Manipulation Intention Understanding for Zero-Shot Composed Image Retrieval

Yuanmin Tang[1,2], **Jing Yu**[3,4*], **Keke Gai**[5,6*] **Gang Xiong**[1,2], **Gaopeng Gou**[1,2], **Meikang Qiu**[7], **Qi Wu**[8]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China
[4]School of Information Engineering, Minzu University of China
[5]School of AI, Beijing Institute of Technology, Beijing 100081, China
[6]Zhongguancun Academy, Haidian, Beijing, China
[7]School of Computer and Cyber Sciences Augusta University Augusta, Georgia, USA
[8]Responsible AI Research Centre, Adelaide University
tangyuanmin@iie.ac.cn, jing.yu@muc.edu.cn, gaikeke@bit.edu.cn, qiumeikang@yahoo.com, qi.wu01@adelaide.edu.au

Abstract

Zero-shot Composed Image Retrieval (ZS-CIR) involves diverse tasks with varied visual manipulation intents across domains, scenes, objects, and attributes. A key challenge is that existing datasets contain limited intent-relevant annotations, making it hard for models to infer human intent from textual modifications. We introduce an intent-centric image–text dataset generated via reasoning by a Multimodal Large Language Model (MLLM) to better train ZS-CIR models for human manipulation intent understanding. Building on this dataset, we propose De-MINDS, a framework that distills the MLLM's reasoning ability to capture manipulation intent and enhance models' comprehension of modified text. A simple mapping network translates image information into language space and combines it with the manipulation text to form a query. De-MINDS then extracts intention-relevant information from this query and encodes it as pseudo-word tokens for accurate ZS-CIR. Across four ZS-CIR tasks, De-MINDS shows strong generalization and improves over existing methods by 2.15% to 4.05%, establishing new state-of-the-art results with comparable inference time.

Dataset — https://github.com/Pter61/De-MINDS

Introduction

Composed Image Retrieval (CIR) (Vo et al. 2019) aims to retrieve a target image that is visually similar to a reference image while incorporating modifications specified by user-provided manipulation text. Unlike traditional image retrieval (Datta et al. 2008), which relies on single-modality features, CIR leverages both visual and textual data to capture user intent more accurately, as shown in Figure 1(c). This dual-modality approach allows users to specify desired changes to reference images, improving search precision and enabling a clearer articulation of user intent.

CIR faces two primary challenges: (1) user intent spans both visual and textual modalities, necessitating a shared semantic space for cross-modal reasoning; (2) interpreting user intent requires deep reasoning since intent is frequently conveyed implicitly, especially through reference images. Various supervised methods address these challenges (Liu et al. 2021; Baldrati et al. 2022) by leveraging extensively annotated triplets to train specialized retrieval models. However, these supervised approaches are labor-intensive and exhibit limited generalizability.

To mitigate these issues, recent advancements introduce Zero-Shot Composed Image Retrieval (ZS-CIR), utilizing pre-trained CLIP models to convert ZS-CIR into traditional text-based retrieval tasks. As depicted in Figure 1(c), these methods map reference images into CLIP's language space as pseudo-tokens, combining them with manipulation text queries to perform zero-shot retrieval via semantic similarity. Despite progress, existing CLIP-based mapping net-

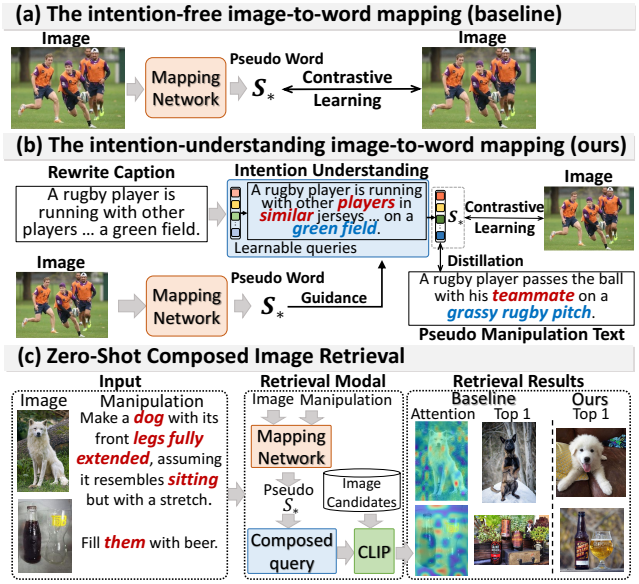

Figure 1: Illustration of our motivation. (a) Intention-free visual mapping. (b) Our intention-understanding visual mapping. (c) ZS-CIR process results from different strategies.

*Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

works struggle to interpret implicit user intentions embedded in manipulation texts accurately. For instance, in Figure 1(c), intents like "make a dog sitting with extended legs" or "fill brown and clear bottles with beer" are implicit and challenging for frozen CLIP models (Tang et al. 2023). Consequently, these models often fail to effectively filter manipulation-relevant information, resulting in inaccurate retrieval outcomes.

In this work, to address this critical gap, we propose **intent-CC3M**, an intention-based dataset designed for training mapping networks to align intention-relevant visual information within the language space. As illustrated in Figure 1, intent-CC3M enhances the widely-used CC3M dataset (Sharma et al. 2018; Saito et al. 2023) with rewritten, redundancy-reduced captions and condensed pseudo-manipulation texts. These refined descriptions, indicative of potential manipulation intents, are generated through chain-of-thought prompting using a Multimodal Large Language Model (MLLM), enabling intent-specific mapping.

Furthermore, we introduce **De-MINDS**, a novel, efficient approach for *un**DE**rstanding **M**anipulation **IN**tention from target **D**escription before **S**earching* for accurate ZS-CIR. De-MINDS leverages rewritten captions and pseudo-manipulation texts to distill MLLM reasoning, interpreting implicit intents guided by reference images. By converting intentions embedded in redundant (*e.g.,* "players in similar orange jersey" for "teammates") or abstract texts (*e.g.,* "green field" for "grassy rugby pitch") into multiple learnable embeddings, De-MINDS effectively enhances CLIP's comprehension of user-specified modifications, significantly improving retrieval accuracy.

Our main contributions are summarized as follows: (1) We introduce **intent-CC3M**, a novel dataset containing pseudo-manipulation texts generated by MLLM reasoning, bridging the gap between pre-training and retrieval in ZS-CIR models. Experiments confirm significant performance gains when baseline models are trained using intent-CC3M. (2) We propose **De-MINDS**, which explicitly extracts manipulation intents under reference image context and represents them as learnable embeddings for improved composed retrieval, offering an intent-aware perspective on image retrieval. (3) Extensive experiments across four ZS-CIR tasks demonstrate that De-MINDS consistently improves CIR performance by 2.15% to 4.05%, achieving new state-of-the-art results with comparable inference efficiency and strong potential for vision-and-language applications.

Related Works

Composed Image Retrieval. Composed Image Retrieval (CIR) combines image and text features for retrieval (Vo et al. 2019), typically using late fusion to integrate visual and language features separately while requiring extensively annotated triplets CIR datasets. (Baldrati et al. 2022; Liu et al. 2021; Zhang et al. 2024). Zero-shot CIR models (Saito et al. 2023; Baldrati et al. 2023; Tang et al. 2024b; Gu et al. 2024; Suo et al. 2024; Li, Ma, and Yang 2025; Yang et al. 2024; Shi et al. 2025; Tang et al. 2024a, 2025b,c,a), mitigate the need for large CIR datasets. The projection-based ZS-CIR methods map reference images into the text space for query construction. However, these methods rely on the pre-trained CLIP, which struggles to understand intentions within manipulation text. To tackle this issue, we propose a novel model that effectively understands these intentions, thereby improving the ZS-CIR model's ability to retrieve images based on manipulation intents accurately. Unlike training-free CIR methods like CIReVL (Karthik et al. 2024) that apply LLMs during inference, introducing non-negligible computational overhead, our model leverages a specific design based on the reasoning capabilities of MLLMs. This design improves accuracy and efficiency without using LLMs during inference, maintaining comparable inference times.

Vision and Language Pre-training Models. Vision and Language Pre-training (VLP) models, like CLIP (Radford et al. 2021), leverage extensive image-text pair training to achieve implicit alignment. Recent VLP advancements (Zhou et al. 2022; Song et al. 2022) utilize static models to integrate encoded image and text features, enabling various zero-shot tasks (Li et al. 2022; Song et al. 2022; Shi et al. 2023). However, current CLIP-based zero-shot learning struggles with manipulation description in CIR tasks, motivating our approach, which enhances CLIP's capabilities of understanding user intentions to modify from fine-grained/long descriptions. Inspired by these, In our work, we utilize multiple learnable queries to guide the extraction of intentions from manipulation text, providing explanatory cues for more accurate ZS-CIR.

Image-text Dataset Enhancement. In the field of vision-language learning, various endeavors (Fan et al. 2024; Lai et al. 2023; Gadre et al. 2024; Nguyen et al. 2024; Chen et al. 2023) aim to enhance caption quality within existing image-text datasets. LaCLIP (Fan et al. 2024) utilizes LLMs to refine raw captions. VeCLIP (Lai et al. 2023) integrates from raw and synthetic sources using LLMs. The latest approach, ShareGPT4V (Chen et al. 2023), leverages MLLMs to generate descriptive captions from deliberate prompts and corresponding image inputs. However, these methods ignore user intentions, which are crucial for CIR tasks. To bridge this gap, we introduce a novel dataset with pseudo-manipulation intentions reasoned by MLLMs.

Methodology

Given a reference image space $\mathcal{I}_r$ and a text description space $\mathcal{T}$, Zero-Shot Composed Image Retrieval (ZS-CIR) (Saito et al. 2023; Baldrati et al. 2023; Tang et al. 2024b) involves user-provided manipulation text $T \in \mathcal{T}$ that describes hypothetical semantic modifications to a reference image $I_r \in \mathcal{I}_r$. The goal is to retrieve a target image from an image database $\mathcal{D} = \{I_1, \ldots, I_n\}$ that reflects the intended context. ZS-CIR methods typically learn a mapping function $f_\theta : \mathcal{I} \to \mathcal{Z}$, where $\mathcal{Z}$ is a predefined text-token embedding space. This function uses intermediate image features from a pre-trained image encoder Ψ_I. The pseudo token embedding $\boldsymbol{S}_* = f_\theta(\Psi_I(I_r))$ is composed with the manipulation text into a natural language prompt P, such as "`a photo of` S_*, $\{T\}$". This query is encoded using a pre-trained text encoder Ψ_T, and retrieval is $\cos(\Psi_I(I_r), \Psi_T(P))$ conducted by computing the cosine similarity.

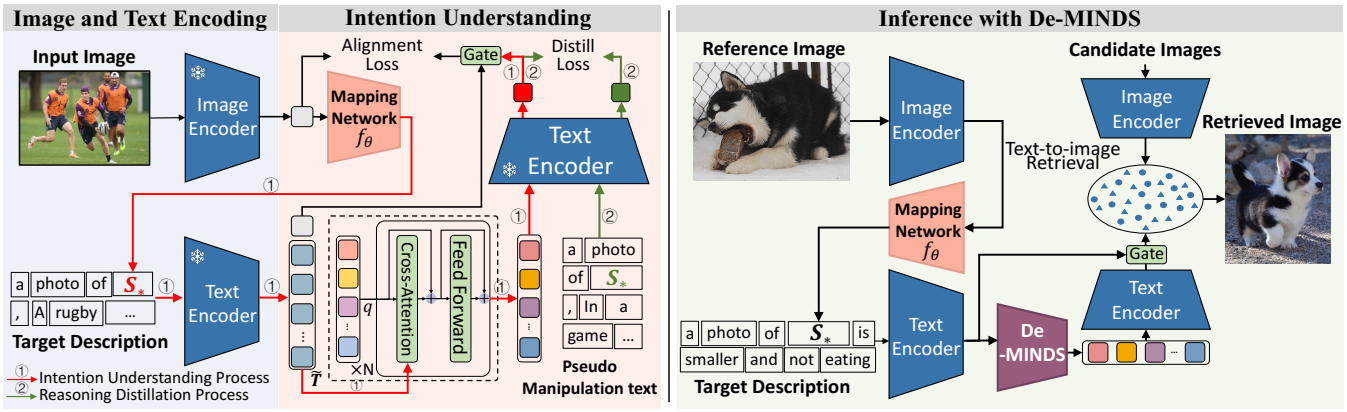

Figure 2: An overview of our De-MINDS. Pre-training (left): Map the image to a pseudo token S_*, then understand the intention. Inference (right): Map the inference image to S_* to construct the target description and understand user intention.

Creating Intention-based Image-text Dataset

To interpret implicit user intentions in manipulation text, we construct an intention-based image-text dataset. To minimize annotation bias and ensure a fair comparison, we augment the popular CC3M dataset using LLaVA (Liu et al. 2024), an open-source Multi-modal Large Language Model (MLLM) with strong vision-language reasoning capabilities. However, accurately inferring manipulation intent from image-text pairs remains challenging due to the limited reasoning capability and hallucinations introduced by LLaVA-generated captions. Specifically, we divide the reasoning process of pseudo-manipulation texts into three stages:

(1) Caption Rewriting. We first leverage MLLMs to capture hidden intentions by rewriting original captions into multi-view visual descriptions suitable for CIR tasks. Given the i-th image I_r^i and its original caption T_{ori}^i from CC3M: $\mathcal{D} = \{(I_r^i, T_{ori}^i), \ldots, (I_r^n, T_{ori}^n)\}$, we prompt LLaVA with a rewriting prompt p_{rew} to generate rewritten captions as:

$$T_{rew}^i = \text{MLLM}(I_r^i, T_{ori}^i, p_{rew}), \tag{1}$$

averaging 65 tokens covering various visual aspects such as objects, scenes, attributes, and domains.

(2) Intention Reasoning. The rewritten captions often contain redundant details irrelevant to manipulation intent (illustrated in Figure 5). To effectively filter these irrelevant details, we again prompt LLaVA with an intention reasoning prompt p_{int} to explicitly infer potential human manipulation intentions from each rewritten caption, formally denoted as:

$$T_{int}^i = \text{MLLM}(I_r^i, T_{rew}^i, p_{int}), \tag{2}$$

This operation resulting in our pseudo-manipulation texts T_{int}^i, averaging 27 tokens.

(3) Similarity-guided Filtering. However, given the limited reasoning ability of LLaVA, some generated pseudo-manipulation texts contain inaccuracies or hallucinations. Thus, we adopt a quality control step, incorporating a similarity-based filtering mechanism. Specifically, we measure the semantic similarity $\text{sim}(I_r^i, T_{int}^i)$ between each reference image I_r^i and the pseudo-manipulation text T_{int}^i using CLIP embeddings as follow:

$$\text{sim}(I_r^i, T_{int}^i) = \cos(\Psi_I(I_r^i), \Psi_T(T_{int}^i)). \tag{3}$$

Image and pseudo-manipulation text pairs with similarity scores below 0.7 undergo up to three regeneration attempts (ablation in supplementary). If similarity remains low, pairs are forwarded to GPT-4o (OpenAI 2022) for higher-quality text generation (manual evaluation in supplementary). Although GPT-4o consistently produces superior texts, applying it directly to all 3M pairs is economically impractical. Therefore, we strategically integrate open-source LLaVA with GPT-4o, leveraging CLIP's semantic knowledge and GPT-4o's robust reasoning capabilities. The resulting dataset is: $\tilde{\mathcal{D}} = \{(I_r^i, T_{ori}^i, T_{rew}^i, T_{int}^i)\}_{i=1}^n$. Detailed prompt templates (*i.e.*, p_{rew} and p_{int}) are provided in the supplementary materials due to space constraints.

Intention-based Image-to-Text Mapping

Since ZS-CIR methods rely on CLIP's text encoder, implicit manipulation intentions in user-provided text pose significant challenges. To effectively interpret these intentions, we propose two complementary modules: the *Manipulation Intention Understanding* module, capturing manipulation intents into pseudo tokens, and the *Reasoning Distillation* module, which distills MLLM-level reasoning capabilities.

Image and Text Encoding. Given a sample $(I_r, T_{ori}, T_{rew}, T_{int})$ from intent-CC3M, we encode image features with CLIP's frozen image encoder Ψ_I, obtaining $\boldsymbol{v} = \Psi_I(I_r) \in \mathbb{R}^d$. A three-layer fully connected network f_θ maps $\boldsymbol{v}$ to a pseudo-token embedding $S_* = f_\theta(\boldsymbol{v})$. We construct a query description P as "`a photo of` S_*, $\{T\}$". As illustrated in Figure 1(b), we consider two intent-understanding scenarios: (1) explicit intention from rewritten captions and (2) implicit intention from original captions. Accordingly, in each training batch, the text T is sampled as 50% original captions T_{ori}, 30% rewritten captions T_{rew}, and 20% pseudo-manipulation texts T_{int}, balancing intent modeling and stabilizing training. The query P is then encoded by the CLIP text encoder Ψ_T, producing token features $\boldsymbol{T} = \{\boldsymbol{t}_i\}_{i=1}^m \subseteq \mathbb{R}^{d \times m}$. The first token $\boldsymbol{t}_1$ is the `[CLS]` embedding summarizing global image–caption context, and the remaining tokens $\tilde{\boldsymbol{T}} = \{\boldsymbol{t}_i\}_{i=2}^m$ represent individual word embeddings.

Manipulation Intentions Understanding. Given the word embeddings of the query descriptions, this module aims to capture different intentions from the manipulation text, thereby enhancing the CLIP text encoder's capability. To capture different intentions, we introduce a set of learnable query embeddings for guidance, denoted as $\boldsymbol{X} = \{\boldsymbol{x}_k\}_{k=1}^{n} \in \mathbb{R}^{d\times n}$, where d is the embedding dimension and n is the number of queries. Each query $\boldsymbol{x}_k$ represents a manipulation intention. As depicted in Figure 2(left), we implement cross-attention mechanisms to extract intention-relevant contextual information from the word embeddings $\tilde{\boldsymbol{T}} = \{\boldsymbol{t}_i\}_{i=2}^{m}$ using the learnable queries $\boldsymbol{X}$. The cross-attention operation involves three primary steps. First, we compute the query, key and value through linear projections, *i.e.,* $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}^Q$, $\boldsymbol{K} = [\boldsymbol{X}, \tilde{\boldsymbol{T}}]\boldsymbol{W}^K$, $\boldsymbol{V} = [\boldsymbol{X}, \tilde{\boldsymbol{T}}]\boldsymbol{W}^V$. $[\boldsymbol{X}, \tilde{\boldsymbol{T}}]$ denotes concatenating the two matrices, which enhances the interaction between learnable queries and word embeddings with better performance. Then, the learnable queries from the current cross-attention block $\boldsymbol{X}^i$ is:

$$\boldsymbol{X}_{att}^i = \mathrm{Att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \textit{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}}\right)\boldsymbol{V}, \quad (4)$$

$$\boldsymbol{X}^i = \mathrm{FFW}(\boldsymbol{X}_{att}^i + \boldsymbol{X}^{i-1}) + \boldsymbol{X}_{att}^i \quad (5)$$

where $\boldsymbol{X}^{i-1}$ are learnable queries from the previous block and $\mathrm{FFW}(\cdot)$ denotes 2-layer feed-forward networks. the refined query embeddings $\boldsymbol{X}$ are then fed into the frozen text encoder Ψ_T of CLIP to extract the intention embedding as $\boldsymbol{t}_* = \Psi_T(\boldsymbol{X}^n) = \{t_*^i\}_{i=1}^{d} \in \mathbb{R}^{d\times 1}$ ($d = 768$).

Reasoning Distillation. Given the intention embedding $\boldsymbol{t}_*$, the AI agent needs to further align with human manipulation intention for distilling the reasoning capability of MLLM. Specifically, we aim to reduce the distance between the intention embedding and the corresponding pseudo-manipulation text's `[CLS]` embedding, representing the MLLM's intention embedding while ensuring each embedding remains distinct and discriminative. Given the intention embeddings $\mathcal{T}_{int} = \{\boldsymbol{t}_*^i\}_{i=1}^{N}$, where N is the number of images in $\tilde{D}$, and the corresponding MLLM's intention embeddings $\tilde{\boldsymbol{t}}_* = \Psi_T(T_{int}) \in \tilde{\mathcal{T}}_{int}$ we employ a symmetric contrastive loss inspired by SimCLR (Chen et al. 2020; Saito et al. 2023) as follows:

$$\mathcal{L}_{distil} = \mathcal{L}_{s2t}(\boldsymbol{t}_*, \tilde{\boldsymbol{t}}_*) + \mathcal{L}_{t2s}(\tilde{\boldsymbol{t}}_*, \boldsymbol{t}_*) \quad (6)$$

The two contrastive loss terms are defined as:

$$\mathcal{L}_{s2t}(\boldsymbol{t}_*, \tilde{\boldsymbol{t}}_*) = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\frac{e^{\tau(\boldsymbol{t}_*^i)^T\tilde{\boldsymbol{t}}_*^i}}{\sum_{j\in\mathcal{B}} e^{\tau(\boldsymbol{t}_*^i)^T\tilde{\boldsymbol{t}}_*^j}}, \quad (7)$$

$$\mathcal{L}_{t2s}(\boldsymbol{t}_*, \tilde{\boldsymbol{t}}_*) = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\frac{e^{\tau(\tilde{\boldsymbol{t}}_*^i)^T\boldsymbol{t}_*^i}}{\sum_{j\in\mathcal{B}} e^{\tau(\tilde{\boldsymbol{t}}_*^i)^T\boldsymbol{t}_*^j}} \quad (8)$$

where B is the number of images in a batch and τ is a temperature hyper-parameter that controls the strength of penalties on hard negative samples.

Cross-Modal Alignment. Given the embedding of user intention, this module aims to form a query embedding optimized for retrieval. Since the nature of CIR, both the reference image and the manipulation intention form a comprehensive context that defines the target image. To dynamically control the influence of manipulation intentions on the retrieval process, we introduce a learnable scalar $gate$ that decides the contribution of the manipulation intention information $\boldsymbol{t}_*$ and integrates the global information $\boldsymbol{t}_{cls}$ to form the final query embedding $\hat{\boldsymbol{t}}$ as follows:

$$\hat{\boldsymbol{t}} = \boldsymbol{t}_{cls} + gate \cdot \boldsymbol{t}_*$$

Then, we aim to match a target image to its paired query embedding while separating unpaired ones. We minimize the symmetric contrastive loss between the image embedding $\boldsymbol{v}$ and the target embedding $\hat{\boldsymbol{t}}$ as follows:

$$\mathcal{L}_{align} = \mathcal{L}_{s2t}(\hat{\boldsymbol{t}}, \boldsymbol{v}) + \mathcal{L}_{t2s}(\boldsymbol{v}, \hat{\boldsymbol{t}}) \quad (9)$$

where $\mathcal{L}_{s2t}$ and $\mathcal{L}_{t2s}$ are two contrastive loss terms as Eq.8. The final loss used to optimize is:

$$\mathcal{L} = \mathcal{L}_{distill} + \mathcal{L}_{align} \quad (10)$$

Inference with De-MINDS. In the inference stage, we compose the reference image with the paired manipulation text and compare the composed query with candidate images for retrieval. As shown in Figure 2 (right), we compose the pseudo token embedding $\boldsymbol{S}_*$ of the image from the mapping network with the manipulation text and feed it to the pre-trained text encoder of CLIP. The result is embedded by the tex encoder and compared to the visual features of candidate images.

Since we focus on studying the manipulation intention understanding searching for ZS-CIR, we utilize the same prompt in the most recent works (Saito et al. 2023; Tang et al. 2024b) for a fair comparison. Since we focus on studying the manipulation intention understanding searching for ZS-CIR, we utilize the same prompt in the most recent works (Saito et al. 2023; Tang et al. 2024b) for a fair comparison. We show prompt examples for different ZS-CIR tasks. In all examples, `[*]` indicates the pseudo token from the mapping network: **(a) Domain conversion** aims to modify the domain of the reference image. The prompt is defined as `a [domain tag] of [*]`; **(b) Sentence manipulation** modifies the reference image based on a sentence. We simply append the sentence with the special token as `a photo of [*], [sentence].`

Experiments

Datasets. We evaluate our model on four ZS-CIR datasets, *i.e.,* ImageNet (Deng et al. 2009; Hendrycks et al. 2021) for domain conversion, CIRR (Liu et al. 2021) and CIRCO (Baldrati et al. 2023) for object/scene manipulation, and Fashion-IQ (Wu et al. 2021) for attribute manipulation. All the dataset settings and evaluation metrics (Recall@K and mAP@K)) follow the recent works (Saito et al. 2023; Tang et al. 2024b; Gu et al. 2024) for a fair comparison.

(1) Domain conversion. This dataset contains 16,983 images from 200 classes across four domains (*cartoon*, *origami*, *toy*, *sculpture*). We adopt prompt (a) at inference. (2) Object/scene manipulation. Each reference image specifies how to modify the object or background. We use prompt

Backbones	Methods	Conferences	Dress		Shrit		TopTee		Average	
			R10	R50	R10	R50	R10	R50	R10	R50
ViT-L/14	Pic2Word†	CVPR 2023	20.0	40.2	26.2	43.6	27.9	47.4	24.7	43.7
	CIReVL†	ICLR 2024	24.6	44.8	29.5	47.4	31.4	53.7	28.6	48.6
	LinCIR†	CVPR 2024	20.9	42.4	29.1	46.8	28.8	50.2	26.3	46.5
	PrediCIR†	CVPR 2025	25.4	49.5	31.8	52.0	33.1	55.4	30.1	52.3
	SEARLE-XL†	ICCV 2023	20.3	43.2	27.4	45.7	29.3	50.2	25.7	46.3
	SEARLE-XL*	–	24.3	46.4	30.9	49.5	31.5	53.1	28.9	49.7
	Context-I2W†	AAAI 2024	23.1	45.3	29.7	48.6	30.6	52.9	27.8	48.9
	Context-I2W*	–	25.3	48.5	31.7	51.4	32.6	55.3	29.9	51.7
	De-MINDS	–	**27.7**	**51.1**	**33.6**	**53.9**	**35.3**	**58.5**	**32.2**	**54.5**
ViT-G/14	CIReVL†	ICLR 2024	27.1	49.5	26.9	45.6	29.3	50.0	25.6	46.2
	LinCIR†	CVPR 2024	38.1	60.9	46.8	65.1	50.5	71.1	45.1	65.7
	PrediCIR†	CVPR 2025	39.7	62.4	48.2	67.4	53.7	73.6	47.2	67.8
	De-MINDS	–	**41.5**	**64.8**	**50.5**	**70.3**	**55.8**	**76.3**	**49.5**	**70.5**

Table 1: Results on Fashion-IQ for attribute manipulation. †indicates results from the original paper.

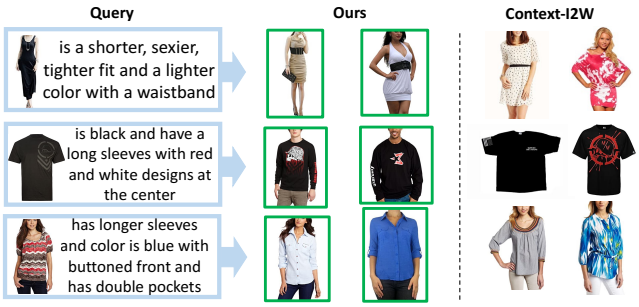

Figure 3: Results of attribute manipulation on FashionIQ.

(b) at inference. (3) Attribute manipulation. This dataset provides diverse descriptions for attribute editing. We also use prompt (b) at inference.

Implementation Details. Generating pseudo-manipulation texts with LLaVA-1.6-13B (Liu et al. 2024) and GPT-4o (OpenAI 2022) on CC3M (Sharma et al. 2018) costs about 649 GPU-hours on five A100 (80G) GPUs and $1,227. For De-MINDS, we use ViT-L/14 CLIP (Radford et al. 2021) and ViT-G/14 from OpenCLIP (Ilharco et al. 2021), trained with AdamW (Loshchilov and Hutter 2018) (learning rate 1×10^{-6}, weight decay 0.1). We set 4 learnable queries and a contrastive batch size of 1024, and initialize the tanh-gating scalar to 0 for stability (Bachlechner et al. 2021). Context-I2W and SEARLE are trained with their original settings, replacing only captions with our pseudo-manipulation texts. All models are trained on four NVIDIA A100 (80G) GPUs, and we report averages over three runs.

Quantitative and Qualitative Results

We compare De-MINDS with several commonly benchmarked ZS-CIR methods, including: 1) **Pic2Word** (Saito et al. 2023): Maps the visual features of a reference image into a pseudo-word token; 2) **SEARLE-XL** (Baldrati et al. 2023): Integrating the pseudo-word token with the caption generated by GPT (Brown et al. 2020); 3) **Context-I2W** (Tang et al. 2024b): Selectively extracts text-relevant visual information from the reference image before mapping; 4) **CIReVL** (Karthik et al. 2024): leverage LLM that infers a target image caption; and 5) **LinCIR** (Gu et al. 2024): Masks subjects in captions for training. 6) **PrediCIR** (Tang et al. 2025b): Predict the target image feature during inference. Since most baselines reported their results on ViT-L/14, we mainly compare results on this backbone and explore the generalization ability of De-MINDS on ViT-G/14.

Overall, baseline models trained with intent-CC3M show consistent performance improvements, and De-MINDS surpasses current ZS-CIR models on both ViT-L/14 and ViT-G/14 backbones. Tables 1 to 4 present the quantitative results, while Figures 3 to 4 display the qualitative results of our model and the most relevant work, Context-I2W. The attribute manipulation task requires accurately localizing specific attributes within the entire image. As demonstrated in Table 1, De-MINDS outperforms existing ZS-CIR models significantly, achieving an average improvement of 2.15% over the State-of-the-Art (SoTA) model, PrediCIR. This improvement can be attributed to De-MINDS' one-stage intention reasoning process, which captures complete reference image content to extract fashion-relevant intentions from manipulation text in CIR. This approach enables more efficient fine-grained intention understanding compared to LLM-based CIR methods (*i.e.,* CIReVL) that rely on captions generated by image captioners, which often lose critical fine-grained visual details. Figure 3 further illustrates how De-MINDS effectively understand complex fashion-relevant attributes in manipulation descriptions, such as a sexier style with a waistband (row 1), black color with a special design in the center (row 2), and longer sleeves with two pockets in blue (row 3), facilitating more accurate searching.

We further assess De-MINDS' capability in foreground/background differentiation and fine-grained image editing through the object/scene manipulation task (Table 3). De-MINDS achieves an average performance improvement of 3.17% on CIRR and 2.43% on CIRCO over each best methods on ViT-L/14. This enhancement is due to De-MINDS' use of the MLLM's distilled reasoning capabilities to extract intention from manipulation text before retrieval,

Backbones	Methods	Conferences	Cartoon		Origami		Toy		Sculpture		Average	
			R10	R50	R10	R50	R10	R50	R10	R50	R10	R50
ViT-L/14	Pic2Word†	CVPR 2023	8.0	21.9	13.5	25.6	8.7	21.6	10.0	23.8	10.1	23.2
	LinCIR	CVPR 2024	9.4	24.2	15.7	26.9	10.8	27.0	11.7	27.9	11.9	26.5
	PrediCIR†	CVPR 2025	14.2	31.9	20.4	34.3	14.7	30.8	16.3	34.9	16.4	33.1
	SEARLE-XL	ICCV 2023	9.6	24.9	16.1	27.3	7.6	25.4	11.3	26.4	11.2	26.0
	SEARLE-XL*	–	11.4	26.9	17.1	29.3	10.0	28.0	13.1	28.9	12.9	28.3
	Context-I2W†	AAAI 2024	10.2	26.1	17.5	28.7	11.6	27.4	12.1	28.2	12.9	27.6
	Context-I2W*	–	12.3	28.4	19.8	31.4	13.6	30.8	14.8	32.5	15.1	30.8
	De-MINDS	–	**17.4**	**35.3**	**24.4**	**38.8**	**18.8**	**35.9**	**20.7**	**39.0**	**20.3**	**37.3**
ViT-G/14	LinCIR	CVPR 2024	13.7	30.2	19.5	32.9	13.8	30.2	15.2	34.0	15.5	31.8
	PrediCIR†	CVPR 2025	15.6	34.6	23.7	37.2	17.2	37.5	19.3	37.8	19.0	36.8
	De-MINDS	–	**19.6**	**38.3**	**27.4**	**41.0**	**20.8**	**40.9**	**22.8**	**41.2**	**22.7**	**40.4**

Table 2: Results on ImageNet for domain conversion. †indicates results from the original paper.

Backbones	Methods	R1	R5	R10
ViT-L/14	Pic2Word†	23.9	51.7	65.3
	CIReVL†	24.6	52.3	64.9
	LinCIR†	25.0	53.3	66.7
	PrediCIR†	27.2	57.0	70.2
	SEARLE-XL†	24.2	52.4	66.3
	SEARLE-XL*	25.8	53.9	68.1
	Context-I2W†	25.6	55.1	68.5
	Context-I2W*	26.5	56.2	69.3
	De-MINDS	**30.0**	**59.7**	**74.5**
ViT-G/14	CIReVL†	34.7	64.3	75.1
	LinCIR†	35.3	64.7	76.1
	PrediCIR†	37.0	66.1	77.9
	De-MINDS	**40.3**	**69.6**	**79.8**

Table 3: Results on CIRR for object manipulation.

Backbones	Methods	mAP@5	mAP@10	mAP@25	mAP@50
ViT-L/14	Pic2Word	8.7	9.5	10.6	11.3
	LinCIR†	12.6	13.6	15.0	15.9
	CIReVL†	18.6	19.0	20.9	21.8
	PrediCIR†	15.7	17.1	18.6	19.3
	SEARLE-XL†	11.7	12.7	14.3	15.1
	SEARLE-XL*	13.8	15.4	16.2	17.4
	Context-I2W	13.0	14.6	16.1	17.2
	Context-I2W*	15.5	17.7	18.8	19.9
	De-MINDS	**21.0**	**22.4**	**22.7**	**23.9**
ViT-G/14	LinCIR†	19.7	21.0	23.1	24.2
	CIReVL†	26.8	27.6	30.0	31.0
	PrediCIR†	23.7	24.6	25.4	26.0
	De-MINDS	**30.8**	**32.2**	**33.9**	**34.3**

Table 4: Results on CIRCO for object manipulation.

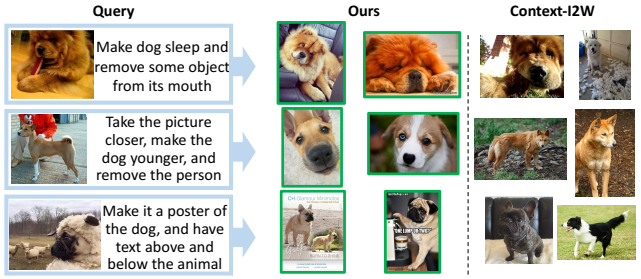

Figure 4: Results of the object manipulation on CIRR

strengthening CLIP's ability to interpret modification intentions accurately. In Figure 4, De-MINDS accurately understands manipulation intentions to remove an overlapping object (row 1), adjust the camera focus (row 2), and modify the style of an image (row 3).

Moreover, in the domain conversion results (Table 2), De-MINDS consistently outperforms existing approaches and notably surpasses the SoTA PrediCIR by an average of 4.05%. It contributes to De-MINDS accurately understanding the intention embedded in abstract texts and maps objects. In contrast, PrediCIR struggles to predict the target features due to its challenge in understanding abstract intention from the user query.

Ablation Study

Following (Saito et al. 2023; Baldrati et al. 2023; Tang et al. 2024b), we evaluate De-MINDS components on CIRR and FashionIQ using ViT-L/14 (Table 5).**(1) Significance of intent-CC3M (models '2-4').** Removing pseudo-manipulation texts (model '2') causes a 4.60% average drop, highlighting intent-CC3M's effectiveness in intention alignment. Without chain-of-thought prompting (model '3') reduces performance by 4.14%, showing CoT's critical role. Without similarity-guided filtering (model '4'), performance decreases by 3.64%, confirming semantic-based filtering's importance. **(2) Key De-MINDS modules (models '5-7').** Excluding De-MINDS intention embeddings (model '5') leads to a 4.80% performance drop, emphasizing their significance. Removing global features (model '6') results in a 2.72% decline, demonstrating the necessity of global context. Direct summation without adaptive gating (model '7') causes a 1.94% drop, highlighting adaptive fusion's importance. **(3) Training strategies (models '8-10').** Using only original captions (model '8') reduces training stability, causing a 2.02% decline. Overlooking distillation loss (model '9') or replacing it with cosine distillation (model '10') reduces performance by 4.32% and 1.94%, respectively, underlining symmetric contrastive loss's efficacy. **(4) Alternative solutions (models '11-14').** Employing LongCLIP

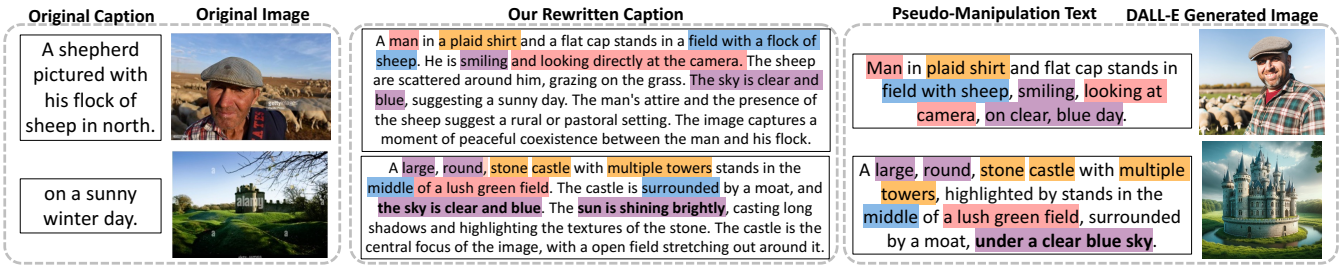

Figure 5: visualization of our intent-CC3M dataset. We leverage DALL-E to generate images for pseudo text. We highlight the intentions embedded in captions. Our pseudo-manipulation text filters the intention-irrelevant details.

	Methods	CIRR R1	CIRR R5	CIRR R10	Fashion-IQ R10	Fashion-IQ R50
1.	full model	30.0	59.7	74.5	32.2	54.5
Significance of intent-CC3M						
2.	w/o intent-CC3M	26.5	55.6	69.4	27.8	48.6
3.	w/o CoT	26.7	56.1	70.0	28.3	49.1
4.	w/o filtering	27.3	56.5	70.5	28.8	49.6
Key modules of De-MINDS process						
5.	w/o De-MINDS	25.9	55.4	69.5	27.6	48.5
6.	w/o global feature	27.7	57.5	70.7	29.5	51.9
7.	w/o gate	28.2	57.7	72.3	30.2	52.8
Training strategies						
8.	w/o construct T	28.4	57.9	72.0	30.0	52.5
9.	w/o distil	26.7	55.9	69.6	28.2	48.9
10.	cos distil	28.4	57.8	72.4	30.1	52.5
Alternative solutions						
11.	Long-CLIP	26.3	54.8	69.1	28.2	49.0
12.	`a photo of` S_*	27.6	57.4	70.6	29.6	51.8
13.	Qwen-VL caption	28.8	58.1	73.2	30.5	53.3
14.	LLM caption	27.0	55.7	69.9	28.5	49.2

Table 5: Ablation study on CIRR and Fashion-IQ.

(model '11') decreases performance by 4.70%, validating our approach's necessity. Missing image-to-text mapping (model '12') results in a 2.78% decline, supporting the value of pseudo-manipulation text. Using Qwen-VL (Wang et al. 2024) generated texts (model '13') decreases performance by 1.40%, indicating superior MLLM models improve quality. Utilizing LLaMA-rewritten captions (Fan et al. 2024) (model '14') causes a 4.12% drop, highlighting our detailed pseudo-manipulation text generation.

Analysis

In this subsection, we provide detailed analyses of De-MINDS's interpretability and efficiency.

Visualization of the intent-CC3M. We analyze visualizations of intentions in the intent-CC3M dataset in Figure 5. Redundant text details can obscure manipulation intentions, *e.g.,* "coexistence between the man and his flock" (line 1), making it harder for models to discern intent. We use MLLMs to filter non-essential information, generating refined descriptions that help models focus on core intent. Abstract captions also contain implicit intentions, *e.g.,* "a sunny winter day" (line 2) implies "a bright sun" and "a clear sky." We use MLLMs to rewrite captions, making these intentions explicit, improving model understanding.

Interpretability of Learnable Query. In Figure 6, we visualize the top two attention words of each learnable query from the last block, demonstrating the distinct focus of the four queries. Specifically, the first two queries mainly focus on object and attribute information, while the last two mostly consider scene and relation information.

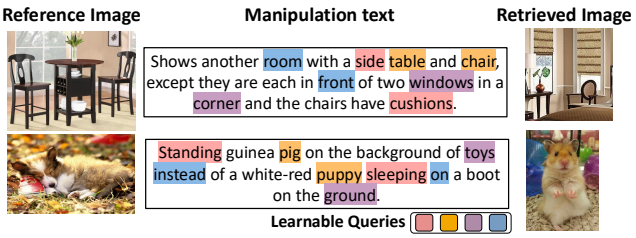

Figure 6: Visualization of the top two attention words for each query, different colors denoting each query result.

Effectiveness and Efficiency Analysis. Our approach achieves significant improvements on four widely compared ZR-CIR tasks from 2.15% to 4.05% over the SoTA models. Designed to understand manipulation intention, the model size of De-MINDS(58.5M) is larger than the simple 3-layer MLP mapping (0.9M) of Pic2Word. Consequently, our training time (20 hours) is 6 hours longer than Pic2Word under the same settings. Notably, our inference time (0.017s) is $\times 58$ faster than CIReVL ($\sim$ 1s), which uses LLM for inference, and only 0.005s slower than Pic2Word. It's worth noting that our model using just 50% of the pre-training data achieves comparable performance to SoTA models.

Conclusion

In this paper, we introduce intent-CC3M, a dataset containing pseudo-manipulation texts generated via MLLM-based chain-of-thought prompting, specifically designed for intention-relevant visual alignment. Leveraging intent-CC3M, we propose De-MINDS, a manipulation intention understanding framework that distills MLLM reasoning to enhance CLIP's interpretation of user intentions in CIR tasks. De-MINDS exhibits strong generalization, substantially outperforming prior methods across four diverse ZS-CIR benchmarks with comparable inference speed. Our approach contributes significantly to intention-aware image retrieval and broader vision-language research.

Acknowledgments

This work was supported by a research grant Project No. E5N00611E5 and Beijing Nova Program under Grant No. 20250484921.

References

Bachlechner, T.; Majumder, B. P.; Mao, H.; Cottrell, G.; and McAuley, J. 2021. Rezero is all you need: Fast convergence at large depth. In *Uncertainty in Artificial Intelligence*, 1352–1361.

Baldrati, A.; Agnolucci, L.; Bertini, M.; and Del Bimbo, A. 2023. Zero-Shot Composed Image Retrieval with Textual Inversion. *arXiv:2303.15247*.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Del Bimbo, A. 2022. Effective Conditioned and Composed Image Retrieval Combining CLIP-Based Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21466–21474.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *Proc. of International Conference on Machine Learning (ICML)*, 1597–1607. PMLR.

Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2): 1–60.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, 248–255.

Fan, L.; Krishnan, D.; Isola, P.; Katabi, D.; and Tian, Y. 2024. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36.

Gadre, S. Y.; Ilharco, G.; Fang, A.; Hayase, J.; Smyrnis, G.; Nguyen, T.; Marten, R.; Wortsman, M.; Ghosh, D.; Zhang, J.; et al. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.

Gu, G.; Chun, S.; Kim, W.; ; Kang, Y.; and Yun, S. 2024. Language-only Efficient Training of Zero-shot Composed Image Retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; Song, D.; Steinhardt, J.; and Gilmer, J. 2021. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8349.

Ilharco, G.; Wortsman, M.; Carlini, N.; Taori, R.; Dave, A.; Shankar, V.; Namkoong, H.; Miller, J.; Hajishirzi, H.; Farhadi, A.; et al. 2021. Openclip. *Zenodo*.

Karthik, S.; Roth, K.; Mancini, M.; and Akata, Z. 2024. Vision-by-Language for Training-Free Compositional Image Retrieval. In *The Twelfth International Conference on Learning Representations*.

Lai, Z.; Zhang, H.; Wu, W.; Bai, H.; Timofeev, A.; Du, X.; Gan, Z.; Shan, J.; Chuah, C.-N.; Yang, Y.; et al. 2023. From scarcity to efficiency: Improving clip training via visual-enriched captions. *arXiv preprint arXiv:2310.07699*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, 12888–12900.

Li, Y.; Ma, F.; and Yang, Y. 2025. Imagine and Seek: Improving Composed Image Retrieval with an Imagined Proxy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3984–3993.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

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.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Nguyen, T.; Gadre, S. Y.; Ilharco, G.; Oh, S.; and Schmidt, L. 2024. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36.

OpenAI. 2022. ChatGPT. GitHub Repository.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.

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.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Annual*

Meeting of the Association for Computational Linguistics, 2556–2565.

Shi, J.; Yin, X.; Chen, Y.; Zhang, Y.; Zhang, Z.; Xie, Y.; and Qu, Y. 2025. Multi-Schema Proximity Network for Composed Image Retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 19999–20008.

Shi, J.; Zhang, Y.; Yin, X.; Xie, Y.; Zhang, Z.; Fan, J.; Shi, Z.; and Qu, Y. 2023. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11218–11228.

Song, H.; Dong, L.; Zhang, W.-N.; Liu, T.; and Wei, F. 2022. CLIP Models are Few-shot Learners: Empirical Studies on VQA and Visual Entailment. arXiv:2203.07190.

Suo, Y.; Ma, F.; Zhu, L.; and Yang, Y. 2024. Knowledge-Enhanced Dual-stream Zero-shot Composed Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26951–26962.

Tang, Y.; Tian, M.; Si, Q.; Mancini, M.; Yu, J.; Gai, K.; Li, L.; Gao, Y.; Song, J.; Zheng, B.; Gou, G.; Xiong, G.; and Wu, Q. 2025a. Boosting Training-Free Composed Image Retrieval with Tools.

Tang, Y.; Yamada, Y.; Zhang, Y.; and Yildirim, I. 2023. When are Lemons Purple? The Concept Association Bias of Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14333–14348.

Tang, Y.; Yu, J.; Gai, K.; Zhuang, J.; Gou, G.; Xiong, G.; and Wu, Q. 2024a. Denoise-i2w: Mapping images to denoising words for accurate zero-shot composed image retrieval. *arXiv preprint arXiv:2410.17393*.

Tang, Y.; Yu, J.; Gai, K.; Zhuang, J.; Xiong, G.; Gou, G.; and Wu, Q. 2025b. Missing Target-Relevant Information Prediction with World Model for Accurate Zero-Shot Composed Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24785–24795.

Tang, Y.; Yu, J.; Gai, K.; Zhuang, J.; Xiong, G.; Hu, Y.; and Wu, Q. 2024b. 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.

Tang, Y.; Zhang, J.; Qin, X.; Yu, J.; Gou, G.; Xiong, G.; Lin, Q.; Rajmohan, S.; Zhang, D.; and Wu, Q. 2025c. Reason-before-Retrieve: One-Stage Reflective Chain-of-Thoughts for Training-Free Zero-Shot Composed Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14400–14410.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6439–6448.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11307–11317.

Yang, Z.; Xue, D.; Qian, S.; Dong, W.; and Xu, C. 2024. LDRE: LLM-based Divergent Reasoning and Ensemble for Zero-Shot Composed Image Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 80–90.

Zhang, K.; Luan, Y.; Hu, H.; Lee, K.; Qiao, S.; Chen, W.; Su, Y.; and Chang, M.-W. 2024. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Conditional Prompt Learning for Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.