OCTOPUS: Multi-Agentic Universal Compositional Visual Retrieval

Zhangtao Cheng[1], Bozhu Zheng[1], Ting Zhong[1], and Fan Zhou[1,2*]

[1] University of Electronic Science and Technology of China, Chengdu, Sichuan, China
[2] Key Laboratory of Intelligent Digital Media Technology of Sichuan Province, Chengdu, Sichuan, China
{zhangtao.cheng,bozhu.zheng}@outlook.com
{zhongting,fan.zhou}@uestc.edu.cn

Abstract. Composed visual retrieval (CVR) aims to locate a target image or video that reflects a user's modification of a reference visual input. Prior works typically design separate models for images and videos, resulting in fragmented paradigms, limited reasoning diversity, and poor adaptability to diverse user intents. We introduce **OCTOPUS**, a novel multi-agentic assistant for tool-integrated progressive self-improvement and user-friendly synergistic compositional retrieval. OCTOPUS features three cooperative agents – a *Perceiver*, a *Creator*, and a *Retriever* – that collectively emulate a human-like process of perception, imagination, and reflection. The Perceiver interprets composed queries, formulates semantic instructions, and refines them through self-reflection. The Creator employs visual imagination and object-level reasoning tools to generate diverse textual and visual proxies that capture missing or ambiguous semantics. The Retriever performs bidirectional retrieval and applies an evaluation mechanism to ensure alignment between results and user intent. Comprehensive experiments on five benchmarks demonstrate that OCTOPUS consistently outperforms both training-free and supervised baselines. The code is available at `https://github.com/zbzzbzzbz/OCTOPUS`.

Keywords: Compositional Visual Retrieval · Zero-shot learning · Mixture of agents · Tool Augmentation

1 Introduction

Composed visual retrieval (CVR), encompassing composed image retrieval (CoIR) [32] and composed video retrieval (CoVR) [38], is a core vision–language task that aims to retrieve a target visual instance (*Tar*) that remains visually consistent with a reference input (*Ref*) while satisfying the semantics of a textual modification (*Mod*). By jointly reasoning over visual and linguistic modalities, CVR provides a foundation for a wide range of real-world applications, including personalized fashion recommendation [9, 19], and event localization in videos [5].

* Corresponding author.

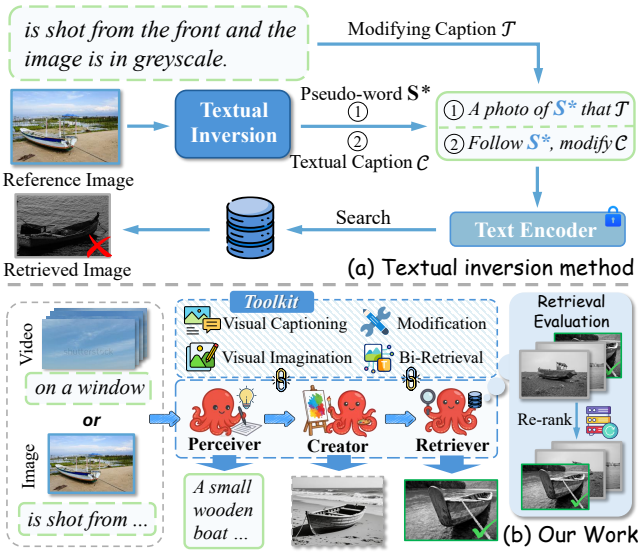


Fig. 1: Previous textual inversion method vs. our work.

Traditional CVR methods commonly adopt a late-fusion strategy [6,38], which combines the semantic representations of the *Ref* and the corresponding *Mod* to perform retrieval. These models are typically trained on human-annotated triplets {*Ref*, *Mod*, *Tar*} that provide explicit supervision. However, constructing such triplets is labor-intensive and domain-specific, thereby restricting the scalability and generalizability of learned models [3]. To overcome these limitations, recent studies [4,32] have explored zero-shot composed retrieval by leveraging pre-trained vision-language models (VLM) such as CLIP [30]. These VLM-based methods employ textual inversion techniques that transform the *Ref* into pseudo-word embeddings used for text-based retrieval, effectively removing the need for annotated triplets. In parallel, emerging works integrate large language models (LLM) to generate semantically enriched target descriptions by reasoning over image captions and modification texts [22], yielding inferred captions that serve as refined retrieval queries.

Despite these task-specific advances, several key challenges remain: (1) **Unifying Fragmented Paradigms**: Existing methods often address task-specific settings in isolation, optimizing for a particular interaction style without providing a unified framework that generalizes across CVR tasks. Such fragmented solutions hinder adaptability to diverse user requirements and limit scalability across both image and video modalities. A universal, task-agnostic system is therefore needed – one that can flexibly adapt to varying modalities, domains, and query complexities. (2) **Embracing Divergent Reasoning**: As illustrated in Fig. 1(a), most existing approaches rely on textual inversion, which tends to discard critical visual cues when compressing the *Ref* into a coarsely edited caption. This limitation makes it difficult to handle manipulations where key visual

elements are missing from the reference. To address this, incorporating divergent reasoning – such as imagining plausible visual scenes [24] or generating diverse textual hypotheses [43] – is crucial for resolving ambiguous and under-specified queries in CVR. (3) **Reimagining User-friendly Retrieval**: The ambiguity of natural language and the intricate nature of real-world visual content make prompt-based retrievers (e.g., CLIP) highly sensitive to wording and template choices. Consequently, users – especially non-experts – often resort to tedious trial-and-error refinements of their queries, leading to inefficient and frustrating retrieval experiences [28].

To address these challenges, we draw inspiration from how humans approach problem-solving: they adapt to complex environments, systematically analyze problems, inform predictions, and iteratively improve through tool use. Motivated by this perspective, we propose **OCTOPUS**, a n**O**vel multi-agenti**C** assistan**T** for tool-integrated pr**O**gressive self-im**P**rovement and **U**ser-friendly **S**ynergistic compositional retrieval. As illustrated in Fig. 1(b), OCTOPUS adaptively interprets diverse user-defined queries and progressively invokes external tools to enhance CVR under zero-shot settings. Specifically, OCTOPUS comprises three training-free agents: (1) **Perceiver**. Analyzes the composed query and formulates a strategic, intent-aligned instruction. It then invokes semantic analysis tools to guide reasoning and self-refinement, producing a concise target instruction for downstream retrieval. (2) **Creator**. Leverages visual divergence tools to enrich interpretation and mitigate ambiguities in the perceiver's predictions. It first generates visual proxies aligned with the composed query and then detects correlations between these proxies and textual instructions to derive diverse textual proxies as semantic augmentations. (3) **Retriever**. Integrates information from both agents to perform bidirectional retrieval and employs evaluation tools to assess candidate instances, verifying their alignment with the user's intended modification. Together, these three agents enable OCTOPUS to operate as a multi-agentic assistant capable of generalizable composed visual retrieval across complex real-world scenarios, mimicking a human-like perception-imagination-reflection process.

We conduct extensive experiments on five benchmark datasets and find that OCTOPUS consistently outperforms all existing CVR methods. For example, on the FashionIQ dataset, OCTOPUS achieves a notable 12.53% average improvement in Recall@10, underscoring its advantage in fine-grained compositional reasoning. These results validate the strong generalization capability of OCTOPUS across diverse domains and retrieval scenarios. To summarize, our key contributions are threefold: (1) **Generalist Compositional Retrieval**. We present the first study on the universal formulation of CVR, capable of handling arbitrary image and video categories within a single unified framework – marking a new and largely unexplored research direction. (2) **Multi-Agentic Framework**. We propose OCTOPUS, the first training-free, multi-agent AI framework for universal compositional retrieval across images and videos, enabled by tool-integrated, progressive self-improvement. (3) **Collaborative Agent Design**. We develop three plug-and-play agents – the Perceiver, Creator, and

Retriever – that interact through a *perception-imagination-reflection* pipeline, enabling *accurate*, *interpretable*, and *user-friendly* retrieval. Finally, OCTOPUS is modular and easily integrated with diverse LLM architectures and retrieval backbones, while remaining compatible with a wide range of external tools.

2 Related Work

Composed image retrieval [34] aims to retrieve a target image given a composed query. In supervised settings, traditional methods [14, 39] typically train specialized CoIR models on large-scale human-annotated triplets. These models commonly project image-text pairs into a shared embedding space using contrastive learning objectives [30, 31] or cross-modal attention mechanisms [11], aligning closely with the principles of compositional learning [21, 27]. However, such methods rely heavily on high-quality annotated data, the collection of which is both labor-intensive and costly. Recently, to address these limitations, many researchers [8,22,35,44] have introduced zero-shot strategies that leverage VLM or LLM to perform text-based image retrieval without requiring human-annotated triplets. Specifically, these methods typically either construct a mapping function that generates pseudo-tokens of reference images, or employ image captioning models to produce textual descriptions of the reference image. These representations are then combined with the textual modifications and used for retrieval.
Composed video retrieval. Recently, researchers have extended the CVR task from the image domain to the video domain [10, 38, 46]. The goal of CoVR is to retrieve target videos by composing a reference video with a textual modification description. Specifically, CoVR [38] introduces a new benchmark, WebVid, which consists of a synthetic training set and a manually curated test set. The authors adopt a supervised strategy, leveraging BLIP to perform cross-modal alignment and fusion for retrieval. CVRDE [37] improves query-specific alignment between source and target videos by utilizing detailed language descriptions of the source content. EgoCVR [16] presents a retrieval benchmark based on egocentric videos and proposes a training-free re-ranking strategy to enhance retrieval performance.

However, existing works tend to address task-specific settings in isolation and often rely on static prompt and prompt-sensitive retrievers, which struggle in complex retrieval scenarios and user-friendly search experience. In contrast, OCTOPUS is a multi-agent framework that embodies the characteristics of an agentic AI assistant, capable of autonomously performing CVR without training or manual engineering, even in complex real-world retrieval scenarios.
Tool-integrated LLMs. Tool-integrated strategies have emerged as a pivotal direction for enhancing the reasoning and planning capabilities of LLMs by enabling interaction with external tools such as search engines [7, 15, 23] and Python interpreters [12]. This tool-use capability bridges the gap between general-purpose LLMs and real-world applications, empowering LLM to acquire external knowledge and perform complex reasoning in dynamic environments [29,33]. Prior work has primarily explored two complementary approaches for enabling tool-use in LLMs: prompt engineering and instruction tuning. Prompt engineering focuses

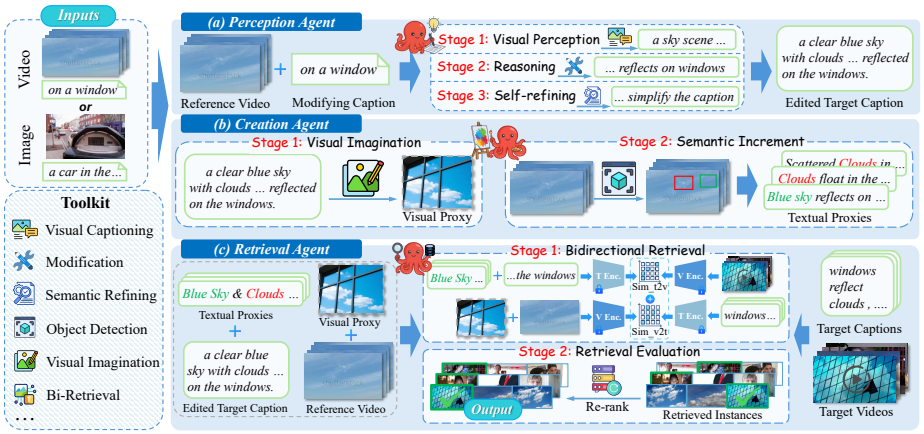


Fig. 2: Overall framework of OCTOPUS. The perceiver interprets the composed query to craft and self-refine an intent-aligned target instruction. The creator invokes visual divergence tools to progressively assess this instruction and resolve ambiguities by generating visual proxies. The retriever first performs bidirectional retrieval, and then assess and adjust the retrieved results aligned with the user's intent.

on designing effective prompts for closed-source LLMs to interact with external tools, following the thought-action format [45]. Specifically, CHAMELEON [26] orchestrates multiple tools using GPT-4 to enhance compositional reasoning, while TOOLCHAIN [48] improves planning efficiency by applying A* search in the action space. In contrast, instruction tuning aims to improve open-source LLMs by tuning them on tool-use demonstrations. Toolformer [33] pioneered this direction by integrating API calls into the language generation process. Notably, in this work, we are the first to introduce a tool-integrated strategy that designs a progressive self-improvement mechanism within a multi-agent framework to enhance the CVR task.

3 Methodology

In this section, we present a detailed description of our OCTOPUS framework. We begin by introducing the necessary preliminaries in Sec. 3.1. Then, we describe the three core components of OCTOPUS: the perception agent (Sec. 3.2), the creation agent (Sec. 3.3), and the retrieval agent (Sec. 3.4). The overall framework of our proposed OCTOPUS is illustrated in Fig. 2.

3.1 Preliminary

Let $\{\mathcal{I}_q^m, \mathcal{T}_q\}$ represent a composed query, where $\mathcal{I}_q^m$ denotes the reference visual input and $\mathcal{T}_q$ is the associated textual instruction. Here, $m \in \{I, V\}$, where I denotes a reference image and V denotes a reference video. The goal of composed

visual retrieval is to retrieve a target instance $\mathcal{I}_t \in \mathcal{D}$ from a large database $\mathcal{D}$ that the retrieved instance reflects the modifications described in $\mathcal{T}_q$, while largely preserving the visual characteristics of the reference input $\mathcal{I}_q^m$. In zero-shot settings, mainstream CVR methods typically design a customized mapping module $\varphi : \mathcal{I}_q^m \rightarrow \mathcal{Z}_q$, which projects the visual input $\mathcal{I}_q^m$ into pseudo-text tokens or a textual description $\mathcal{Z}_q$. A manually crafted prompt (e.g., "*a photo of* $\{\mathcal{Z}_q\}$ *that* $\{\mathcal{T}_q\}$") is then applied to combine the inverted representation $\mathcal{Z}_q$ and the modifying text $\mathcal{T}_q$, thereby producing a composed caption $\mathcal{T}_t$. This caption $\mathcal{T}_t$ is subsequently used for target retrieval via a VLM-based retriever, which encodes the caption using a text encoder Ψ_t and the candidate instance $\mathcal{I} \in \mathcal{D}$ using a vision encoder Ψ_v. The matching score $\mathcal{S}$ is computed using cosine similarity: $\mathcal{S} = \cos\_\text{sim}(\Psi_v(\mathcal{I}), \Psi_t(\mathcal{T}_t))$.

3.2 Perception Agent

Stage 1: Visual Perception. To unify different modality inputs in the CVR task, we adopt a language-for-vision mechanism inspired by prior works [22, 43], which extracts a holistic understanding of visual semantics from the *Ref* in linguistic form. The core insight behind this mechanism lies in the compositional nature of language. In other words, leveraging textual descriptions facilitates the integration of semantics across diverse input types and modalities. Specifically, in the CoIR task, given a reference image $\mathcal{I}_q^I$, the Perceiver invokes a *visual captioning tool* $\Phi_t^{\mathcal{C}}$ to extract global visual semantics and generate a corresponding language description: $\mathcal{C}_q^I = \Phi_t^{\mathcal{C}}(\mathcal{I}_q^I)$. Analogously, in the CoVR task, given a reference video $\mathcal{I}_q^V$, we uniformly sample N keyframes $F_q = \{f_1, f_2, \ldots, f_N\}$ and apply the same tool to obtain the textual description $\mathcal{C}_q^V = \Phi_t^{\mathcal{C}}(F_q)$. Such a strategy provides users with clearer insights into the retrieval process, thereby enabling potential human intervention.

Stage 2: Reasoning. In real-world scenarios, the CVR task often involves complex scenes with multiple objects and nuanced modifications, driven by diverse user intentions. Ideally, an intelligent semantic reasoning tool should effectively interpret the user's modification intent and adjust the reference visual caption $\mathcal{C}_q^m$ accordingly. Recent advancements in LLM, particularly in content understanding and compositional reasoning, offer a promising foundation for semantic manipulation within the language space. Motivated by this, we design an LLM-based *semantic modification tool* $\Phi_t^{\mathcal{M}}$ to produce cohesive and semantically unified target captions. Specifically, the Perceiver a executes the tool to combine the dense caption $\mathcal{C}_q^m$ with the modification text $\mathcal{T}_q$, accurately interpret the intended transformation, and generate an initial target caption: $\hat{\mathcal{T}}_t = \Phi_t^{\mathcal{M}}(\mathcal{P} \circ \mathcal{C}_q^m \circ \mathcal{T}_q)$, where $\mathcal{P}$ denotes a simple prompt template following [22]. Notably, this prompt is largely task-agnostic, enabling the model to generalize well across diverse retrieval scenarios.

Stage 3: Self-refining. Due to the inherent ambiguity of user intent descriptions, some generated captions may fail to capture the user's implicit intent when relying solely on simple prompts [8, 35]. To address this issue, we introduce a reflection

mechanism [20] based on LLM, and design a *semantic refining tool* $\Phi_t^{\mathcal{R}}$ to self-refine the initial caption $\hat{\mathcal{T}}_t$ and produce an intention-aligned caption for retrieval. The Perceiver employs $\Phi_t^{\mathcal{R}}$ to filter out inaccurate content in $\mathcal{C}_q^m$ and identify the most relevant manipulated elements aligned with the user's intent. This tool is designed to highlight key decisions that preserve the coherence and context of the *Ref* while fulfilling the manipulation intent, thereby providing a logical bridge between the original content and the final description. The refinement process is formulated as: $\mathcal{T}_t = \Phi_t^{\mathcal{R}}(\mathcal{P}_r \circ \hat{\mathcal{T}}_t \circ \mathcal{C}_q^m)$, where $\mathcal{T}_t$ denotes the refined caption, and $\mathcal{P}_r$ is a reflection prompt template. Notably, this step also helps mitigate hallucination issues that may arise during the reasoning phase.

3.3 Creation Agent

Stage 1: Visual Imagination. An agentic assistant for the CVR task should possess the ability to adaptively predict key visual elements (e.g., objects, scenes, attributes) that are missing from the reference input, guided by the manipulation text. To this end, we introduce a *visual imagination tool* $\Phi_t^{\mathcal{I}}$ to produce a visual proxy $\mathcal{V}^P$, i.e., a synthesized image generated by a controllable diffusion model [42], that aligns with the user-provided query, thereby providing additional visual cues to enhance retrieval. Specifically, given the generated *Tar* caption $\mathcal{T}_t$ and the *Ref* caption $\mathcal{C}_q^m$, the Creator invokes the visual generation tool $\Phi_t^{\mathcal{G}}$ to first infer a plausible overall scene and layout for the visual proxy, enabling better alignment with the generative model's training distribution. A visual proxy is then synthesized using a generative model (e.g., Qwen-Image [40] or Nano Banana). Since CVR is inherently a fuzzy retrieval task [43], it is unnecessary to strictly preserve the visual details of the *Ref* during the generation process. This insight motivates us to construct generative prompts by jointly reasoning over both the generated *Tar* caption and the *Ref* caption. Moreover, the Creator is model-agnostic and flexibly accommodate various generative tools. To minimize computational overhead, we primarily employ image generation tools. Video generation tools are avoided due to their substantial time cost.

Stage 2: Semantic Increment. To capture the diverse possible semantics of the composed target, the Creator invokes an *object detection tool* $\Phi_t^{\mathcal{O}}$ to decompose the visual proxy and *Ref* into a set of fine-grained, interpretable regions and recognize salient objects potentially omitted from the target caption $\mathcal{T}_t$. Directly incorporating all detected objects into $\mathcal{T}_t$ would inevitably introduce noise (e.g., irrelevant objects). To address this issue, we design an object filtering strategy in $\Phi_t^{\mathcal{O}}$ that leverages an LLM to focus on specific objects and their attributes. This process identifies objects that should be present in the target image ("existent objects") and those that should not ("nonexistent objects") by comparing the generated content with the composed query. After extracting the set of existent objects $\mathcal{O}$, we construct a simple prompt $\mathcal{P}_o$, inspired by [43], and combine it with the target caption $\mathcal{T}_t$ and the identified objects $\mathcal{O}$ as input to the LLM. This enables reasoning over object-level semantics and generates a set of diverse textual proxies: $\mathcal{T}_t^P = \left\{ T_i^A = \Phi_t^{\mathcal{O}}(\mathcal{T}_t \circ \mathcal{O} \circ \mathcal{P}_o) \;\middle|\; 0 \le i < K \right\}$, where K is the number of textual proxies.

3.4 Retrieval Agent

Stage 1: Bidirectional Retrieval. After obtaining the diverse textual proxies $\mathcal{T}_t^P$ and visual proxy $\mathcal{V}^P$, a key challenge lies in effectively integrating both sources of semantic knowledge for retrieval. A straightforward strategy is to directly combine the text-based similarity with the visual similarity based on textual and visual proxy. However, directly using the visual proxy features often places excessive emphasis on irrelevant elements (e.g., background clutter), which can overshadow the fine-grained semantics captured in the text descriptions. To address this, we design a bidirectional retrieval strategy that disentangles semantic relevance between the query and the target from the perspectives of both visual and textual modalities, and balances the two retrieval signals using an adaptive gating mechanism.

Specifically, the retrieval agent use a VLM-based *bi-retrieval tool* $\Phi_t^{\mathcal{R}}$ (i.e., CLIP) to automatically compute semantic similarity for retrieval. For the visual proxy $\mathcal{V}^p$, we calculate the vision-to-text similarity between the visual proxy and the textual caption of candidate visual instances in the database. The vision encoder $\Psi_v(\cdot)$ of $\Phi_t^{\mathcal{O}}$ is used to extract visual features of $\mathcal{V}^p$, while the text encoder $\Psi_t(\cdot)$ is used to encode the textual captions of candidate visual instances. The similarity score $\mathcal{S}_{v2t}$ is calculated:

$$\mathcal{S}_{v2t} = \frac{\Psi_t(T)^\top \Psi_i\left(\mathcal{V}_p\right)}{\|\Psi_t(T)\| \cdot \|\Psi_i\left(\mathcal{V}_p\right)\|}. \tag{1}$$

Analogously, we calculate the text-to-vision similarity score $\mathcal{S}_{t2v}$ between the textual proxies $\mathcal{T}_t^P$ and candidate visual instances, following [43]. The rationale behind this bidirectional retrieval design is based on the cross-modal alignment ability of VLM. Furthermore, we design a gating strategy [24] to balance retrieval results as follows:

$$S_f = \lambda \mathcal{S}_{t2v} + (1-\lambda)\mathcal{S}_b, \quad \mathcal{S}_b = \mathcal{S}_{t2v} * \mathcal{S}_{v2t}, \tag{2}$$

where $\mathcal{S}_b$ represents the balanced similarity and λ is the balancing parameter.

Stage 2: Retrieval Verification. Due to the limitations of VLMs in capturing nuanced details, certain retrieval outcomes may not accurately reflect user intent. Thus, we design a retrieval verification strategy that involves an overall evaluation of the retrieved candidate instances and further re-ranks them according to evaluation scores. Specifically, the Retriever invokes an LLM-based *evaluation tool* $\Phi_t^{\mathcal{E}}$ to perform pairwise comparisons between the *Ref* and the top-ranked candidate instances $\mathcal{V}_r$. By integrating the composed query and the descriptions of the retrieved candidates, the evaluation tool $\Phi_t^{\mathcal{E}}$ adaptively determines whether the candidate instances approximately meet the user's intent, offering binary results (Yes or No) along with necessary justifications. This process sequentially assesses each candidate until one meets the criteria or the threshold α is reached, which represents the maximum number of candidate instances to evaluate. If a suitable candidate is found, it is re-ranked to the top. We also consider different forms of user-edited intent, such as descriptions of desired changes, which are designed into carefully designed prompts $\mathcal{P}_v$. This process can be summarized as:

$$f = \{f_j\}_{j=1}^{\alpha} = \Phi_t^{\mathcal{E}}(\mathcal{V}_r, C_{1j}, \mathcal{P}_v),$$
$$\{C_{21}, C_{22}, \ldots, C_{2\alpha}\} = \text{argsort}_{\downarrow}(f), \tag{3}$$

where $f \in \mathbb{R}^{1\times\alpha}$ denotes the binary results for candidate instances, and $\{C_{21}, C_{22}, \ldots, C_{2\alpha}\}$ is the final ranking of candidates. With such design, OCTOPUS leverages multimodal collective reasoning to ensure that the top-ranked candidates satisfies user intent both in fine-grained details and overall semantic alignment.

4 Experiments

In this section, we first introduce the five evaluation benchmarks and experimental setup (Sec. 4.1). Quantitative results and comparisons with baselines are presented in Sec. 4.2. A detailed ablation study is provided in Sec. 4.3, followed by additional analyses and discussions in Sec. 4.4.

4.1 Experimental Settings

Datasets and Metrics. For effective evaluation, we adopt five benchmarks spanning three distinct tasks (i.e., CoIR and CoVR) [22, 38, 46]. Specifically, we use: (1) *Natural image domain*: CIRCO [4] and CIRR [25]; (2) *Fashion image domain*: FashionIQ [41]; and (3) *Video domain*: WebCVR [38] and FineCVR [46]. CIRCO contains 123,403 images (4.53 ground-truth images per query). CIRR includes 36,554 triplets from 21,552 images. FashionIQ has three categories (Dress, Shirt, Toptee) with 30,135 triplets and 77,683 images. WebCVR provides 2.5K annotated triplets, and FineCVR contains 1,000,028 training and 6,364 test triplets.

Following prior works [32,38], we use Recall@k (R@k) as the evaluation metric on CIRR, FashionIQ, WebCVR, and FineCVR, and mean average precision@k (mAP@k) for CIRCO, where each query has multiple ground-truth targets.

Baselines. To enable a comprehensive comparison, we select a set of competitive baselines. For the CoIR task, **Pic2Word** [32], and **LinCIR** [13] project image features into pseudo-word tokens, which are then fused with the text prompt for retrieval via CLIP. **CIReVL** [22], **LDRE** [43], **AutoCIR** [8], **IP-CIR** [24], and **OSrCIR** [35] leverage LLM by combining the caption of the reference image with the textual modification, utilizing the language reasoning capabilities of LLMs to guide retrieval. For the CoVR task, **CoVR** [38], **CVRDE** [37], and **CVRDM** [36] jointly encode the reference video and textual instruction to produce a unified representation for retrieval.

Implementation. For the retriever, we employ ViT-B/32 and ViT-L/14 CLIP models from OpenAI [30], along with the ViT-G/14 CLIP from OpenCLIP [18]. The feature dimensions are as follows: 512 for B/32, 768 for L/14, and 1024 for G/14. By default, we use GPT-4o-Mini [1] as our LLM, with additional ablations considering both Qwen2.5-VL [2], MiniGPT-4 [47] GPT-4o [17]. For visual generation, we use Qwen-Image [40] to produce visual proxy.

Table 1: Comparison results on the WebVid, FineCVR and CIRCO test sets. The best and second-best scores are highlighted in bold and underlined, respectively. '-' indicates results not reported in the original paper.

Backbone	Model	Training	WebVid				FineCVR				CIRCO			
			Recall@k				Recall@k				mAP@k			
			$k=1$	$k=5$	$k=10$	$k=50$	$k=1$	$k=5$	$k=10$	$k=50$	$k=5$	$k=10$	$k=25$	$k=50$
			CoIR											
ViT-B/32	LinCIR	✓	29.84	53.80	63.86	86.26	6.08	21.29	31.69	59.88	3.83	3.92	4.42	4.72
	CIReVL	✗	32.07	56.89	67.82	88.72	8.78	24.04	33.08	61.08	14.94	15.42	17.00	17.82
	LDRE	✗	31.17	56.81	67.66	88.29	6.52	17.65	25.20	49.10	17.96	18.32	20.21	21.11
	AutoCIR	✗	33.91	61.08	70.99	89.51	9.38	25.80	36.61	65.07	18.82	19.41	21.38	22.32
	IP-CIR	✗	37.20	62.88	73.49	91.58	7.34	21.07	30.12	55.58	–	–	–	–
	OSrCIR	✗	33.63	58.38	69.58	88.37	9.02	25.11	34.57	61.33	18.04	19.17	20.94	21.85
			CoVR											
	CoVR	✗	31.56	53.64	63.08	83.01	13.84	37.65	50.05	77.51	4.55	4.83	5.54	5.98
	CVRDE	✗	31.17	55.32	66.01	85.87	15.16	39.71	53.11	81.11	3.55	4.00	4.56	4.94
	CVRDM	✗	31.95	57.05	66.91	86.88	16.37	40.96	54.26	81.71	6.55	6.97	7.89	8.48
	OCTOPUS	✗	**57.83**	**79.17**	**85.16**	**95.73**	**19.56**	**48.33**	**62.27**	**87.88**	**21.62**	**22.28**	**25.34**	**26.47**
			CoIR											
ViT-L/14	Pic2Word	✓	48.47	74.67	83.75	95.34	7.68	23.15	33.22	59.46	8.72	9.51	10.64	11.29
	LinCIR	✓	46.20	74.47	82.62	95.97	9.15	25.86	36.57	64.28	12.59	13.58	15.00	15.85
	CIReVL	✗	38.76	64.84	74.94	92.56	10.07	25.85	36.38	62.49	18.57	19.01	20.89	21.80
	LDRE	✗	41.20	70.01	78.86	93.77	8.33	22.03	30.26	54.46	23.35	24.03	26.44	27.50
	AutoCIR	✗	45.61	71.93	80.85	93.70	11.17	30.20	41.34	68.40	24.05	25.14	27.35	28.36
	IP-CIR	✗	45.46	71.81	79.87	94.36	9.43	24.72	34.51	59.05	26.43	27.41	29.87	31.07
	OSrCIR	✗	40.80	66.60	76.59	91.86	11.25	27.92	38.70	65.93	23.87	25.33	27.84	28.97
			CoVR											
	CoVR	✗	37.27	63.04	72.59	88.29	16.33	40.89	54.18	79.90	6.38	7.11	8.21	8.82
	CVRDE	✗	37.24	65.27	74.86	91.35	16.33	40.52	55.09	81.29	5.56	6.35	7.49	8.09
	CVRDM	✗	39.23	67.07	76.86	91.66	17.25	42.57	56.33	82.23	9.90	10.59	11.99	12.63
	OCTOPUS	✗	**67.15**	**85.28**	**90.49**	**97.73**	**20.16**	**48.81**	**62.57**	**87.21**	**28.18**	**28.66**	**31.09**	**32.26**

4.2 Main Results

In Table 1, we present comparative evaluations on the WebVid, FineCVR, and CIRCO datasets, highlighting the ability of OCTOPUS to distinguish between foreground and background elements and to perform fine-grained visual editing via object- and scene-level manipulations across both image and video retrieval scenarios. Across all CLIP-based retrieval variants, OCTOPUS consistently outperforms all competitive CVR models in every setting, without requiring any task-specific training. Notably, with the ViT-B/32 backbone, OCTOPUS achieves a substantial average improvement of 17.81% across all metrics on the WebVid and FineCVR datasets, which are characterized by less explicit and noisier textual instructions, compared to the best-performing baselines. Furthermore, on the CIRCO dataset, which provides clean editing annotations and multiple ground-truth target images, OCTOPUS achieves a mAP@5 of 21.62%, outperforming the best training-free method (AutoCIR) by 14.87%, and the second-best baseline (OSrCIR) by 19.84%. Notably, we observe that CoVR models generally exhibit better performance on CoVR datasets, but underperform compared to CoIR models in CoIR scenarios. A similar trend is observed for CoIR models, further reinforcing our motivation to develop a unified framework for the CVR task. Nevertheless, our OCTOPUS consistently surpasses all CVR models by a considerable margin, demonstrating the effectiveness of our multi-agent framework across diverse retrieval scenarios.

Table 2: Results on the FashionIQ dataset. The best and second-best scores are shown in bold and underlined, respectively.

Backbone	Model	Training	Shirt		Dress		Toptee		Average	
			Recall@k		Recall@k		Recall@k		Recall@k	
			$k=10$	$k=50$	$k=10$	$k=50$	$k=10$	$k=50$	$k=10$	$k=50$
ViT-B/32	*CoIR*									
	LinCIR	✓	17.81	33.61	14.97	34.16	20.04	40.29	17.61	36.02
	CIReVL	✗	28.36	47.84	25.29	46.36	31.21	53.85	28.29	49.35
	LDRE	✗	27.38	46.27	19.97	41.84	27.07	48.78	24.81	45.63
	AutoCIR	✗	<u>32.43</u>	<u>51.67</u>	26.52	46.36	33.96	56.09	30.97	51.37
	IP-CIR	✗	26.20	44.79	22.16	43.97	26.67	47.62	25.01	45.47
	OSrCIR	✗	31.16	51.13	<u>29.35</u>	<u>50.37</u>	<u>36.51</u>	<u>58.71</u>	<u>32.34</u>	<u>53.40</u>
	CoVR									
	CoVR	✗	21.00	36.90	18.00	35.89	23.61	40.69	20.87	37.83
	CVRDE	✗	16.73	30.96	9.97	24.34	16.83	30.55	14.51	28.62
	CVRDM	✗	25.12	42.30	18.64	37.48	26.77	45.89	23.51	41.89
	OCTOPUS	✗	**36.41**	**56.62**	**32.37**	**53.99**	**41.66**	**62.42**	**36.81**	**57.67**
ViT-L/14	*CoIR*									
	Pic2Word	✓	26.20	43.60	20.00	40.20	27.90	47.40	24.70	43.70
	LinCIR	✓	29.10	46.81	20.92	42.44	28.81	50.18	26.28	46.49
	CIReVL	✗	29.49	47.40	24.79	44.76	31.36	53.65	28.55	48.57
	LDRE	✗	31.04	51.22	22.93	46.76	31.57	53.64	28.51	50.54
	AutoCIR	✗	<u>34.00</u>	<u>53.43</u>	24.94	45.81	33.10	55.58	30.68	51.60
	IP-CIR	✗	31.40	49.16	25.58	48.43	32.63	53.69	29.87	50.43
	OSrCIR	✗	33.17	52.03	<u>29.70</u>	<u>51.81</u>	<u>36.92</u>	<u>59.27</u>	<u>33.26</u>	<u>54.37</u>
	CoVR									
	CoVR	✗	29.24	46.37	23.60	43.63	31.16	49.77	28.00	46.59
	CVRDE	✗	23.50	39.01	12.59	29.25	20.55	36.66	18.88	34.97
	CVRDM	✗	33.27	49.21	23.85	43.28	32.99	52.83	30.04	48.44
	OCTOPUS	✗	**39.70**	**58.93**	**31.14**	**53.20**	**40.18**	**60.68**	**37.00**	**57.60**

We further evaluate our model's capability on attribute manipulation tasks using the FashionIQ dataset, which requires accurate localization and understanding of fine-grained fashion attributes (e.g., style and color), as shown in Table 2. When using both ViT-B and ViT-L backbones, OCTOPUS significantly outperforms all strong CVR methods across all metrics, achieving an average gain of +12.53% in Recall@10 and +6.96% in Recall@50. This improvement is attributed to OCTOPUS's enhanced ability to use the creation agent to generate visual proxies that incorporate target attributes missing from the reference image, guided by the manipulation intent. Additionally, OCTOPUS employs the retriever agent to assess and adjust on retrieval outcomes, thereby reducing semantic bias and ultimately enhancing CLIP-based retrieval performance.

4.3 Ablation Studies

Following prior works [22, 43], an ablation study is to evaluate each module using the ViT-L/14 on the WebVid and CIRCO datasets.

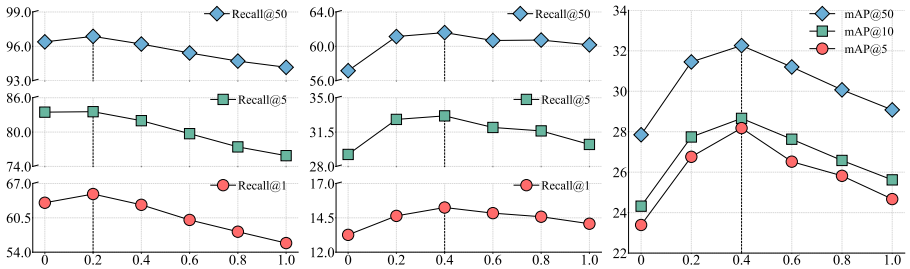
Table 3: Ablation study of the core components within OCTOPUS.

Method	WebVid				CIRCO			
	k=1	k=5	Avg.	Δ	k=5	k=10	Avg.	Δ
1. **OCTOPUS**	**67.15**	**85.28**	76.21	0.00	**28.18**	**28.66**	28.42	0.00
Impact of Perception Agent								
2. w/o Reasoning	60.34	81.01	70.68	-5.53	26.40	27.51	26.95	-1.47
3. w/o Refining	60.45	80.70	70.57	-5.65	24.67	25.62	25.14	-3.28
Impact of Creation Agent								
4. w/o T-Proxy	62.06	81.28	71.67	-4.55	23.84	25.02	24.43	-3.99
5. w/o V-Proxy	61.90	81.01	71.45	-4.77	25.26	26.24	25.75	-2.67
6. w/o Creator	59.44	79.21	69.32	-6.90	22.84	23.76	23.30	-5.12
Impact of Retrieval Agent								
7. w/o Bi-Retrieval	61.51	81.48	71.49	-4.73	26.91	27.48	27.19	-1.23
8. w/o Verification	52.70	78.58	65.64	-10.58	26.36	27.30	26.83	-1.59
Impact of different LLMs								
9. MiniGPT-4	60.88	80.54	70.71	-5.51	23.48	24.25	23.87	-4.55
10.Qwen-VL	65.00	83.56	74.28	-1.94	27.06	28.13	27.60	-0.82
11.GPT-4o	67.31	85.90	76.60	+0.39	28.45	29.10	28.77	+0.35

(1) Models '2-3' analyze the impact of key modules in the perception agent: Removing the modification tool during the reasoning stage (model '2') results in a 3.94% average performance drop. Additionally, removing the refinement tool from the perceiver (model '3') leads to a 7.28% average performance decline compared to the full OCTOPUS. These results underscore the importance of LLM-driven reasoning and refinement mechanisms for accurately interpreting manipulation intentions in composed queries.

(2) Models '4-6' evaluate the impact of the creation agent: Skipping the semantic increment process (model '4') leads to a 7.96% average performance drop. Removing the visual proxy generation (model '5') results in a 5.59% decline. Furthermore, removing the entire creation agent (model '6') yields an 11.43% performance decrease. These results validate our motivation for embracing divergent reasoning and further support the view that the CVR task is inherently a fuzzy retrieval problem. Importantly, the findings highlight the effectiveness of generating diverse textual and visual proxy via visual divergence tools in enhancing CVR.

(3) Models '7-8' evaluate the impact of each component in the retrieval agent: In model '7', we replace our bidirectional retrieval mechanism with a simple retrieval strategy (i.e., text-to-vision and vision-to-vision), resulting in a performance drop of 2.97%. This highlights the importance of cross-modal alignment in CLIP-based retrieval. Similarly, removing the retrieval verification step (model '8') leads to a more substantial performance decline of 7.56%. This indicates that the verification step plays a critical role in compensating for the lack of task-specific alignment in the CLIP, particularly in the video domain, where CLIP may lack sufficient domain-specific understanding.



(a) WebVid **(b)** Toptee **(c)** CIRCO

Fig. 3: Sensitivity analysis of the hyperparameter λ.

(4) Models '9-11' analyze the impact of the choice of LLM: We observe that open-source LLM (e.g., MiniGPT-4 (Model '9') and Qwen-VL (Model '10')) achieve retrieval performance that is nearly comparable to the best-performing baseline. However, there remains a performance gap of 11.61% and 2.71% compared to GPT-4o-mini. Notably, GPT-4o-mini performs comparably well, with only a 0.87% drop in performance, while offering significantly higher efficiency than GPT-4o. This result indicates that OCTOPUS is modular and can be seamlessly integrated with diverse LLM architectures, demonstrating its potential to further enhance CVR through the use of more powerful LLM.

4.4 Further Analysis

• **Effects of the weight** λ. Since our OCTOPUS incorporates textual and visual proxies, we conduct a sensitivity analysis to evaluate the impact of the weighting factor λ on performance. The results on the WebVid, CIRCO and the Toptee category of the FashionIQ dataset are shown in Fig. 3. Notably, the performance of OCTOPUS peaks at an optimal value of λ, beyond which increasing the weight leads to a performance decline. This suggests that while the visual proxy contributes positively to retrieval, overemphasizing it can overshadow important textual semantics. We observe that smaller values of λ yield better results. In particular, performance remains relatively stable within the range of $\lambda \in [0.2, 0.4]$ for Toptee and CIRCO, and $\lambda = 0.2$ offers the best result on WebVid. These findings validate our design of integrating proxy information for enhancing CVR.
• **Qualitative Results**. Fig. 4 visualizes successful retrievals where the instructions impact different semantic elements, such as color change, and object insertion. Compared to two baselines (i.e., LDRE and OSrCIR), we observe that these methods, even when equipped with the LLM's reasoning, often fail to attend to the relevant visual aspects. In Case (b), both models overlook critical terms, e.g., "v-neck" and "higher neckline", highlighting the limitations of fixed retrieval strategies in handling complex queries involving multiple objects and nuanced attribute modifications. Additionally, OSrCIR neglects terms like "leaf clover" and "no text" in Case (c), leading to worse results than LDRE, which benefits from divergent composed descriptions. In contrast, OCTOPUS effectively

(a) WebVid **(b)** Toptee **(c)** Shirt

Fig. 4: Qualitative comparison of three retrieval mechanisms based on the Top-4 results on the WebVid and FashionIQ datasets. Reference visual input and target instances are highlighted with purple and green outlines, respectively.

Fig. 5: Visualization of a failure case from the FashionIQ validation set. The Top-2 retrieved results are displayed.

captures all intended modifications and accurately retrieves the target instances. This success is attributed to a creator and retriever, which enable the generation of textual and visual proxies to offer complementary visual cues and self-reflect on refining retrieval outputs.

• **Analysis of Failure Cases**. Fig. 5 visualizes exemplary failure cases of OCTOPUS. Since OCTOPUS primarily performs CVR in the language domain, these errors are generally interpretable by users. Specifically, we observe that OCTOPUS generates appropriate retrieval instructions aligned with the target image; however, it still retrieves incorrect results. This discrepancy is likely due to conceptual misalignment between the LLM and the retrieval model. The retrieval backbone struggles to interpret fashion-specific terms generated by the LLM, such as "exotic patterns". Replacing these with simpler terms (e.g., "bold floral print") leads to more accurate retrieval.

• **Scalability Analysis**. To validate the scalability of our design, we select two strong baselines representing different text-based retrieval paradigms: LinCIR (a textual inversion method) and OSrCIR (an LLM-induced method), and integrate our creator and retriever into each. As shown in Fig. 6, we observe consistent performance improvements across both models on the WebVid dataset. These results demonstrate the effectiveness and scalability of our tool-augmented, agent-based design in capturing contextual knowledge from textual and visual proxies, and in self-reflecting on retrieval results.

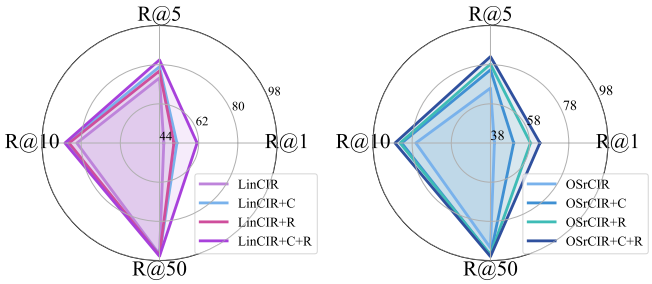


Fig. 6: Performance improvements of several baselines (i.e., LinCIR and OSrCIR) when augmented with the creation agent (C) and the retriever agent (R).

• **Efficiency Analysis**. Our OCTOPUS consistently outperforms LLM-induced methods (IP-CIR, AutoCIR, and OSrCIR), while maintaining interactivity with an average inference time of 5.3 seconds per query. Notably, OCTOPUS achieves 84.33% faster inference than IP-CIR (33.8s) and 12.67% faster than AutoCIR (6.07s), while being slower than OSrCIR (0.6s). Despite this, our model offers a significantly more favorable trade-off between latency and retrieval performance. Compared to textual inversion methods, our model achieves higher retrieval accuracy without requiring any training, although inference remains approximately 40× slower. Since over 95% of OCTOPUS's runtime is spent on external API calls, further latency reductions could be achieved through faster APIs.

5 Conclusion

In this study, we introduce OCTOPUS, a novel AI agent framework for generic, flexible, and interpretable compositional retrieval across any image and video domains. OCTOPUS comprises three agents, a perceiver, a creator, and a retriever, that collaboratively invoke tools to iteratively extract and refine task-relevant information, mimicking a human-like perception-imagination-reflection process. Extensive experiments on five benchmark datasets demonstrated the superior performance of our framework compared to existing methods. In future work, we plan to design a lightweight multi-agentic framework to reduce computational overhead making it more suitable for real-world retrieval systems. Additionally, we aim to extend OCTOPUS to other domains, such as visual super-resolution and autonomous driving.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62572097), the Key Scientific and Technological Research Project of Henan Province (Grant No. 262102210054), and the Natural Science Foundation of Henan Province (Grant No. 262300420698).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Bai, Y., Xu, X., Liu, Y., Khan, S., Khan, F., Zuo, W., Mong, R.S.M., Feng, C.M.: Sentence-level prompts benefit composed image retrieval. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024)
4. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-shot composed image retrieval with textual inversion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15338–15347 (2023)
5. Chen, H., Wang, X., Chen, H., Zhang, Z., Feng, W., Huang, B., Jia, J., Zhu, W.: Verified: A video corpus moment retrieval benchmark for fine-grained video understanding. Advances in Neural Information Processing Systems (NeurIPS) **37**, 40393–40406 (2024)
6. Chen, Y., Bazzani, L.: Learning joint visual semantic matching embeddings for language-guided retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 136–152. Springer (2020)
7. Cheng, Z., Lang, J., Zhong, T., Zhou, F.: Seeing the unseen in micro-video popularity prediction: Self-correlation retrieval for missing modality generation. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 142–152 (2025)
8. Cheng, Z., Ma, Y., Lang, J., Zhang, K., Zhong, T., Wang, Y., Zhou, F.: Generative thinking, corrective action: User-friendly composed image retrieval via automatic multi-agent collaboration. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). pp. 334–344 (2025)
9. Cheng, Z., Zhang, J., Xu, X., Trajcevski, G., Zhong, T., Zhou, F.: Retrieval-augmented hypergraph for multimodal social media popularity prediction. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 445–455 (2024)
10. Cheng, Z., Zheng, B., Zhong, T., Zhou, F.: Evidential matching, uncertainty calibration: Towards robust composed video retrieval with noisy triplets. In: Proceedings of the ACM Web Conference. pp. 2373–2383 (2026)
11. Delmas, G., de Rezende, R.S., Csurka, G., Larlus, D.: ARTEMIS: attention-based retrieval with text-explicit matching and implicit similarity. In: Proceedings of the International Conference on Learning Representations (ICLR) (2022)
12. Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., Neubig, G.: Pal: Program-aided language models. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 10764–10799. PMLR (2023)
13. Gu, G., Chun, S., Kim, W., Kang, Y., Yun, S.: Language-only efficient training of zero-shot composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13225–13234 (2024)
14. Han, X., Wu, Z., Huang, P.X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., Davis, L.S.: Automatic spatially-aware fashion concept discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1463–1471 (2017)

15. Hong, R., Lang, J., Zhong, T., Wang, Y., Zhou, F.: Tameing long contexts in personalization: Towards training-free and state-aware mllm personalized assistant. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 452–463 (2026)
16. Hummel, T., Karthik, S., Georgescu, M.I., Akata, Z.: Egocvr: An egocentric benchmark for fine-grained composed video retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 1–17 (2024)
17. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
18. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip, `https://doi.org/10.5281/zenodo.5143773`
19. Islam, S.M., Joardar, S., Sekh*, A.A.: A survey on fashion image retrieval. ACM Computing Surveys **56**(6), 1–25 (2024)
20. Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., Fung, P.: Towards mitigating llm hallucination via self reflection. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2023)
21. Karthik, S., Mancini, M., Akata, Z.: Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9336–9345 (2022)
22. Karthik, S., Roth, K., Mancini, M., Akata, Z.: Vision-by-language for training-free compositional image retrieval. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024)
23. Komeili, M., Shuster, K., Weston, J.: Internet-augmented dialogue generation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). pp. 8460–8478 (2022)
24. Li, Y., Ma, F., Yang, Y.: 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). pp. 3984–3993 (2025)
25. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2125–2134 (2021)
26. Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K.W., Wu, Y.N., Zhu, S.C., Gao, J.: Chameleon: Plug-and-play compositional reasoning with large language models. Advances in Neural Information Processing Systems (NeurIPS) **36**, 43447–43478 (2023)
27. Mancini, M., Naeem, M.F., Xian, Y., Akata, Z.: Open world compositional zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5222–5230 (2021)
28. Pavlichenko, N., Ustalov, D.: Best prompts for text-to-image models and how to find them. In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). pp. 2067–2071 (2023)
29. Qin, Y., Hu, S., Lin, Y., Chen, W., Ding, N., Cui, G., Zeng, Z., Zhou, X., Huang, Y., Xiao, C., et al.: Tool learning with foundation models. ACM Computing Surveys **57**(4), 1–40 (2024)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)

31. Roth, K., Vinyals, O., Akata, Z.: Integrating language guidance into vision-based deep metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16177–16189 (2022)
32. Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T.: Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19305–19314 (2023)
33. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems (NeurIPS) **36**, 68539–68551 (2023)
34. Song, X., Lin, H., Wen, H., Hou, B., Xu, M., Nie, L.: A comprehensive survey on composed image retrieval. ACM Transactions on Information Systems (TOIS) (2025)
35. Tang, Y., Zhang, J., Qin, X., Yu, J., Gou, G., Xiong, G., Lin, Q., Rajmohan, S., Zhang, D., Wu, Q.: 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). pp. 14400–14410 (2025)
36. Thawakar, O., Demidov, D., Thawkar, R., Anwer, R.M., Shah, M., Khan, F.S., Khan, S.: Beyond simple edits: Composed video retrieval with dense modifications. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV) (2025)
37. Thawakar, O., Naseer, M., Anwer, R.M., Khan, S., Felsberg, M., Shah, M., Khan, F.S.: Composed video retrieval via enriched context and discriminative embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26896–26906 (2024)
38. Ventura, L., Yang, A., Schmid, C., Varol, G.: Covr: Learning composed video retrieval from web video captions. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 38, pp. 5270–5279 (2024)
39. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J.: Composing text and image for image retrieval-an empirical odyssey. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6439–6448 (2019)
40. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
41. Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., Feris, R.: 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 (CVPR). pp. 11307–11317 (2021)
42. Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H.: Diffusion models: A comprehensive survey of methods and applications. ACM computing surveys **56**(4), 1–39 (2023)
43. Yang, Z., Xue, D., Qian, S., Dong, W., Xu, C.: Ldre: Llm-based divergent reasoning and ensemble for zero-shot composed image retrieval. In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). pp. 80–90 (2024)
44. Yang, Z., Pang, W., Yuan, Y.: Xr: Cross-modal agents for composed image retrieval. In: Proceedings of the ACM Web Conference 2026. pp. 2071–2082 (2026)

45. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023)
46. Yue, W., Qi, Z., Wu, Y., Sun, J., Wang, Y., Wang, S.: Learning fine-grained representations through textual token disentanglement in composed video retrieval. In: Proceedings of the International Conference on Learning Representations (ICLR). vol. 3, p. 4 (2025)
47. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024)
48. Zhuang, Y., Chen, X., Yu, T., Mitra, S., Bursztyn, V.S., Rossi, R.A., Sarkhel, S., Zhang, C.: Toolchain*: Efficient action space navigation in large language models with a* search. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024)