to *move the dog outdoors*. Our fusion strategy leverages the superior user intent-aware capability of LVLM and successfully recalls the target image.

To harness CIR-LVLM's reasoning capabilities and help it grasp CIR task patterns, we design a novel hybrid intent instruction module consisting of two kinds of prompts, providing explicit intent guidance at two levels: **(1) Task prompt:** We propose a task prompt to help the model discern user intent comprehensively at the task-level for the CIR task. This prompt provides detailed guidance on the task requirements, enabling the LVLM to accurately obtain the desired information from both image and text within the given context. Given the differing requirements when processing query and target images, we have designed specific task prompts for each process. **(2) Instance-specific soft prompt:** We further employ a learnable prompt pool that automatically selects appropriate soft prompts based on the input image and text for each instance to provide Instance-level guidance. Different from using a universal prompt for all instances (Zhou et al. 2022; Shin et al. 2020), this process allows for adaptive knowledge consolidation over changing user intent in each instance by mapping similar CIR instances onto similar prompts, maintaining the user intent-aware ability across various instance-specific requirements.

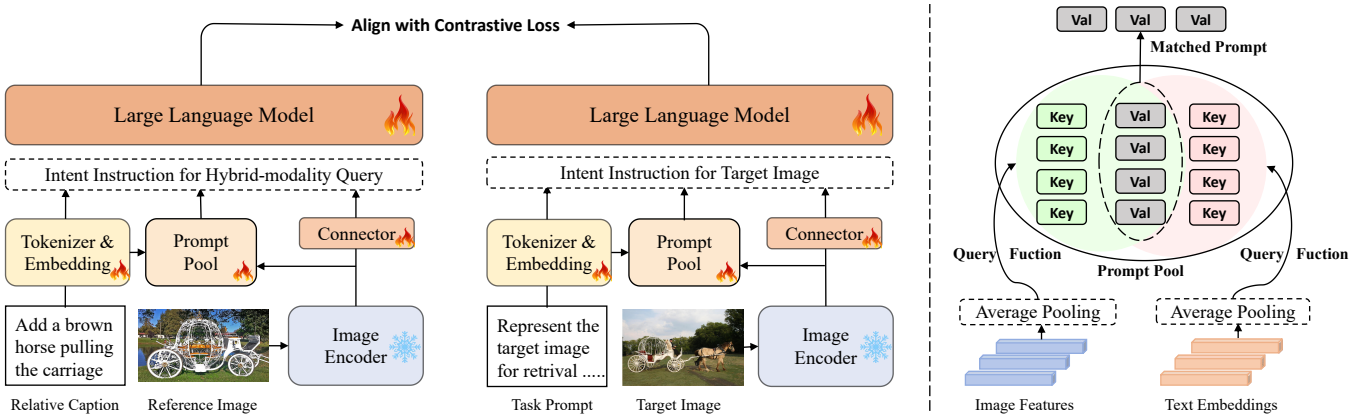
In summary, the main contributions of our paper include:
(1) To the best of our knowledge, we are the first to explore adopting the LVLM as a user intent-aware encoder in the CIR task, which accurately capture user intentions while maintaining acceptable inference efficiency
(2) We propose a novel hybrid intent instruction module tailored for the CIR task, providing explicit intent guidance at both task and instance levels.
(3) Extensive experiments conducted on three benchmarks demonstrate that CIR-LVLM outperforms the state-of-the-art CIR methods. Furthermore, our method can be easily transferred to the latest state-of-the-art models as they evolve, achieving even higher performance.
(4) We are the first to clearly show that in multimodal retrieval tasks that require reasoning (e.g., CIR), the LVLMs have the potential to surpass the VLPMs.

Related Work

Composed Image Retrieval

The prevalent contemporary CIR methods predominantly leverage the advancements in VLPMs (Radford et al. 2021; Li et al. 2022) as the foundational encoders and propose various strategies to adapt them to CIR task. Among them, CASE (Levy et al. 2023) and Re-ranking (Liu et al. 2023b) adapt the early fusion strategy, leveraging the unimodal encoder of BLIP to fuse the information of the query. These methods enhance modality fusion by finer textual-visual interaction at the token-patch level. However, since the textual-visual interaction only relies on the intermediate cross-attention layers in the unimodal encoder, these methods tend to fail to extract visual information contained in the reference image but not mentioned in the relative caption. Another category of approaches introduces a textual inversion module to transform reference image into its pseudo-word embedding (Gal et al. 2022; Saito et al. 2023; Tang et al. 2024) or a sentence-level prompt (Bai et al. 2023b), which is then concatenated with the relative caption for text-to-image retrieval. However, due to the limitation of the text encoder, the user intent is still hard to be captured.

Recently some methods have tried to use LLM to recognize user intent for zero-shot retrieval. CIReVL (Karthik et al. 2023) and GRB (Sun, Ye, and Gong 2023) employ a caption model to generate captions for reference images and utilize LLMs to identify user intentions and revise these captions. However, this method faces several challenges: 1. It is prone to compatibility issues between different modules. 2. The multi-pass decoding design of the caption model and LLM results in inefficiency and hallucination issues (Huang et al. 2024; Wang et al. 2023). These limitations restrict their application in real-world scenarios. In this work, we fine-



(a) Overview of CIR-LVLM

(b) Details of Prompt Pool

Figure 2: (a) Illustration of the architecture overview of our proposed model. All the parameters are shared between the query and target image. The intent instructions are used to form the inputs of LLM. The details of intent instructions can be found in Fig.3. (b) Details of the prompt pool. We select prompts according to both visual features and text embeddings.

tuned the LVLM as a user intent-aware encoder to improve efficiency and reduce hallucination issues.

Large Vision-Language Model

With the rapid advancement of large language models (Brown et al. 2020; Touvron et al. 2023; OpenAI 2023), researchers are increasingly exploring the integration of multimodal knowledge into these models. Large vision-language models (LVLMs) (Li et al. 2023; Liu et al. 2023a; Dai et al. 2024; Zhu et al. 2023a; Bai et al. 2023a) have emerged as a prominent avenue for enhancing instruction-following capabilities through the incorporation of visual instruction tuning. With the unique model architecture, these models show exceptional performance across various visual tasks (Agrawal et al. 2019; Hudson and Manning 2019; Mishra et al. 2019). In the text retrieval domain, fine-tuning LLMs as retrievers has proven to be effective, showcasing promising results (Ma et al. 2023; Muennighoff 2022; Muennighoff et al. 2024). However, few studies have shed light on employing LVLMs as encoders in the multimodel retrieval domain. Considering the effectiveness of LVLM in understanding and following user modification intent, in this work, we aim to explore the possibility of employing LVLM as the user intent-aware encoder in CIR task.

Prompt Tuning

Prompt tunings (Han et al. 2022; Jin et al. 2021; Zang et al. 2022) is an efficient, low-cost way of adapting a pre-trained foundation model to new downstream tasks. In this work, We devise the task (hard) prompt to clarify the task requirement of CIR and add soft prompts to further stimulate the reasoning ability of LLMs. Prior related efforts include L2P (Wang et al. 2022), which demonstrates the potential of learnable prompts stored in a shared pool to enable continual learning without a rehearsal buffer. We also use a prompt pool to allocate soft prompts for each instance. Considering the multimodal input in CIR tasks, we select prompts using both visual and textual inputs. Such instance-specific soft prompts, compared to a universal prompt for all instances (Zhou et al. 2022; Shin et al. 2020), allow the model to focus on user intent at the instance level.

Method

Preliminary

Task Definition. Given a hybrid-modality query $Q = \{I_r, T\}$, where I_r denotes the reference image and T denotes the relative caption, and a candidate set $D = \{I_t^1, I_t^2, ..., I_t^{N_D}\}$ consisting of N_D images, the goal of CIR is to identify the k target images from the candidate set D that are most relevant to the query Q, with $k \ll N_D$.

Challenge. The core challenge of the CIR task is to modify the image content based on the relative caption while retaining as much of the reference image content as possible. Achieving such precise modifications requires the model to not only have a comprehensive understanding of the image but also possess strong reasoning abilities to accurately discern the user's intended modifications.

In this work, we aim to leverage the advanced reasoning and instruction comprehension capabilities of LVLMs to accurately capture user intent and enhance performance on CIR tasks, while maintaining acceptable inference efficiency. We introduce CIR-LVLM, a novel framework that fine-tunes the LVLM to function as a user intent-aware encoder, extracting representative embeddings of both hybrid-modality queries and candidate images. Thus, CIR-LVLM need to meets three key challenges: **(1)** How to effectively extract and process visual features? **(2)** How to accurately understand the complex user requirements in CIR tasks? **(3)** How to effectively aggregate and utilize the output features of LVLM?

Model Architecture

Overview. As shown in Fig.2, we deploy the Connector containing a set of learnable query embeddings to adaptively capture the desired visual content and map it into a sentence-level prompt. This component ensures that the LLM can

comprehensively perceive and understand the visual information. Then, we use LLM to identify implicit user intent from images and text, obtaining rich-content embeddings. However, since CIR tasks involve three inputs and are inherently complex, fully leveraging the model's ability to infer user intentions requires the model to accurately understand the relationships among these inputs. To achieve this, we propose a novel hybrid intent instruction module that includes both a task prompt and an instance-specific soft prompt, providing two levels of guidance. Finally, to facilitate the LLM's transition from a generative model to an encoder, we experimented with three pooling strategies to aggregate features. In the subsequent sections, we will delve deeper into the details of these components.

Hybrid Intent Instruction Module. CIR is a complex task that imposes different requirements to the model when processing queries and target images: deriving query embedding involves integrating information from both the reference image and relative caption, while deriving target embedding requires comprehensive extraction of visual details from the target image. To address these complexities and bolster the CIR-LVLM's understanding of the CIR task requirements, we design two hybrid intent instructions tailored for this context, offering explicit guidance from perspectives of task and instance. As depicted in Fig.3, this module emphasizes three critical elements for crafting effective prompts: Task Input, Task Prompt and Instance-Specific Soft Prompt. For hybrid-modality query, task input includes a reference image and a relative caption. When processing a target image, the task input consists solely of the target image.

Task Prompt. The task prompt is designed to provide task-level intent guidance, clarifying the task requirements and aiding the model in comprehensively discerning user intent at the task level. Importantly, we have crafted distinct task prompts for the hybrid-modality query and the target image to cater to their respective specific needs. A detailed and well-crafted prompt can help the model better utilize its reasoning capabilities, learn higher-level concepts, and obtain more accurate representations for the input.

Instance-specific Soft Prompt. The instance-specific soft prompt refines the LLM's focus toward task-specific nuances. Previous works (Zhou et al. 2022; Shin et al. 2020) focus on using a universal prompt for all instances to improve the performance of pre-trained models. However, each instance in CIR contains subtle differences in user intent. For example, the instance shown in Fig.2 involves the addition of objects, whereas the instance shown in Fig.1 includes object removal, scene changes, and other complex modifications. To address these variations and provide instance-level guidance for CIR-LVLM, we propose a shared pool of prompts that adaptively selects an instance-specific soft prompt.

Ideally, we want the model to leverage related past experiences, where similar input tend to retrieve the same group of prompts from the pool. However, since both the image and relative caption in CIR are crucial for prompt selection, we design two keys for each prompt, as shown in Fig.2 (b). The prompt pool is thus defined as:

$$P_K = \{(k_1^{img}, k_1^{text}, P_1), \cdots, (k_M^{img}, k_M^{text}, P_M)\} \quad (1)$$

Intent Instruction for Hybrid-modality Query

***

[Task Input]:
Reference image: <img>{Reference Image}</img>
Relative caption: <relcap>{Relative Caption}</relcap>

***

[Soft Prompt]: Soft Prompt:<sp>{Instance-Specific Soft Prompt}</sp>

***

[Task Prompt]: The relative caption expresses the user's modification intent for the reference image. Please carefully observe the contents of the reference image and treat the relative caption as a user instruction to make corresponding modifications to the contents of the reference image. Then, generate a representation of the modified image to retrieve the target image. Please be sure to emphasize the user's modification intention and make sure that the representation contains all the content in the relative caption.

Intent Instruction for Target Image

***

[Task Input]:
Target image: <img>{Target image}</img>

***

[Soft Prompt]: Soft Prompt:<sp>{Instance-Specific Soft Prompt}</sp>

***

[Task Prompt]: Represent the target image for a retrieval task, highlighting the main features and any prominent details. Ensure that the representation comprehensively reflects the visual characteristics and content of the target image.

Figure 3: Illustration of intent instructions for the hybrid-modality query and the target image. An intent instruction consists of three components: (1) Task Input, (2) Task Prompt, and (3) Instance-Specific Soft Prompt.

where M is the length of prompt pool, $P_m \in \mathbb{R}^{L_p \times d_t}$ is a learnable soft prompt with pre-defined length L_p and the embedding size d_t, and k_m^{img} and k_m^{text} are learnable embeddings that represent the key for image with the shape of $\mathbb{R}^{d_i}$ and the key for text with the shape of $\mathbb{R}^{d_t}$, respectively. We denote the set of all keys by $K = \{< k_m^{img}, k_m^{text} >\}_{i=1}^{M}$.

As shown in Fig.2 (b), our approach calculates the distance between image features and text embeddings with their corresponding keys, recalls the $top{-}\mathcal{K}$ closest prompts, and concatenates them in order of proximity to form an instance-specific soft prompt. we use average pooling to aggregate the image features extracted from a frozen vision encoder as $q(I)$, where I can be either a reference image I_r or a target image I_t. Similarly, the text embedding $q(T)$ is obtained by averaging the text input embeddings of LLM. Note that T represents the task prompt in Fig.3 for the target image and the relative caption for the query. We then get a subset $K_{(I,T)}$ of $top - \mathcal{K}$ keys selected from K by simply solving the objective:

$$K_{(I,T)} = \underset{\{s_i\}_{i=1}^{\mathcal{K}} \in [1:M]}{\arg\min} \sum_{i=1}^{\mathcal{K}} \Upsilon(q(I), k_{s_i}) + \Upsilon(q(T), k_{s_i}) \quad (2)$$

where $\Upsilon(\cdot)$ could be the cosine distance or another appropriate metric. This approach ensures that the chosen prompts are those most aligned with both the visual and textual inputs, allowing the model to respond more accurately and contextually to the given inputs. To distinguish the instance-specific soft prompt, we add two special tokens: $< sp >$ and

Method	FashionIQ									Shoes			
	Shirt		Dress		Tops&Tees		Avg.						
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	R_{mean}	R@1	R@10	R@50	R_{mean}
TIRG (Vo et al. 2019)	13.10	30.91	14.13	34.61	14.79	34.37	14.01	33.30	23.66	12.60	45.45	69.39	42.48
RR (Santoro et al. 2017)	18.33	38.63	15.44	38.08	21.10	44.77	18.29	40.49	29.39	12.31	45.10	71.45	42.95
VAL (Chen, Gong, and Bazzani 2020)	22.38	44.15	22.53	44.00	27.53	51.68	24.15	46.61	35.38	17.18	51.52	75.83	48.18
CoSMo (Lee, Kim, and Han 2021)	24.90	49.18	25.64	50.30	29.21	57.46	26.58	52.31	39.45	16.72	48.36	75.64	46.91
CLVC-Net (Wen et al. 2021)	28.75	54.76	29.85	56.47	33.50	64.00	30.70	58.41	44.56	17.64	54.39	79.47	50.50
ARTEMIS (Delmas et al. 2022)	21.78	43.64	27.16	52.40	29.20	54.83	26.05	50.29	38.17	18.72	53.11	79.31	50.38
FashionVLP (Goenka et al. 2022)	30.73	58.02	30.41	57.11	33.67	64.48	31.60	59.87	45.735	-	49.08	77.32	-
AMC (Zhu et al. 2023b)	30.67	59.08	31.73	59.25	36.21	66.60	32.87	61.64	47.255	19.99	56.89	79.27	52.05
DCNet (Kim et al. 2021)	23.95	47.30	28.95	56.07	30.44	58.29	27.78	53.89	40.84	-	53.82	79.33	-
Clip4Cir (Baldrati et al. 2022)	39.99	60.45	33.81	59.40	41.41	65.37	38.32	61.74	50.03	21.42	56.69	81.52	53.21
PLIR (Zhao, Song, and Jin 2022)	39.45	61.78	33.60	58.90	43.96	68.33	39.02	63.00	51.01	22.88	58.83	84.16	55.29
CASE (Levy et al. 2023)	48.48	70.23	47.44	69.36	50.18	72.24	48.79	70.68	59.74	-	-	-	-
TG-CIR (Wen et al. 2023)	52.60	72.52	45.22	69.66	56.14	77.10	51.32	73.09	62.20	25.89	63.20	85.07	58.05
Re-ranking (Liu et al. 2023b)	50.15	71.25	48.14	71.43	55.23	76.80	51.17	73.13	62.15	-	-	-	-
SPRC (Bai et al. 2023b)	55.64	73.89	49.18	72.43	59.35	78.58	54.92	74.97	64.85	-	-	-	-
CIR-LVLM	**58.59**	**75.86**	**50.42**	**73.57**	**59.61**	**78.99**	**56.21**	**76.14**	**66.17**	**31.40**	**70.20**	**88.91**	**63.51**

Table 1: Comparison with the state-of-the-art methods on the Fashion-IQ and Shoes dataset. where R_{mean} indicates the average results across all the metrics. The best results are in boldface, while the second-best results are underlined.

$< /sp >$. Thus, the total soft prompt sequence is:

$$S =< sp > [P_{s_1}], [P_{s_2}], ..., [P_{s_\mathcal{K}}] < /sp > \quad (3)$$

Each instance can be assigned to multiple prompts, and through the prompt pool and query-key matching mechanism, different categories of knowledge can be learned.

Representation in CIR-LVLM. Previous work on the CIR task often uses a bi-directional encoder-only model, taking the representation of the prepended [CLS] token to represent the input. However, due to the causal attention mask in an auto-regressive decoder transformer, only the last token has attended to all tokens in a sequence. To account for this information mismatch, following (Muennighoff 2022), we adopt a position-weighted mean pooling method:

$$V = \sum_{i=1}^{k} w_i LLM([t_1], ..., [t_k])[i], \text{ where } w_i = \frac{i}{\sum_{j=1}^{k} j} \quad (4)$$

where k is the sequence length, $LLM(\cdot)$ represents the LLM decoder and w_i represents the positional weight which ensures that tokens appearing later have higher weights. $\{[t_1], [t_2], ..., [t_k]\}$ is input sequence that integrated with the intent instructions.

Learning Objective. The goal of training our model for CIR is to match the representation V_Q of the hybrid-modality query (I_r, T) with the representation V_{I_t} of the target image I_t. At each iteration, we have a mini-batch $\{(V_Q^{(i)}, V_{I_t}^{(i)})\}_{i=1}^{N_B}$, where $(V_Q^{(i)}, V_{I_t}^{(i)})$ denotes the i-th pair of (hybrid-modality query, target image), and N_B denotes the mini-batch size. Following (Vo et al. 2019), we define the batch-based classification loss for model training:

$$L = \frac{1}{N_B} \sum_{i=1}^{N_B} -\log \frac{\exp(\lambda * Sim(V_Q^{(i)}, V_{I_t}^{(i)}))}{\sum_{j=1}^{N_B} \exp(\lambda * Sim(V_Q^{(j)}, V_{I_t}^{(j)}))} \quad (5)$$

where $Sim(\cdot)$ denotes the cosine similarity function, and λ denotes a temperature parameter.

Experiments

Datasets and Evaluation Metrics

We make performance evaluations on three CIR benchmarks, including two fashion-domain datasets **Fashion-IQ** (Wu et al. 2021) and **Shoes** (Guo et al. 2018), as well as an open-domain dataset **CIRR** (Liu et al. 2021).

Following previous works (Delmas et al. 2022), we adopt the Recall@K (R@K) as the evaluation metric, which refers to the fraction of queries for which the correct item is retrieved among the top K results. We also report R_{mean}, the mean of all R@K values, to evaluate the overall retrieval performance for Fashion-IQ and Shoes datasets. For CIRR dataset, thanks to its unique design, we can additionally report Recall$_{subset}$@K where the task is to retrieve the correct image from six curated samples, and the average score of Recall@5 and Recall$_{subset}$@1 as in (Liu et al. 2021).

Implementation Details

We initialize our models with the Qwen-VL-Chat checkpoint and train on 8 × 80G A800 GPUs. A challenge in fine-tuning LLMs for retrieval is the high GPU memory costs associated with contrastive learning. To address this, we employ recent memory efficiency solutions, including LoRA (Hu et al. 2021), flash attention (Dao 2023), and gradient checkpointing to reduce GPU memory usage.

We train our model for a maximum of 10 epochs with the Adam (Kingma and Ba 2014) optimizer (with the learning rate 6e-4 for the prompt pool and 2e-4 for the rest). The temperature parameter and batch size are tailored for each dataset: 100 and 512 for Fashion-IQ, 130 and 800 for CIRR, and 70 and 512 for Shoes. For the parameters $L_p, \mathcal{K}, M$, we

Method	Recall@K				Recall$_{subset}$@K			Avg.
	K=1	K=5	K=10	K=50	K=1	K=2	K=5	
TIRG	14.61	48.37	64.08	90.03	22.67	44.97	65.14	35.52
CIRPLANT	19.55	52.55	68.39	92.38	39.20	63.03	79.49	45.88
ARTEMIS	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
LF-CLIP	33.59	65.35	77.35	95.21	62.39	81.81	92.02	72.53
CLIP4CIR	38.53	69.98	81.86	95.93	68.19	85.64	94.17	69.09
BLIP4CIR	40.15	73.08	83.88	96.27	72.10	88.27	95.93	72.59
DRA	39.93	72.07	83.83	96.43	71.04	87.74	94.72	71.55
CASE	48.00	79.11	87.25	97.57	75.88	90.58	96.00	77.50
TG-CIR	45.25	78.29	87.16	97.30	72.84	89.25	95.13	75.57
CoVR-BLIP	49.69	78.60	86.77	94.31	75.01	88.12	93.16	76.80
Re-ranking	50.55	81.75	89.78	97.18	80.04	91.90	96.58	80.90
SPRC	51.96	82.12	89.74	97.69	**80.65**	92.31	96.60	81.38
CIR-LVLM	**53.64**	**83.76**	**90.60**	**97.93**	79.12	**92.33**	**96.67**	**81.44**

Table 2: Comparison with the state-of-the-art methods on the CIRR dataset, where Avg. indicates the average results of Recall@5 and Recall$_{subset}$@1. The best results are in boldface, while the second-best results are underlined.

set these to 5, 8, and 45 respectively for the Fashion-IQ and Shoes datasets, and to 5, 12, and 55 for the CIRR dataset.

Comparison with State-of-the-Art Methods

Table 1 presents the comparative results on the **Fashion-IQ** and **Shoes** datasets. It can be observed that: **(1)** Our proposed CIR-LVLM consistently achieves the highest recall across all evaluation metrics on both the Fashion-IQ and Shoes datasets. This performance emphasizes the effectiveness of fine-tuning a LVLM as a user intent-aware encoder, and the significant impact of our intent instructions comprising both the task prompt and instance-specific soft prompt. **(2)** Compared with TG-CIR, which is the best model among the models employing the late-fusion strategy, our model leads to an improvement of 3.97% and 5.46% in terms of the R_{mean} metric on the Fashion-IQ and Shoes dataset, respectively. CIR-LVLM's superior performance is attributed to its ability to automatically select appropriate visual features through the Connector, enabling the LLM to capture and understand the key information in images. **(3)** Compared to the best model among the models employing the early-fusion strategy, i.e. Re-ranking, our model achieves a 4.02% improvement in R_{mean} metric on the Fashion-IQ dataset. This is mainly because CIR-LVLM is more powerful in comprehensively perceiving the information of the reference image and more sensitive to user intent conveyed in the relative caption. **(4)** It is worth noting that, our model still outperforms SPRC across nine evaluation metrics on the Fashion-IQ dataset, even if it applies a similar textual inversion module as CIR-LVLM. This further illustrates the powerful user intent-aware abilities of LVLM in CIR task and the effectiveness of our proposed intent instructions.

When applied to the open-domain dataset CIRR, CIR-LVLM still demonstrates compelling results, summarized in Table 2. Similar conclusions can be drawn from Table 2, indicating that CIR-LVLM can accurately discern user intent in complex and dynamic open-world scenarios, demonstrating good generalization capabilities. However, we note that even though CIR-LVLM has made significant improvements on all other metrics, it fails to outperform SPRC on the Recall$_{subset}$@1 metrics. This shortfall suggests that our model may struggle with distinguishing target images that are semantically or visually similar to the reference image.

Method	Avg.		
	R@10	R@50	R_{mean}
Encoder			
A.1 CLIP	35.78	54.90	45.34
A.2 BLIP-2	43.32	66.39	54.85
Task Prompt			
B.1 No Prompt	53.98	73.91	63.84
B.2 Brief Prompt	54.98	75.10	65.04
B.3 Detailed Prompt	55.82	75.19	65.50
Soft Prompt			
C.1 Universal	54.80	75.10	64.95
C.2 Instance-specific (Ours)	**56.21**	**76.14**	**66.17**

Table 3: Ablation studies on the effects of the encoder and intent instruction. The best results are in boldface.

Inference Efficiency Analysis

Given that CIR models are typically employed in real-time e-commerce scenarios, maintaining acceptable inference speeds is crucial. In this section, we compare the inference speeds of CIR-LVLM with SPRC (Bai et al. 2023b), which uses a lightweight encoder, and CIReVL (Karthik et al. 2023), which also incorporates an LLM as a component. For a fair comparison, we replaced the LLM used in CIReVL with Qwen-7B instead of ChatGPT. We calculated the average time for each model to infer a single query on an A800 GPU, with the following results: SPRC took **0.035s**, CIR-LVLM took **0.08s**, and CIReVL took **1.38s**.

While CIR-LVLM's larger parameter size slightly increases inference time compared to lightweight encoders like SPRC, the difference is negligible in practical applications. In contrast, CIReVL, with a similar parameter size, significantly slows down inference due to its multi-pass decoding design. CIR-LVLM employs a single-pass encoding design, which eliminates the efficiency issues of LVLM, enabling it to be effectively applied to CIR tasks.

Ablation Study

Discussion on Different Retrieval Models

To verify the effectiveness of leveraging the LLM as an encoder, in the first three rows of Table 3, we summarize the recalls of our method across different retrieval model configurations. It can be seen that a more powerful retrieval model leads to higher performance (refer to **A.1-2**) due to the enhanced representation capabilities. Besides, when we replace the LLM in CIR-LVLM with the text encoder of CLIP or BLIP-2, a significant decline in the R_{mean} metrics could be observed. This suggests that compared to regular text encoders, the fine-tuned LLM is more advantageous in discerning user intent and extracting the desired information.

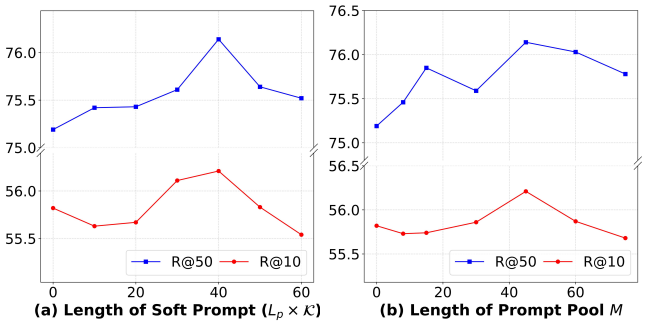


Figure 4: Influence of (a) length of soft prompt and (b) length of prompt pool.

Discussion on Task Prompt

Here we investigate the effectiveness of the guidance from task prompt in Table 3. To minimize noise interference, we opted to exclude the soft prompt from all experiments. As depicted in **B.1-2** of Table 3, when we provide a brief prompt for the model, the model's performance improves significantly than providing no prompt. This suggests that when applying LVLM directly to the CIR task, the model will be confused about the task requirements, cause the requirements for the model are different when processing the query and the target image. Providing the model with a prompt containing the task requirements can alleviate this problem effectively. As shown in **B.3** of Table 3, it's evident that a more detailed task prompt leads to higher performance as it helps the model understand the CIR task more comprehensively. Interestingly, this impact becomes even more pronounced when LLaVA is selected as our backbone model, illustrating the crucial and indispensable role that prompts play when applying LVLMs to CIR tasks.

Discussion on Prompt Pool

We further discuss the effectiveness of the prompt pool. We first compare the performance of our method when applying a universal soft prompt and an instance-specific soft prompt in Table 3 **C.1-2**. It can be observed that the instance-specific soft prompt achieved a significant improvement as it learned subtle differences between each instance. We then discuss the effect of the length of soft prompt ($P = L_p \times \mathcal{K}$) and the length of prompt pool M in Fig.4. The results show that as the increase in P and M, the recall gradually increases, indicating that the additional instance-specific soft prompt for detailed guidance is critical for CIR. Additionally, we can observe from this figure that when $P = 40$ and $M = 45$ our method obtains the highest results, after which it begins to decline due to the excessively long soft prompts or an oversized prompt pool.

Interpretability of CIR-LVLM

In the right columns of the first and third rows of Fig.5, we visualize the attention maps of CIR-LVLM for the images. In the first example, our model successfully focuses its attention on the dress itself while largely ignoring the background. Due to the mention of "have a print" in the relative

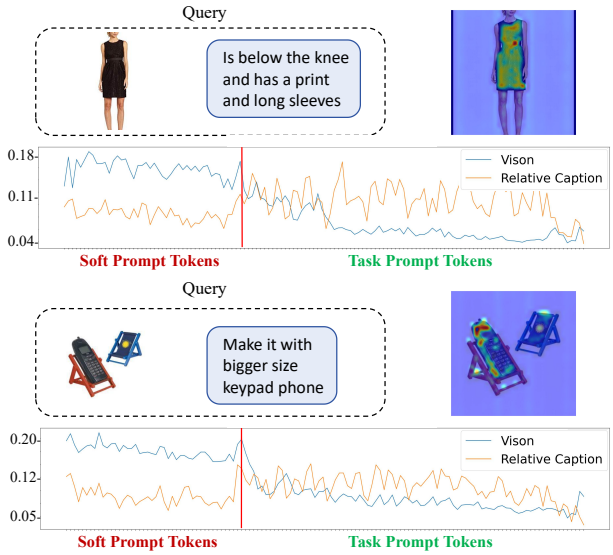


Figure 5: Attention map visualization (right side of the first and third rows) and the sum of the attention weights over all the visual or relative caption tokens for the soft prompt and hard prompt. See Fig. 3 for details of task prompt.

caption, the model pays strong attention to the surface of the dress. Similarly because the relative caption also mentions "long sleeves", the model increases its attention to the arm area. A similar conclusion can be observed in the second example, where CIR-LVLM correctly understands the user's intent and focuses on the desired image regions.

In the line charts of Fig.5, we analyze the attention weights of each token in the soft prompt and task prompt with respect to the overall visual tokens and the overall relative caption tokens. We can observe that the soft prompt places more attentions on the visual tokens, while the task prompt is more concerned about the relative caption. This conclusion is intuitive, as the task prompt primarily explains how to utilize the relative caption to accomplish the task, whereas the recall process of the soft prompt is partially driven by visual features. This finding partly explains the complementary roles these prompts play in CIR, which leads to a more comprehensive understanding of user intent.

Conclusion

The successful application of LVLMs in various visual tasks has initiated interest in retrieval tasks. In this work, we apply LVLMs to the CIR task for the first time as a user intent-aware encoder. Specifically, we utilize LVLM to provide a unified processing framework for composed querying and leverage its superior reasoning capabilities to understand and implement the user's modification intent. Furthermore, we devise novel intent instructions consisting of the task prompt and the instance-specific soft prompt to provide detailed guidance at two levels. Our work shows that in multimodal retrieval tasks that require reasoning, LVLMs have the potential to surpass VLMs, offering new directions for multimodal retrieval.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (62376274, 62437002) and Beijing Natural Science Foundation (L233008).

References

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8948–8957.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023a. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Bai, Y.; Xu, X.; Liu, Y.; Khan, S.; Khan, F.; Zuo, W.; Goh, R. S. M.; and Feng, C.-M. 2023b. Sentence-level prompts benefit composed image retrieval. *arXiv preprint arXiv:2310.05473*.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Del Bimbo, A. 2022. Conditioned and Composed Image Retrieval Combining and Partially Fine-Tuning CLIP-Based Features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4959–4968.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, Y.; Gong, S.; and Bazzani, L. 2020. Image search with text feedback by visiolinguistic attention learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3001–3011.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Dao, T. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Delmas, G.; de Rezende, R. S.; Csurka, G.; and Larlus, D. 2022. ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity. *arXiv preprint arXiv:2203.08101*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems*, 26: 2121–2129.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gao, D.; Jin, L.; Chen, B.; Qiu, M.; Li, P.; Wei, Y.; Hu, Y.; and Wang, H. 2020. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2251–2260.

Goenka, S.; Zheng, Z.; Jaiswal, A.; Chada, R.; Wu, Y.; Hedau, V.; and Natarajan, P. 2022. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14105–14115.

Gordo, A.; Almazán, J.; Revaud, J.; and Larlus, D. 2016. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*, 241–257.

Guo, X.; Wu, H.; Cheng, Y.; Rennie, S.; Tesauro, G.; and Feris, R. 2018. Dialog-based interactive image retrieval. *Advances in Neural Information Processing Systems*, 31: 676–686.

Han, X.; Zhao, W.; Ding, N.; Liu, Z.; and Sun, M. 2022. Ptr: Prompt tuning with rules for text classification. *AI Open*, 3: 182–192.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13418–13427.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Jin, W.; Cheng, Y.; Shen, Y.; Chen, W.; and Ren, X. 2021. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*.

Karthik, S.; Roth, K.; Mancini, M.; and Akata, Z. 2023. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*.

Kim, J.; Yu, Y.; Kim, H.; and Kim, G. 2021. Dual compositional learning in interactive image retrieval. In *AAAI Conference on Artificial Intelligence*, 1771–1779.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, S.; Kim, D.; and Han, B. 2021. Cosmo: Content-style modulation for image retrieval with text feedback. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 802–812.

Levy, M.; Ben-Ari, R.; Darshan, N.; and Lischinski, D. 2023. Data Roaming and Early Fusion for Composed Image Retrieval. *arXiv preprint arXiv:2303.09429*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1096–1104.

Liu, Z.; Rodriguez-Opazo, C.; Teney, D.; and Gould, S. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2125–2134.

Liu, Z.; Sun, W.; Teney, D.; and Gould, S. 2023b. Candidate Set Re-ranking for Composed Image Retrieval with Dual Multi-modal Encoder. *arXiv preprint arXiv:2305.16304*.

Ma, X.; Wang, L.; Yang, N.; Wei, F.; and Lin, J. 2023. Fine-tuning llama for multi-stage text retrieval. *arXiv preprint arXiv:2310.08319*.

Mishra, A.; Shekhar, S.; Singh, A. K.; and Chakraborty, A. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, 947–952. IEEE.

Muennighoff, N. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Muennighoff, N.; Su, H.; Wang, L.; Yang, N.; Wei, F.; Yu, T.; Singh, A.; and Kiela, D. 2024. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*.

OpenAI. 2023. ChatGPT. https://chat.openai.com/.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Saito, K.; Sohn, K.; Zhang, X.; Li, C.-L.; Lee, C.-Y.; Saenko, K.; and Pfister, T. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19305–19314.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. *Advances in Neural Information Processing Systems*, 30: 4974–4983.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Sun, S.; Ye, F.; and Gong, S. 2023. Training-free Zero-shot Composed Image Retrieval with Local Concept Reranking. *arXiv preprint arXiv:2312.08924*.

Tang, Y.; Yu, J.; Gai, K.; Zhuang, J.; Xiong, G.; Hu, Y.; and Wu, Q. 2024. Context-I2W: Mapping Images to Context-dependent Words for Accurate Zero-Shot Composed Image Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5180–5188.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.-J.; Fei-Fei, L.; and Hays, J. 2019. Composing text and image for image retrieval-an empirical odyssey. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6439–6448.

Wang, J.; Zhou, Y.; Xu, G.; Shi, P.; Zhao, C.; Xu, H.; Ye, Q.; Yan, M.; Zhang, J.; Zhu, J.; et al. 2023. Evaluation and Analysis of Hallucination in Large Vision-Language Models.(Aug. *arXiv preprint arxiv:2308.15126*.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 139–149.

Wen, H.; Song, X.; Yang, X.; Zhan, Y.; and Nie, L. 2021. Comprehensive linguistic-visual composition network for image retrieval. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 1369–1378.

Wen, H.; Zhang, X.; Song, X.; Wei, Y.; and Nie, L. 2023. Target-guided composed image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, 915–923.

Wu, H.; Gao, Y.; Guo, X.; Al-Halah, Z.; Rennie, S.; Grauman, K.; and Feris, R. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11307–11317.

Zang, Y.; Li, W.; Zhou, K.; Huang, C.; and Loy, C. C. 2022. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*.

Zhao, Y.; Song, Y.; and Jin, Q. 2022. Progressive Learning for Image Retrieval with Hybrid-Modality Queries. *arXiv preprint arXiv:2204.11212*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhu, H.; Wei, Y.; Zhao, Y.; Zhang, C.; and Huang, S. 2023b. Amc: Adaptive multi-expert collaborative network for text-guided image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6): 1–22.