Disentangling Semantics via Concept Quantization for Zero-Shot Composed Image Retrieval

Yusong Qiu*[1]
saramanda@sjtu.edu.cn
Yue Xu*[1]
silicxuyue@sjtu.edu.cn
Siqi Liu[1]
magi-yunan@sjtu.edu.cn
Junfu Pu†[2]
jevinpu@tencent.com
Ying Shan[2]
yingsshan@tencent.com
Yong-Lu Li†[1]
yonglu_li@sjtu.edu.cn

[1] Shanghai Jiao Tong University
[2] ARC Lab, Tencent IEG

Abstract

Composed Image Retrieval (CIR) requires models to retrieve images that match both a reference image and a textual instruction, posing significant challenges in compositional reasoning and cross-modal alignment. While recent zero-shot CIR (ZS-CIR) methods leverage text inversion to bridge vision and language, they often encode the reference image into holistic, entangled representations, making it difficult to selectively manipulate static properties versus functional semantics. In this work, we tackle this challenge by an explicit disentanglement of visual representations into their fundamental concepts: attributes and affordances. Our novel framework leverages vector-quantized codebooks to learn a discrete semantic vocabulary for each of these concepts. Then, these concept codes are dynamically modulated by an instruction-aware conditioning module, generating a query embedding precisely aligned with the editing intent. Experiments demonstrate that our method achieves state-of-the-art performance on CIR benchmarks.

1 Introduction

Composed Image Retrieval (CIR) [30, 49] bridges the gap between image retrieval and AI-generated content (AIGC) by enabling instruction-driven image edits. It receives an image and a relative caption to simulate image edits directly in the retrieval space. Further, Zero-Shot CIR (ZS-CIR) [3] elevates this challenge by generalization to unseen editing instruc-

∗ Equal contribution. † Corresponding authors.

© 2026. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

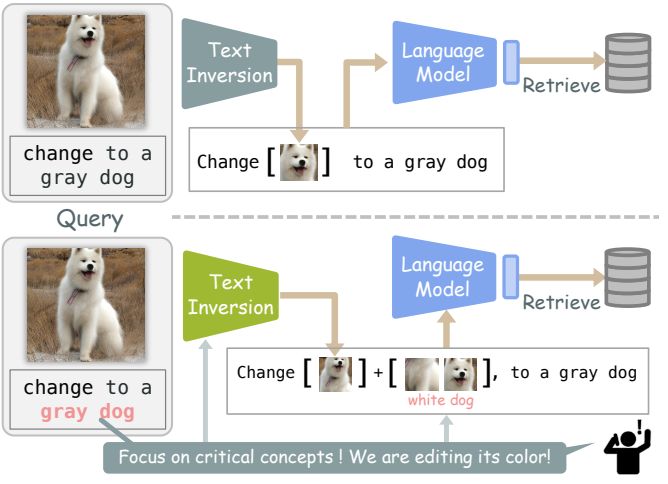

Figure 1: Comparison of vanilla text inversion (top) to our concept-augmented model with relative conditioning (bottom).

tions without supervised triplet data. ZS-CIR advances retrieval systems and the broader goal of building generalizable VLMs.

Among the predominant ZS-CIR paradigms, text inversion [11] is highly competitive, compressing the image into pseudo-text tokens compatible with language encoders. Though intuitive, this static, instruction-agnostic inversion process suffers from representation collapse, where fine-grained visual details are sacrificed to fit into a holistic, entangled embedding. This static, instruction-agnostic process inevitably discards critical information and conflates distinct concepts—objects, properties, and functions, preventing the model from identifying which specific visual components the instruction should modify.

This bottleneck motivates an investigation of learning finer-grained concepts. For example, FTI4CIR [27] takes a first step by exploiting the attribute concept with BLIP [23] model. We argue, however, that a more complete conceptual framework is crucial. This raises three critical questions for a fine-grained ZS-CIR paradigm: (1) What concepts should be considered for CIR? (2) How can we learn robust concept tokenizers in a zero-shot setting? (3) How do we best leverage these concepts for the retrieval task?

To bridge this gap, we ground our approach in two fundamental concepts of object and scene: attributes (properties like color and material, *e.g.* "blue") and affordances (functional or contextual roles, *e.g.* "seating"). These concepts align with how humans describe objects and scenes: attributes emerge during static human perception, and affordances describe the potential interaction between the human and scene. By modeling both, we create a structured tokenization that enriches semantic expressiveness while maintaining VLM-compatibility.

Then, to learn robust concept tokenizers in a zero-shot setting, we build upon a language-only paradigm [13] to exploit the massive linguistic prior in LLMs. We model the attribute and affordance concepts with two discrete semantic codebooks and learn them via vector quantization. This process acts as a powerful regularizer, forcing the model to map diverse inputs onto a finite set of prototypes. This mechanism inherently denoises the concept extraction and creates a robust and disentangled latent space for true compositional reasoning.

With the concept modules, we could guide the text inversion model to enhance its representation quality. Given the limited model capacity, the task-agnostic text inversion could waste its capacity on irrelevant details. *E.g.*, an instruction "make the sofa blue" requires focusing on the sofa's color attribute, not the background. To prioritize task-critical information, we identify key concepts in the *editing instruction* itself and use them to condition the text inversion model. This conditioning dynamically amplifies pertinent semantics, ensuring the final embeddings are compact and precisely aligned with the editing intent.

We translate these insights to introduce **ConCIR** (**Con**cept **C**omposed **I**mage **R**etrieval), integrating: (1) **Concept Augmented Text Inversion** (**CATI**): learns two discrete semantic codebooks (attributes and affordances) via quantization, with a continuous token for global context. (2) **Relative Concept Conditioning** (**RCC**): modulates the model on the concepts in the relative caption, ensuring the inverted embedding retains the most critical details.

ConCIR generates a rich, disentangled semantic representation that adapts to specific edit instructions. As a modular method, it can be adapted to various text-inversion approaches. Experiments show that ConCIR significantly outperforms state-of-the-art methods, achieving gains of over +3% on popular CIR benchmarks like CIRCO [3] and GeneCIS [48]. Our contributions are threefold: (1) We identify a critical gap in ZS-CIR: the need to model fine-grained concepts for complete compositional reasoning. (2) We propose a novel quantization-based tokenizer that learns robust, discrete, and denoised concept codebooks. (3) We introduce a dynamic conditioning mechanism that leverages these concepts for task-specific, instruction-aware retrieval, leading to SOTA performance.

2 Related Work

2.1 Composed Image Retrieval

Composed Image Retrieval (CIR) [49] aims to retrieve target images by modifying a reference image according to language instructions. Unlike conventional image retrieval, which relies on direct similarity matching, CIR requires understanding compositional reasoning over visual and textual modalities to simulate human-like editing. Thus, annotated triplets are needed to train CIR models. ComposedAE [1], an autoencoder-based method, is proposed to learn the compositional features of images and text. CLIP4Cir [2] shows a CLIP [37]-based alignment framework by fine-tuning the pre-trained text encoders.

While recent vision-language models have advanced CIR, Zero-Shot Composed Image Retrieval (ZS-CIR) presents a much harder challenge: models must generalize to CIR tasks without access to supervised triplet data. Representative methods for ZS-CIR based on textual inversion [11] combine image and text features into a single text query for retrieval. Pic2Word [39] and SEARLE [3] use image-based training, while LinCIR [15] first proposes text-based training and alleviates the modal gap. ContextI2W [43] learns multiple queries and uses both image and caption to enrich the information in textual inversion tokens. Besides, some attempts like RTD [4] partially finetune CLIP to tackle the task discrepancy. De-MINDS [42] builds a linguistic dataset to guide the text encoder for the editing intention. MoTaDual [22] leverages learnable tokens as inputs of both the text and image encoders. These tokens are trained using LLM-generated text triplets. InstructCIR [8] uses LLaVA [29] and an LLM to generate a large dataset from existing image-text pairs for model training.

Beyond textual inversion, other approaches have also been explored in different directions. SCOT [17] and another recent study [46] similarly leverage text triplets generated

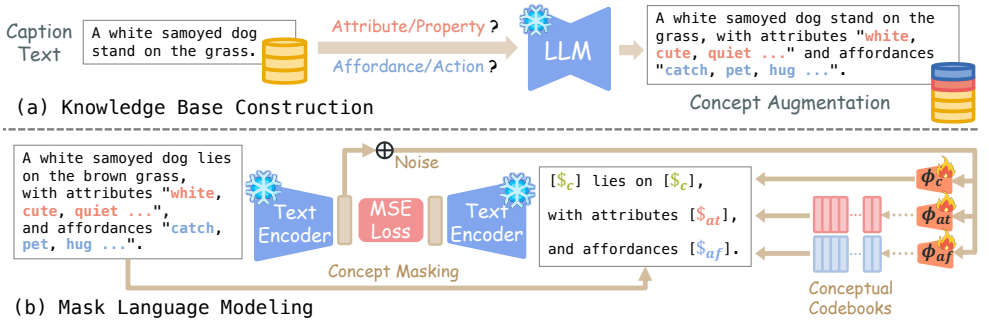

Figure 2: Overview of concept-augmented pretraining. (a) We use LLMs to augment the text caption from the existing corpus, especially focusing on the object concepts, including attributes and affordances. (b) We use a language-only pretraining paradigm following [15]. The key concepts are masked from the captions and recovered by three text inversion models $\phi_c, \phi_{at}, \phi_{af}$, representing context, attribute, and affordance.

by large language models (LLMs) and fine-tune the Q-Former within the pre-trained BLIP-2 [24] framework, exploring the capabilities of multimodal models in the zero-shot CIR domain, although both training and inference are relatively computationally intensive. Recently, larger LLMs and MLLMs are adopted to solve ZS-CIR tasks during inference time, and achieve good performance due to their large capacity and zero-shot reasoning ability [16, 25, 44]. However, these methods violate the basic protocol of ZS-CIR that only a frozen CLIP model can be used during inference, thus they are beyond our scope.

2.2 Object and Scene Concepts

Object-Centric (OC) learning aims to learn visual representations by viewing images as combinations of different objects. Many methods focus on implicit object-centric representation, especially in a self-supervised paradigm. Slot Attention [31] demonstrates remarkable success in OC learning. DINO [32] employs the ViT structure [7] and attention mechanism to extract the multi-level information of each object in the image through self-supervised learning. In contrast, object and scene concepts could also be learned explicitly. Object concept learning (OCL) [26] divides vision's understanding of objects and scenes into hierarchies and proposes fine-grained object **attributes** and **affordances** recognition. *Attributes* describe the objective physical properties of objects, including, *e.g.*, shape, color and texture. Attributes have long served as fine-grained discriminative features for object categorization [9] and as intermediate representations for downstream perceptual tasks [20, 33, 34]. This motivates attribute large-scale dataset construction [19, 28]. *Affordances*, originating from ecological psychology and cognitive science [12], refer to the perceived functions of objects, which emerged as a critical research focus in scene understanding [18, 45, 54] and robotic manipulation [35, 36]. CIR tasks require precise scene-object editing, so it is critical to understand both object attributes and affordances comprehensively.

3 Method

3.1 Preliminaries

Composed Image Retrieval (CIR) aims to retrieve an image $\mathbf{I}_t$ from a database using a multimodal query: a reference image $\mathbf{I}_r$ and a relative caption $\mathbf{T_r}$. The task requires finding the most relevant image to I_r after the modifications of T_r. Further, Zero-Shot CIR addresses scenarios where no supervised triplets $\{\mathbf{I}_r, \mathbf{T_r}, \mathbf{I}_t\}$ are available. Models should leverage pretrained vision-language models (*e.g.* CLIP [37]) to generalize to unseen compositions.

Recently, **Text Inversion** advances ZS-CIR. It bridges visual and textual modalities by mapping reference image features $\mathcal{F}_v(\mathbf{I}_r)$ into the token embedding space of a pretrained text encoder. A lightweight projection network ϕ generates a pseudo-word token $\$ = \phi(\mathcal{F}_v(\mathbf{I}_r))$, treated as a representation of the whole image in the text template, *e.g.*,

$$\mathbf{q} = \mathcal{F}_t(\texttt{"a photo of [\$] that [}\mathbf{T}_r\texttt{]"}), \tag{1}$$

where text encoder $\mathcal{F}_t$ remains frozen, while the projection network ϕ is trained via contrastive learning on image-text pairs. The prompt $\mathbf{q}$ is encoded as $\mathbf{q}$ to retrieve $\mathbf{I}_t$ in the database via feature similarity.

The fundamental limitations of text inversion methods stem from the constrained capacity of the inverted token, with two critical challenges: (1) how to maximize the semantic richness of image representations, and (2) how to prioritize instruction-relevant information while discarding less critical details. Hence, we propose a concept-centric approach that decomposes visual concepts into hierarchical levels of abstraction, encoding finer-grained concepts that correlate with editing intents. This introduces the following key components.

3.2 Concept Augmented Text Inversion

We first propose **C**oncept **A**ugmented **T**ext **I**nversion (**CATI**) to address the semantic incompleteness of conventional text inversion through a hierarchical concepts decomposing.

Fine-grained Concepts. To enrich the semantic knowledge in the image representation, we leverage the fine-grained concepts. In cognitive science, there are two primary concepts pertaining to a visual entity: (1) **attributes** are the static visual properties of an entity (*e.g.* color, texture, material). These attributes constitute an **object-centric** and an objective description of objects and scenes, and constitute the majority of visual concepts. (2) **affordances** are the dynamic properties of an entity, capturing functional or contextual relations (*e.g.*, a chair "can be sat on"). These are **human-centric** and subjective concepts, which complement attributes and are critical for the human-computer interaction (HCI), including the retrieval system. Therefore, we incorporate both the attributes and affordances as fine-grained concepts for our model. These concepts are rooted in their role as foundational units of object and scene semantics [26], and are also well-studied in cognition [12, 35, 36, 45] and vision research [10, 20, 34, 38, 50, 55, 56], offering a structured vocabulary for multimodal reasoning and aligning with human priors and cognition. The ablation experiments in the next sections also show the necessity of both types of concepts.

Knowledge Base Construction. As shown in Fig. 2 (a), we first build **text-only** concept learning corpus based on regular ZS-CIR pretraining data (*e.g.* CC3M [40]). We use LLMs to augment the regular image description text with fine-grained concepts [1]. For each caption,

[1] The instruction for concept description generation is given in the supplementary materials.

we extract the top 5 attributes and affordances with Qwen 3 [53]. By leveraging LLMs, we extract these concepts without costly supervision, ensuring the scalability of our method.

Here the adoption of an LLM-augmented, text-only training corpus is a choice grounded in the constraints and common practices of ZS-CIR. While some prior works leverage multimodal models like BLIP [23] to extract continuous image attributes, using **text-only LLMs** for generating linguistic augmentation remains a widely recognized and equitable strategy within the pure language-only ZS-CIR paradigm [15, 22]. Notably, many recent advances exploit heavy large multimodal models even in inference time, which is far beyond the scope of ZS-CIR task. So we are not adopting these techniques, and we do not compare against these methods. This adherence to the language-only domain ensures a fair comparison against leading ZS-CIR methods and maintains the zero-shot nature of the task.

Semantic Concept Quantization. As shown in Fig. 2 (b), the concept-augmented caption $\mathbf{T}$ is first encoded into a global text feature vector $\mathbf{f}_t = \mathcal{F}_t(\mathbf{T})$ by a frozen CLIP text encoder $\mathcal{F}_t$. Our goal is to decompose this holistic feature $\mathbf{f}_t$ into a set of structured, disentangled, and robust concept tokens. This process faces two challenges: the noise inherent in the LLM-generated corpus and the discrete nature of visual concepts. To address this, we introduce a **Semantic Quantization** framework. Instead of mapping $\mathbf{f}_t$ directly to continuous embeddings, we convert it to three tokens: a continuous context token $\$_c$, and two *discrete* tokens $\$_{at}$ and $\$_{af}$ derived from vector quantization (VQ). Visual concepts like color or functional roles are inherently categorical in human language. By discretizing the latent space into a finite vocabulary of semantic prototypes, we force the model to discard continuous visual noise (e.g., lighting variations) and focus on high-level symbolic meanings.

First, we use three encoder modules (implemented as MLPs) $\phi_c, \phi_{at}, \phi_{af}$ to project the global feature $\mathbf{f}_t$ into concept-specific latent spaces:

$$\mathbf{z}_c = \phi_c(\mathbf{f}_t), \quad \mathbf{z}_{at} = \phi_{at}(\mathbf{f}_t), \quad \mathbf{z}_{af} = \phi_{af}(\mathbf{f}_t). \tag{2}$$

Then, the context token $\$_c$ is simply the continuous output $\mathbf{z}_c$. But for attributes and affordances, we introduce two independent, learnable, discrete codebooks: $\mathcal{C}_{at} = \{\mathbf{e}_{at}^{(k)}\}_{k=1}^{K}$ and $\mathcal{C}_{af} = \{\mathbf{e}_{af}^{(k)}\}_{k=1}^{K}$, where K is the codebook size (a hyperparameter) and $\mathbf{e} \in \mathbb{R}^D$ are the "semantic prototypes". The continuous latents $\mathbf{z}_{at}$ and $\mathbf{z}_{af}$ are then quantized by finding the nearest codebook vector:

$$\begin{aligned} \$_{at} &= \mathbf{e}_{at}^{(k)}, \quad \text{where } k = \arg\min_j \|\mathbf{z}_{at} - \mathbf{e}_{at}^{(j)}\|_2^2; \\ \$_{af} &= \mathbf{e}_{af}^{(k)}, \quad \text{where } k = \arg\min_j \|\mathbf{z}_{af} - \mathbf{e}_{af}^{(j)}\|_2^2. \end{aligned} \tag{3}$$

This quantization forces the model to map noisy, diverse inputs ($\mathbf{z}$) onto a finite set of clean, learned prototypes ($\mathbf{e}$), acting as a powerful regularizer and denoiser.

Learning Objective. Our framework is trained end-to-end using a language-only mask modeling objective [15], combined with the VQ-loss [47]. **(1) Mask Language Modeling Loss ($\mathcal{L}_{MLM}$):** We borrow the name from language models. Here the three concept adaptors ϕ_c, ϕ_{at}, ϕ_{af} are learned by mask modeling following [15] in a language-only manner. For each concept-rich caption, the core concepts, including context description, attributes, and affordance, are masked and replaced by the textual concept embeddings $\$_c, \$_{at}, \$_{af}$. Both the original and the masked captions are fed into the same frozen CLIP text encoder and constrained by the MSE loss between features. During training, noise from $\mathcal{N}(0,1) \times \mathcal{U}(0,1)$ is added to $\mathbf{f}_t$ to alleviate the modality gap [15], allowing the encoders ϕ to also consume

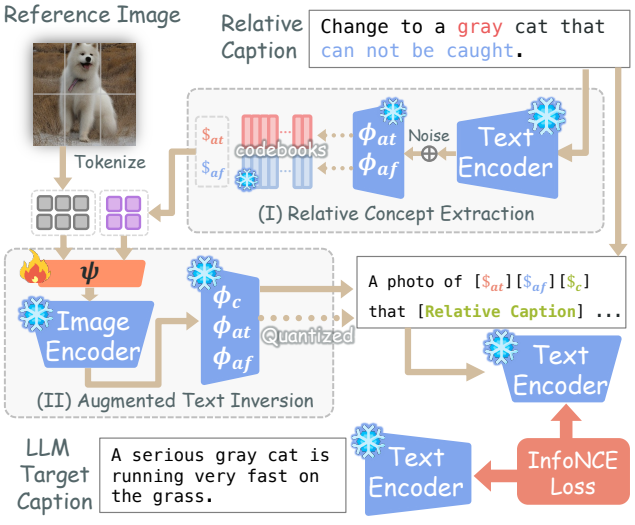

Figure 3: Overview of relative concept conditioning. We use the text inversion models $\phi_c, \phi_{at}, \phi_{af}$ to extract the concept tokens as the condition of the image encoder. This injects the preference for image editing into the representation model.

image features during text inversion. **(2) VQ-Loss ($\mathcal{L}_{VQ}$):** This loss trains the codebooks and encoders. It consists of a *codebook loss* (pulling the chosen prototype $\mathbf{e}$ towards the encoder output $\mathbf{z}$) and a *commitment loss* (pulling the encoder output $\mathbf{z}$ towards the prototype $\mathbf{e}$). For each concept (*e.g.*, attribute):

$$\mathcal{L}_{VQ-at} = \underbrace{\|\text{sg}(\mathbf{z}_{at}) - \mathbf{e}_{at}^{(k)}\|_2^2}_{\text{Codebook Loss}} + \beta \underbrace{\|\mathbf{z}_{at} - \text{sg}(\mathbf{e}_{at}^{(k)})\|_2^2}_{\text{Commitment Loss}}, \tag{4}$$

where "sg" is stop-gradient operation and β is a weighting parameter.

The final loss for CATI is: $\mathcal{L}_{CATI} = \mathcal{L}_{MLM} + \lambda_1 \mathcal{L}_{VQ-at} + \lambda_2 \mathcal{L}_{VQ-af}$, where λ_1, λ_2 are loss weights. We train on the CC3M-595k corpus [29, 40] following [22], **strictly preventing data contamination**.

The final resultant tokens $\$_c, \$_{at}, \$_{af}$ form a structured representation of the image, which are now **discrete semantic prototypes** selected from our learned codebooks. This is a crucial distinction from prior work, providing a robust, denoised, and structured representation that aligns with human perception and is perfectly suited for compositional edits.

3.3 Relative Caption Conditioning

While CATI provides robust, discrete concept tokens, naively inverting all of them results in semantic redundancy, where the model's capacity is diluted by visual information irrelevant to the current editing intent. To prioritize instruction-critical semantics, we introduce a **Relative Concept Conditioning (RCC)** module. As depicted in Fig. 3, this module dynamically modulates the visual features based on the concepts extracted from the relative caption $\mathbf{T_r}$.

Conditioning. Our conditioning process (Fig. 3 (I)) re-leverages our CATI framework in a novel, zero-shot manner. First, the relative caption $\mathbf{T_r}$ is encoded by the frozen text

encoder $\mathcal{F}_t$ to produce a feature $\mathbf{f}_r$. To bridge the modality gap between these text features for future fusion with image features, we apply noise injection [15] to the latent feature. Then the feature is ready for image models. It is passed through our **frozen** CATI encoders ϕ_{at} and ϕ_{af}, and quantized by the frozen concept codebooks to extract attribute and affordance concept tokens $\$_{at}, \$_{af}$. These adaptors are reused from the CATI stage and remain frozen, as they have already acquired robust semantic extraction capabilities. To use the concept tokens as image encoder conditions, we use a learnable module ψ to fuse the tokens to the image patch tokens $\mathbf{P} \in \mathbb{R}^{N \times d}$ (N is the number of patches and d is the dimension).

The conditioning is performed by a learnable module ψ (Fig. 3 (II)), implemented as a compact cross-attention mechanism. We treat the image patch tokens $\mathbf{P} \in \mathbb{R}^{N \times d}$ as *queries* and the concatenated, instruction-derived concept tokens as *keys* and *values*:

$$\mathbf{Q} = \mathbf{P}\mathbf{W}_Q, \mathbf{K} = [\$_{at}; \$_{af}]\mathbf{W}_K, \mathbf{V} = [\$_{at}; \$_{af}]\mathbf{W}_V, \tag{5}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable projection matrices and $[\cdot;\cdot]$ denotes concatenation. The cross-attention output $\mathbf{P}'$ is computed with a residual connection:

$$\mathbf{P}' = \sigma\left(\mathbf{Q}\mathbf{K}^\top/\sqrt{d}\right)\mathbf{V} + \mathbf{P}, \tag{6}$$

where σ is the Softmax function. This allows each image patch to dynamically attend to the instruction's concepts. By learning from the curated dataset, the cross-attention mechanism accounts for both positive and negative correlations of the instruction to the query image. Finally, the updated tokens $\mathbf{P}'$ are then passed to the image encoder, yielding the conditioned visual feature $\mathcal{F}_v(\mathbf{P}')$.

Domain Alignment with Contrastive Learning. The conditioned image feature $\mathcal{F}_v(\mathbf{P}')$ is now fed into our full CATI pipeline (Sec. 3.1) to generate the final, editing-aware prompt. Specifically, $\mathcal{F}_v(\mathbf{P}')$ is processed by all three concept adaptors (ϕ_c, ϕ_{at}, ϕ_{af}) and the respective codebooks for the concept tokens $\$_c, \$_{at}, \$_{af}$. These are embedded into the prompt:

$$\mathbf{T}_{\text{edit}} = \texttt{"a photo of [\$}_c\texttt{][\$}_{at}\texttt{][\$}_{af}\texttt{] that [}\mathbf{T}_r\texttt{]"}. \tag{7}$$

The text encoder $\mathcal{F}_t$ processes $\mathbf{T}_{\text{edit}}$ to produce the final query embedding $\mathbf{q}$.

To optimize the only learnable module in this stage (the cross-attention ψ), we employ contrastive learning. We build the pretrain database following [22]. For each caption, we use LLM to generate a relative caption and also imagine a target caption after editing. The query $\mathbf{q}$ generated by ConCIR is aligned with the embedding of the target caption $\mathbf{k} = \mathcal{F}_t(T_{target})$ using the InfoNCE loss:

$$\mathcal{L}_{RCC}(\mathcal{B}) = -\sum_i \left[\log \sigma_i\left(\frac{\mathbf{k}_j^\top \mathbf{q}_j}{\tau}\right) + \log \sigma_i\left(\frac{\mathbf{q}_j^\top \mathbf{k}_j}{\tau}\right)\right], \tag{8}$$

where σ_i is the i-th value of Softmax and $\mathcal{B}$ is a minibatch. This loss aligns the query $\mathbf{q}_i$ with its target $\mathbf{k}_i$ against all other negatives $\mathbf{k}_j$ ($\forall j \neq i$) in the batch, teaching ψ to fuse the instruction's concepts into the visual features effectively.

3.4 Inference

Putting the two stages together, we propose Concept Composed Image Retrieval (**ConCIR**). ConCIR yields concept tokens and uses conditioning to generate a retrieval-ready query

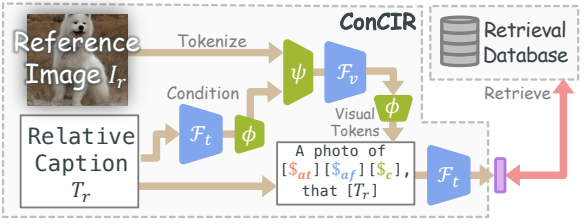

Figure 4: Inference process of ConCIR.

Method	Latency (ms)		Parameters (M)	
	Feature	Total	Trainable	Total
Pic2Word	19.8	297.4	1.0	408.8
LinCIR	20.0	297.6	13.5	421.3
MoTaDual	20.5	298.1	14.3	422.1
ConCIR	31.3	308.9	22.5	430.3

Table 1: Inference efficiency.

embedding $\mathbf{q}$. As in Fig. 4, during inference, the relative caption $\mathbf{T}_r$ is fed into the text encoder and concept adaptors for $\$_c, \$_{at}, \$_{af}$ as conditions. The reference image $\mathbf{I}_r$ is then tokenized and fused with the conditions with cross-attention and converted to tokens for the query template. With the relative caption, we extract the query embedding of the prompt for retrieval. Please refer to the supplementary for the pseudocode of the method pipeline.

Inference Efficiency. While we introduce additional computation, the overhead remains modest, as shown in Tab. 1. Our additional trainable components consist of the three concept encoders ($\phi_c, \phi_{at}, \phi_{af}$), the two discrete codebooks ($\mathcal{C}_{at}, \mathcal{C}_{af}$), and the lightweight cross-attention module (ψ).

We compare the efficiency of SOTA methods in Tab. 1 with the CLIP-L backbone. For each query, ConCIR adds around 10 ms of latency and 8 M parameters over the LinCIR baseline. This overhead is negligible for practical deployment (e.g., a 3.8% relative latency increase on COCO) but is critical for achieving significant accuracy gains.

This small overhead is worthwhile, as it enables fine-grained concept reasoning capability, **which is not achievable by merely increasing the model size**. As we demonstrate in our ablation study (Sec. 4.3), ConCIR outperforms a baseline LinCIR model that is given a comparable parameter budget, justifying our structured approach.

4 Experiments

4.1 Implementation Details

We experiment on mainstream ZS-CIR benchmarks including CIRCO [3], GeneCIS [43], CIRR [30] and FashionIQ [51], and we adhere to a strict **zero-shot protocol**: models can not train on labeled CIR triplets in the dataset and rely solely on pretrained VLMs.

For model training, we leverage pretrained VLMs to ensure zero-shot: (1) CLIP large (CLIP-L), with weights from [37], (2) CLIP giga (CLIP-G), with weights from [6]. The

Method (w/ CLIP-L)	CIRCO				GeneCIS		
	mAP@5	mAP@10	mAP@25	mAP@50	R@1	R@2	R@3
Pic2Word [39]	8.72	9.51	10.64	11.29	11.16	21.47	30.38
SEARLE [3]	11.68	12.73	14.33	15.12	12.26	22.11	31.30
MTCIR [5]	10.36	11.63	12.95	13.67	-	-	-
CompoDiff [14]	12.55	13.36	15.83	16.43	14.88	25.26	37.24
FTI4CIR[27]	15.05	16.32	18.06	19.05	-	-	-
LinCIR [15]	12.59	13.58	15.00	15.85	12.19	22.76	32.38
+ CATI (**Ours**)	14.64	15.77	17.62	18.54	14.91	26.53	36.57
+ CATI + RCC (**Ours**)	**16.38**	**17.39**	**19.08**	**20.07**	**15.79**	**27.20**	**37.48**
MoTaDual [22]	20.42	21.62	23.99	25.00	16.20	27.70	37.58
+ CATI (**Ours**)	21.00	22.36	24.64	25.70	16.36	**27.91**	38.11
+ CATI + RCC (**Ours**)	**21.14**	**22.31**	**24.76**	**25.75**	**16.64**	27.88	**38.34**

Table 2: Performance with CLIP-L backbone on CIRCO [3] and GeneCIS [28] benchmarks. The best results are bolded, and the second best are underlined.

Method (w/ CLIP-G)	CIRCO			
	mAP@5	mAP@10	mAP@25	mAP@50
Pic2Word† [39]	5.54	5.59	6.68	7.12
SEARLE† [3]	13.20	13.85	15.32	16.04
CompoDiff [14]	15.33	17.71	19.45	21.01
LinCIR [15]	19.71	21.01	23.13	24.18
+ ConCIR (**Ours**)	**20.95**	**22.31**	**24.52**	**25.59**
MoTaDual [22]	28.36	29.60	32.19	33.27
+ ConCIR (**Ours**)	**29.02**	**30.73**	**33.47**	**34.65**

Table 3: Performance with CLIP-G backbone on CIRCO [3]. (†: reproduced by [15]).

concept adaptors ϕ_c, ϕ_{at}, and ϕ_{af} are three-layered MLPs, and the first two linear layers are shared. The condition adaptor ψ is a cross-attention. We train on a subset of the image captioning dataset CC3M with 595 K samples [29, 40], avoiding any exposure to CIR triplets. Please refer to the supplementary for more details.

4.2 Results

We evaluate ConCIR against leading zero-shot CIR methods on CIRCO and GeneCIS. As a plug-and-play, we embed our system into state-of-the-art text inversion methods LinCIR [15] and MoTaDual [22]. We show the results with the CLIP-L backbone in Tab. 2. With the LinCIR baseline, ConCIR achieves mAP@5 of 16.38% and recall 15.79% on the CIRCO and GeneCIS dataset [2] respectively, significantly outperforming the baselines by **3%**. ConCIR also enhances the state-of-the-art method MoTaDual by nearly 1%. We also show the results with the CLIP-G backbone in Tab. 3, where ConCIR shows consistent improvement on the CIRCO dataset. This gain stems from its hierarchical tokenization, which preserves fine-grained concepts. For example, in queries like "change the sofa to leather", ConCIR's

[2]Please refer to the Supplementary for the full GeneCIS results.

Method (w/ CLIP-L)	CIRR					
	R@1	R@5	R@10	R_s@1	R_s@2	R_s@3
Pic2Word [39]	23.90	51.70	65.30	53.76	74.46	87.08
SEARLE [3]	24.24	52.48	66.29	53.76	75.01	88.19
ContextI2W [43]	25.60	55.10	68.50	-	-	-
KEDs [41]	26.40	54.80	67.20	-	-	-
MTCIR [5]	25.52	54.58	67.59	55.64	77.54	89.47
CompoDiff [14]	18.24	53.14	**70.82**	57.42	77.10	87.90
MoTaDual [22]	**27.28**	56.39	-	58.12	78.05	89.83
+ ConCIR (**Ours**)	**27.28**	**58.02**	70.70	**61.59**	**81.01**	**91.13**

Table 4: Performance with CLIP-L backbone on CIRR [30]. $\mathbf{R_s}$@k (k=1, 2, 3) are the recalls on the subset task.

Method (w/ CLIP-L)	shirt		dress		toptee		average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Pic2Word [39]	26.20	43.60	20.00	40.20	27.90	47.40	24.70	43.70
SEARLE [3]	26.89	45.58	20.48	43.13	29.32	49.97	25.56	46.23
LinCIR [15]	29.10	46.81	20.92	42.44	**28.81**	50.18	26.28	46.49
+ ConCIR (**Ours**)	**29.69**	**47.99**	**21.72**	**43.48**	28.45	**50.48**	**26.62**	**47.32**

Table 5: Performance with CLIP-L backbone on FashionIQ [51] benchmark.

attribute tokens explicitly encode material semantics, whereas baselines conflate texture with unrelated properties.

On the CIRR dataset (Tab. 4). Our method brings 1.63% R@5 improvement to the MoTaDual baseline. However, the performance gain is not as consistent as on other datasets. This is due to the potential bias of the CIRR dataset, which has been emphasized in previous work [3, 15, 21]. We provide a deeper analysis of the modality redundancy in CIR benchmarks in Sec. 4.5. On FashionIQ (Tab. 5), ConCIR consistently improves upon the LinCIR, demonstrating its effectiveness on domain-specific datasets.

4.3 Ablation Study

We validate the components of ConCIR. While the ablation of each module has been shown in Tab. 2, more experiments are listed in Tab. 6 on CIRCO.

Model Components. Tab. 2 shows the ablation of our model components (CATI and RCC). CATI and RCC both improve the baseline, while combining them (full ConCIR) achieves the most significant enhancement. And we also find that in some settings, RCC achieves comparable performance to full ConCIR, indicating that during image tokenization, it is critical to focus on the key concepts for CIR.

Concept Tokens (A). During RCC, we use the concept tokens as the condition of the image encoder. We compare different choices of concept tokens in Tab. 6 (A1-A5). Both the combination of $\phi_{at} + \phi_{af}$ (A4) and $\phi_c + \phi_{at} + \phi_{af}$ (A5) achieve leading performance compared to a single concept token, since the finer-grained knowledge of concepts is important for

Category	Variant	mAP@5	mAP@10
Default	All components	**16.38**	**17.39**
Concept Tokens	A1: Only $\$_c$	15.68	16.45
	A2: Only $\$_{at}$	16.07	16.93
	A3: Only $\$_{af}$	14.98	16.18
	A4: $\$_{at}$+$\$_{af}$	**16.38**	**17.39**
	A5: $\$_c$+$\$_{at}$+$\$_{af}$	16.15	16.97
LinCIR + Extra ϕ	B1: 1 ϕ	12.59	13.58
	B2: 3 ϕ's	12.39	13.58
	B3: 4 ϕ's	11.58	12.74
	B4: CATI (Ours)	**14.64**	**15.77**
ϕ Module in CATI	C1: Linear	10.60	11.40
	C2: MLP	13.23	14.39
	C3: Shared MLP	**14.64**	**15.77**
Prompts	D1: A photo of [$\$_c$] that $\mathbf{T}_r$	13.83	14.68
	D2: A photo of [$\$_c$] have attributes [$\$_{at}$] and affordances [$\$_{af}$], that $\mathbf{T}_r$	14.50	15.23
	D3: [$\$_{at}$] [$\$_{af}$] [$\$_c$] that $\mathbf{T}_r$	16.12	17.06
	D4: A photo of [$\$_{at}$] [$\$_{af}$] [$\$_c$] that $\mathbf{T}_r$	**16.38**	**17.39**
Codebook in CATI	E1: w/o codebooks	13.40	14.33
	E2: codebook size 8	13.53	14.60
	E3: codebook size 64	**14.64**	**15.77**
	E4: codebook size 2048	14.42	15.52
LLMs	F1: Qwen3:8b [53]	**16.38**	**17.39**
	F2: Qwen2.5:7b [52]	15.88	16.73
	F3: Llama3.1:8b [13]	16.14	17.09

Table 6: Ablation study on CIRCO with LinCIR. Gray and bold indicate the best.

image editing. Interestingly, the A4 settings yield better performance than A5, suggesting that selective concept focus is more critical for precise editing than holistic retention.

LinCIR with Extra ϕ (B). For a fair comparison with similar parameter size overhead, we extend LinCIR [15] with multiple text inversion adaptors ϕ. We show the comparison with CATI in Tab. 6 B1-B4. With comparable or more trainable parameters, ablation experiments B2 and B3 exhibit performance degradation, while CATI improves the LinCIR by over 1.5% mAP. This justifies the incorporation of fine-grained concept knowledge in ConCIR, which is more computationally efficient than the brute-force stack model parameters.

Architecture of concept adaptor ϕ (C). As shown in Tab. 6 (C1-C3), a non-linear MLP model is necessary for concept extraction. In practice, we share the parameters of the first layer in ϕ's for efficiency (C3), which could also enhance the overall performance.

Prompt Engineering (D). We explore the impact of different retrieval prompts (D1-D4) and list some of them in Tab. 6(D1-D4). Prompt D1 shows that removing pseudo-tokens for attributes and affordances degrades performance (-2.41% mAP@10). Explicitly including tokens for attributes and affordances in the prompt D2 also degrades performance (-1.24% mAP@10), as this format is inconsistent with the training data. In contrast, the simplest prompt construction (D3) has a relatively minor impact on the results compared to our default configuration for ConCIR (D4) (-0.03% mAP@10).

Codebook in CATI (E). We provide a comparative analysis of the scores obtained when the codebook module is omitted, wherein the output of the MLP is directly used as the token representations for attributes and affordances, bypassing the quantization process of the



Figure 5: Examples of retrieval results for ConCIR and state-of-the-art CIR methods.

codebook. Not only in the CIRCO dataset, the integration of the codebook could consistently improve the model, showing the discrete nature of visual concepts. We compare CATI with codebook sizes 8, 64, and 2048 (E2–E4). Notably, size 64 yields the best results, while size 8 shows limited gains, and size 2048 saturates with a marginal drop—indicating that a too-small codebook lacks capacity, whereas an overly large one introduces redundant parameters and diminishes effectiveness.

LLM Comparison (F). The LLM employed in our task was Qwen3:8B [53]. We also compare the performance of models based on other LLMs. The quality of data produced by different LLMs varied, leading to certain fluctuations in the evaluation scores. Furthermore, we measure the quality and hallucination of LLMs by randomly selecting 100 generated data samples for CATI training and conducting a manual inspection to assess the accuracy of the generated attributes and affordances. The accuracy rates were 83% (415/500) for attributes and 80.6% (403/500) for affordances, respectively, showing the good quality of LLM-generated data. The hallucination issue could be further alleviated by our concept quantization mechanism acting as a filter for semantics learning.

4.4 Visualization

We also qualitatively examine the proposed approach. To illustrate ConCIR's advantages over text inversion baselines, we show some examples of composed image retrieval tasks in Fig. 5. Among the failed cases for the MoTaDual and LinCIR methods, ConCIR could accurately capture the attribute, affordance, or relation of the object or scene. For example, in the first case, both MoTaDual and LinCIR fail to capture the color attribute "white" of the bear. In the second case, MoTaDual misses the "pink" attribute while LinCIR omits the boat and water. Overall, ConCIR could handle complex queries with more details, and we believe that the explicit learning of object and scene concepts is key for out-of-distribution content

Method	CIRR val						CIRCO val			
	R@1	R@5	R@10	R_s@1	R_s@2	R_s@3	mAP@5	mAP@10	mAP@25	mAP@50
LinCIR	**25.07**	**54.36**	**67.83**	57.62	77.80	88.64	**10.61**	**12.27**	**13.82**	**14.52**
+ Image-Only	5.62	21.65	31.95	21.14	41.47	61.35	1.21	1.59	2.12	2.36
+ Caption-Only	19.92	43.79	55.20	**61.78**	**80.55**	**90.24**	2.59	3.14	3.67	4.00
ConCIR (**Ours**)	**24.49**	**52.31**	**66.01**	54.75	76.06	88.02	**16.92**	**18.09**	**20.14**	**20.84**
+ Image-Only	5.84	22.22	33.87	22.17	43.24	63.09	2.76	3.21	4.09	4.56
+ Caption-Only	19.92	43.79	55.20	**61.78**	**80.55**	**90.24**	2.59	3.14	3.67	4.00
MoTaDual	**28.06**	**58.26**	**71.42**	62.62	81.70	90.55	**21.90**	**23.05**	**24.78**	**25.64**
+ Image-Only	5.62	20.83	32.46	20.90	41.02	61.42	2.43	2.75	3.58	4.02
+ Caption-Only	24.61	49.94	62.59	**65.92**	**83.52**	**91.87**	4.92	5.27	5.85	6.18

Table 7: Modality redundancy for CIRR and CIRCO validation sets (CLIP-L backbone). **Image-Only**: retrieve images without relative caption. **Text-Only**: retrieve images without reference image. The CIRR dataset has a higher modality redundancy level.

retrieval, such as the CIR task.

4.5 Modality Redundancy

We discuss an inconsistency in model performance on CIRR, which indicates the **modality redundancy** of the benchmarks themselves. For CIR task, an ideal composite query should necessitate **both** modalities to effectively retrieve the desired target. Any bias towards a specific modality leads to a biased model evaluation and preference for model overfitting the modality. Therefore we conducted image-only and text-only retrieval experiments on the CIRR and CIRCO validation sets in Tab. 7. Compared to CIRCO, the CIRR dataset shows a much higher level of redundancy, particularly in the subset retrieval task: using only the relative caption yields competitive performance on CIRR, even surpassing composite image-text (CIR) methods. This observation aligns with the findings of [21] and is consistent with the design strategies employed in CIRCO [3] to mitigate modality redundancy.

In comparison, ConCIR primarily focuses on enriching information in image tokens rather than optimizing text representations. Consequently, ConCIR achieves a strong performance on CIRCO, where joint modality understanding is critical. And it does not yield significant improvements on CIRR, where textual redundancy alone often suffices for competitive retrieval. This further underscores the impact of dataset design on method evaluation and highlights the need for modality-balanced benchmarks in CIR research.

5 Conclusion

We presented ConCIR, a zero-shot composed image retrieval framework that integrates hierarchical object and scene concepts to address the semantic incompleteness and domain gap of the conventional method. ConCIR generates query embeddings that align with instructions by decomposing images into context, attribute, and affordance concepts and dynamically conditioning the image encoder. Experiments validate that explicit modeling of concept hierarchies is critical for image edits, bringing over 3% performance gain on multiple benchmarks. Our work advances ZS-CIR toward human-like compositional reasoning, with broader implications for instruction-guided vision-language systems.

6 Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (U25A20442), Shanghai Municipal Science and Technology Major Project (2025SHZDZX-025G14), and CCF-Tencent Rhino-Bird Open Research Fund.

References

[1] Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, pages 1140–1149, 2021.

[2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21466–21474, 2022.

[3] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15338–15347, 2023.

[4] Jaeseok Byun, Seokhyeon Jeong, Wonjae Kim, Sanghyuk Chun, and Taesup Moon. Reducing task discrepancy of text encoders for zero-shot composed image retrieval, 2024.

[5] Junyang Chen and Hanjiang Lai. Pretrain like your inference: Masked tuning improves zero-shot composed image retrieval, 2023.

[6] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[8] Yiqun Duan, Sameera Ramasinghe, Stephen Gould, and Ajanthan Thalaiyasingam. Scaling Prompt Instructed Zero Shot Composed Image Retrieval with Image-Only Data. *arXiv e-prints*, art. arXiv:2504.00812, April 2025. doi: 10.48550/arXiv.2504.00812.

[9] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1778–1785. IEEE, 2009.

[10] David F Fouhey, Xiaolong Wang, and Abhinav Gupta. In defense of the direct perception of affordances, 2015.

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

[12] James J Gibson. The ecological approach to the visual perception of pictures. *Leonardo*, 11(3):227–235, 1978.

[13] Aaron Grattafiori, Abhimanyu Dubey, and Abhinav et al. Jauhri. The Llama 3 Herd of Models, 2024.

[14] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion, 2023.

[15] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13225–13234, 2024.

[16] Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, Trishul Chilimbi, and Abhinav Shrivastava. Collm: A large language model for composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3994–4004, 2025.

[17] Bhavin Jawade, Joao V. B. Soares, Kapil Thadani, Deen Dayal Mohan, Amir Erfan Eshratifar, Benjamin Culpepper, Paloma de Juan, Srirangaraj Setlur, and Venu Govindaraju. SCOT: Self-Supervised Contrastive Pretraining For Zero-Shot Compositional Retrieval. *arXiv e-prints*, art. arXiv:2501.08347, January 2025. doi: 10.48550/arXiv.2501.08347.

[18] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–251, 2018.

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73, May 2017. ISSN 0920-5691. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

[20] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pages 951–958. IEEE, 2009.

[21] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data Roaming and Quality Assessment for Composed Image Retrieval, March 2023.

[22] Haiwen Li, Fei Su, and Zhicheng Zhao. Motadual: Modality-task dual alignment for enhanced zero-shot composed image retrieval, 2024.

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

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

[25] Wei Li, Hehe Fan, Yongkang Wong, Yi Yang, and Mohan S Kankanhalli. Improving context understanding in multimodal large language models via multimodal composition learning. In *ICML*, volume 3, page 7, 2024.

[26] Yong-Lu Li, Yue Xu, Xinyu Xu, Xiaohan Mao, Yuan Yao, Siqi Liu, and Cewu Lu. Beyond object recognition: A new benchmark towards object concept learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20029–20040, 2023.

[27] Haoqiang Lin, Haokun Wen, Xuemeng Song, Meng Liu, Yupeng Hu, and Liqiang Nie. Fine-grained textual inversion network for zero-shot composed image retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 240–250. ACM, 2024.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, May 2014.

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[30] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134, October 2021.

[31] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33: 11525–11538, 2020.

[32] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision, 2023.

[33] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[34] Peixi Peng, Yonghong Tian, Tao Xiang, Yaowei Wang, Massimiliano Pontil, and Tiejun Huang. Joint semantic and latent attribute modelling for cross-class transfer learning. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1625–1638, 2017.

[35] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 3406–3413. IEEE, 2016.

[36] Lerrel Pinto, Dhiraj Gandhi, Yuanfeng Han, Yong-Lae Park, and Abhinav Gupta. The curious robot: Learning visual representations via physical interactions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 3–18. Springer, 2016.

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

[38] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 186–201. Springer, 2016.

[39] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314, 2023.

[40] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

[41] Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26951–26962, 2024.

[42] Yuanmin Tang, Jing Yu, Keke Gai, Gang Xiong, Gaopeng Gou, and Qi Wu. Manipulation intention understanding for accurate zero-shot composed image retrieval. 2024.

[43] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. 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, pages 5180–5188, 2024.

[44] Yuanmin Tang, Jue Zhang, Xiaoting Qin, Jing Yu, Gaopeng Gou, Gang Xiong, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Wu. Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14400–14410, 2025.

[45] Spyridon Thermos, Georgios Th Papadopoulos, Petros Daras, and Gerasimos Potamianos. Deep affordance-grounded sensorimotor object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6167–6175, 2017.

[46] Rong-Cheng Tu, Wenhao Sun, Hanzhe You, Yingjie Wang, Jiaxing Huang, Li Shen, and Dacheng Tao. Multimodal Reasoning Agent for Zero-Shot Composed Image Retrieval. *arXiv e-prints*, art. arXiv:2505.19952, May 2025. doi: 10.48550/arXiv.2505.19952.

[47] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[48] Sagar Vaze, Nicolas Carion, and Ishan Misra. GeneCIS: A Benchmark for General Conditional Image Similarity, June 2023.

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

[50] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2596–2605, 2017.

[51] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback, May 2019.

[52] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[53] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[54] Bangpeng Yao, Jiayuan Ma, and Li Fei-Fei. Discovering object functionality. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2512–2519, 2013.

[55] Yibiao Zhao and Song-Chun Zhu. Scene parsing by integrating function, geometry and appearance models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3119–3126, 2013.

[56] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 408–424. Springer, 2014.