DIFFUSION PRIOR-DRIVEN PSEUDO TARGET EXTRACTION FOR TRAINING-FREE ZERO-SHOT COMPOSED IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

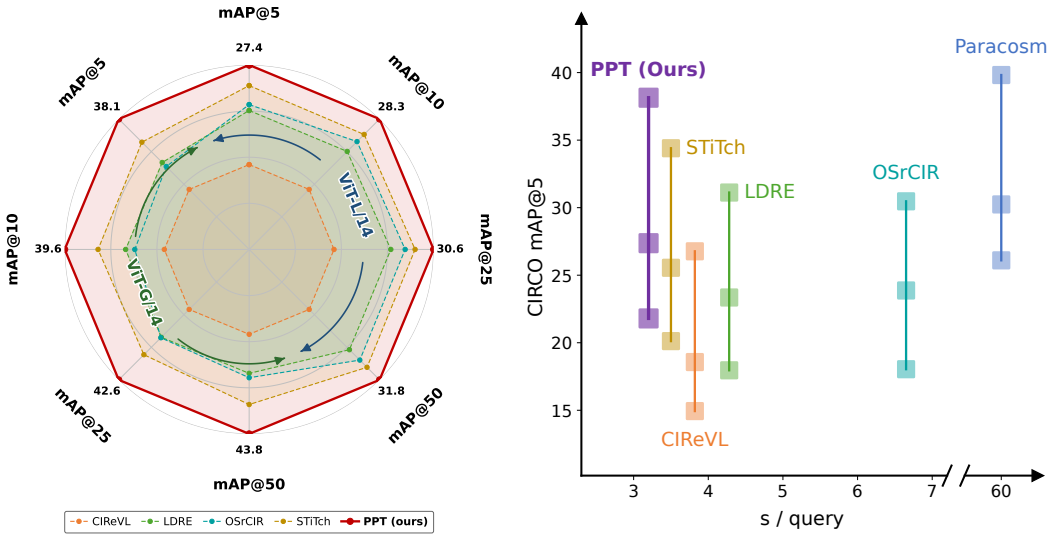
Zero-shot Composed Image Retrieval (ZS-CIR) aims to retrieve relevant images from a gallery using both a reference image and manipulation text without requiring task-specific training. Recent approaches imagine the target by generating a pseudo target image in pixel space with an image generation model such as Qwen-Image-Edit, and then extract its embedding with the CLIP image encoder. However, the pseudo target image may introduce redundant image details that mislead retrieval, and the image generation and re-encoding incur substantial computational overhead. To address these issues, we propose Diffusion Prior-driven Pseudo Target Extraction (PPT), a training-free framework that extracts the pseudo target embedding directly with a diffusion prior in the CLIP image embedding space. Without any diffusion decoder or CLIP image encoder at query time, PPT achieves competitive performance on object/scene manipulation and attribute manipulation tasks with much lower latency. Our code will be publicly released upon publication.

1 INTRODUCTION

Composed image retrieval (CIR) (Vo et al., 2019; Liu et al., 2021) aims to retrieve a target image visually similar to a reference image while incorporating the visual modifications specified in a manipulation text. CIR is critical in various practical applications, including internet search and e-commerce applications (Wu et al., 2021; Liu et al., 2021). Different from conventional content-based image retrieval (Gordo et al., 2016; Huang, 1997; Deng et al., 2018; Gordo et al., 2020; Chen et al., 2025c) and text-to-image retrieval (Rasiwasia et al., 2010; Wang et al., 2017; Tian et al., 2025; Wang et al., 2022; Li et al., 2025a; Cui et al., 2024; An et al., 2023; Jiang et al., 2025), where the query comes from a single modality, CIR requires jointly reasoning over the visual content to be preserved and the textual modification to be applied, making it a more flexible yet more challenging retrieval paradigm (Du et al., 2025; Wan et al., 2025). Supervised CIR methods require extensive triplet annotations for training (Vo et al., 2019; Baldrati et al., 2022), which is labor-intensive and limits their generalizability. Recently, zero-shot CIR (ZS-CIR) methods have gained significant attention (Saito et al., 2023; Baldrati et al., 2023), which leverage CLIP (Radford et al., 2021) to cast this task into text-image retrieval.

Query generation methods in ZS-CIR can be training-based or training-free. Textual inversion methods like Pic2Word (Saito et al., 2023) and SEARLE (Baldrati et al., 2023) train a mapping network to map the reference image into a pseudo token in the language space and concatenate this token with the manipulation text through a static template to form the composed query. Two-stage training-free methods, such as CIReVL (Karthik et al., 2024) and LDRE (Yang et al., 2024), caption the reference image with a captioner (e.g., BLIP-2 (Li et al., 2023)) and then use an LLM (Brown et al., 2020) to infer the composed query by integrating the manipulation text. One-stage training-free approaches, such as OSrCIR (Tang et al., 2025b) and STiTch (Li et al., 2026), use an MLLM (Liu et al., 2023) to reason over the reference image and the manipulation text jointly and generate the composed query in one step. More recently, Paracosm (Wang et al., 2026) further imagines the target by generating a pseudo target image and re-encoding it with the CLIP image encoder. Although imagining the target brings visual evidence into the query, doing so in pixel space leads to two issues.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) (b)

Figure 1: **(a)** On all mAP@k metrics of the CIRCO test set with both the ViT-L/14 and the ViT-G/14 backbones, PPT outperforms the training-free methods that do not synthesize images, including the one-stage methods STiTch and OSrCIR and the two-stage methods LDRE and CIReVL. **(b)** CIRCO mAP@5 versus inference latency of one-stage (STiTch, OSrCIR and Paracosm) and two-stage (CIReVL and LDRE) training-free methods. The latencies of PPT and Paracosm are measured by us on a single NVIDIA A100 GPU, and those of the other methods are taken from STiTch (Li et al., 2026), measured on a single NVIDIA A6000 GPU; the accuracy of Paracosm is taken from its paper (Wang et al., 2026). The dots of each method correspond, from bottom to top, to the ViT-B/32, ViT-L/14 and ViT-G/14 backbones, whose latencies are nearly identical. The x-axis is broken between 7 s and 60 s.

(1) **Pseudo target images carry redundant details.** Rendering the imagined target in pixels forces the generator to commit to details that the multimodal query does not specify, such as background, layout and text, and these details are then encoded into the pseudo target embedding. As illustrated in Fig. 2(a) and Fig. 3, such details can mislead retrieval.

(2) **Pixel-space imagination is inefficient.** Generating the pseudo target image and re-encoding it with the CLIP image encoder adds substantial computation at query time (Wang et al., 2026; 2025a). As shown in Fig. 1b, Paracosm takes 60 s per query, more than an order of magnitude longer than PPT, which never leaves the embedding space.

Our Solution. To tackle these challenges, we propose Diffusion Prior-driven Pseudo Target Extraction (PPT), a training-free approach that imagines the target directly in the CLIP image embedding space. PPT uses a diffusion prior to map the textual composed query generated by an MLLM to a pseudo target image embedding, and aggregates it with the text embeddings for retrieval, without any image decoding or re-encoding. As shown in Fig. 1a, PPT surpasses STiTch, the strongest training-free baseline that does not synthesize images, on every mAP@k metric of CIRCO with both ViT-L/14 and ViT-G/14. Fig. 1b shows that PPT has the lowest latency among all compared methods; compared with Paracosm, which attains higher accuracy by synthesizing images for both the queries and the whole gallery, PPT is about 19$\times$ faster. Our contributions are as follows:

- We identify two limitations of imagining the pseudo target in pixel space for ZS-CIR: the redundant image details introduced by rendering, and the computational overhead of image generation and re-encoding.
- We propose PPT, a simple yet effective training-free approach that extracts the pseudo target embedding directly with a diffusion prior and aggregates it with the textual composed query and the manipulation text.

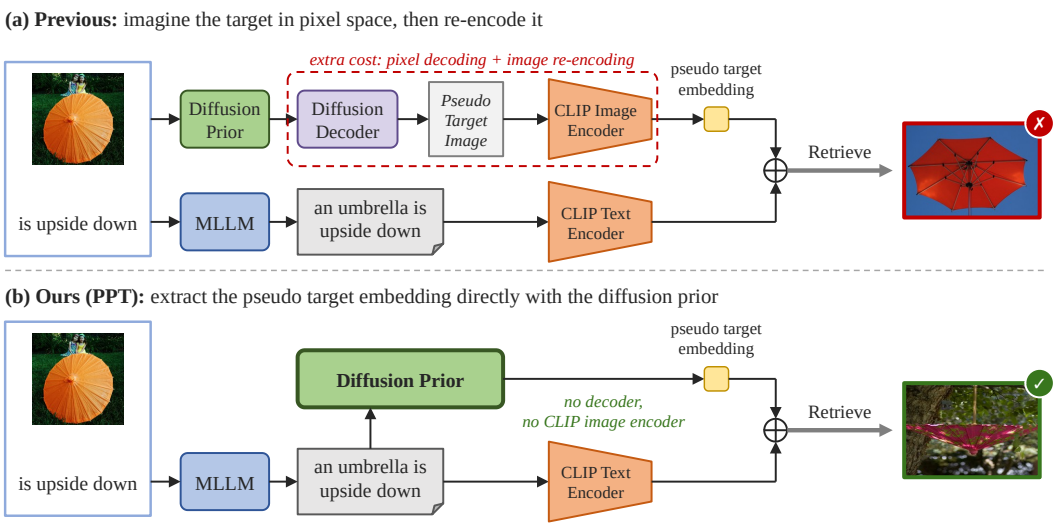


Figure 2: Motivation of PPT. **(a)** Existing pseudo-target methods imagine the target in pixel space: the imagined target is decoded into a pseudo target image, which is then re-encoded by the CLIP image encoder. The decoded image carries details that the multimodal query does not specify, which can mislead retrieval, and the decoder together with the image encoder adds inference cost. **(b)** PPT extracts the pseudo target embedding directly with the diffusion prior, so neither the diffusion decoder nor the CLIP image encoder is needed at query time.

- Extensive experiments on object/scene and attribute manipulation benchmarks demonstrate that PPT achieves a favorable trade-off between retrieval accuracy and efficiency.

2 Background and Motivation

In this section, we provide a brief overview of the foundational concepts in this study in Sec. 2.1, and highlight key observations in Sec. 2.2, which offers valuable insights for our approach. The introduction of related work is presented in the Appendix E due to the page limit.

2.1 Preliminaries

MLLM-based ZS-CIR pipeline. The MLLM-based ZS-CIR pipeline typically adopts an MLLM such as Qwen2.5-VL, which takes the reference image I_r and the manipulation text T_m as input and generates a textual description T_c of the target image by aggregating the visual and textual content:

$$\mathrm{T}_c = \mathcal{F}(\mathrm{I}_r, \mathrm{T}_m), \tag{1}$$

where $\mathcal{F}$ denotes the MLLM.

Zero-Shot CIR (ZS-CIR). Recent studies (Saito et al., 2023; Baldrati et al., 2023) remove the need for triplet annotations by applying CLIP (Radford et al., 2021) to ZS-CIR via text-image retrieval. Current approaches predominantly fall into textual-inversion-based, two-stage training-free and one-stage training-free strategies.

- **Textual-inversion-based.** These methods map the reference image into a pseudo language token by training a mapping network, and then concatenate this token with the manipulation text to form a composed query, pioneered by Pic2Word (Saito et al., 2023) and SEARLE (Baldrati et al., 2023). Pic2Word uses image-text pairs and unlabeled image datasets to train the mapping network. SEARLE also introduces the open-domain benchmark CIRCO. However, textual inversion methods require training on large-scale image-text data.

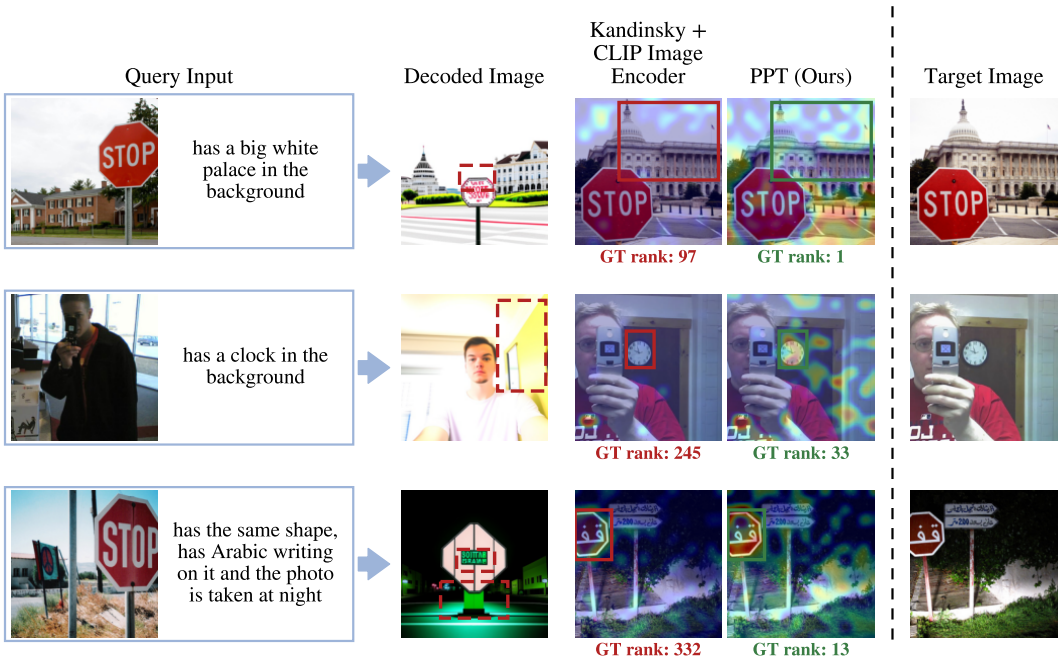


Figure 3: Redundant image details introduced by decoding the pseudo target into pixels (CIRCO validation set, ViT-G/14). Both routes start from the same diffusion-prior sample of the same query: *Kandinsky + CLIP image encoder* (Tab. 4) renders it with the Kandinsky 2.2 decoder and re-encodes the result with the CLIP image encoder, whereas PPT uses the prior embedding directly. Dashed red boxes mark details that the decoder invents or corrupts (garbled sign text, an over-exposed wall, a glowing base), and the clock requested in the second query is missing altogether. The heatmaps are Grad-CAM maps of the cosine similarity between each pseudo target embedding and the ground-truth target, with both maps of a row on one colour scale; solid boxes mark the region the manipulation text refers to. GT rank is the rank of the annotated target among the 123,403 gallery images when the pseudo target embedding alone is used as the query. Over all 220 validation queries, the median rank is 24 for PPT versus 49 for Kandinsky + CLIP image encoder, and PPT ranks the target higher on 124 of them; the three examples are drawn from the 24 queries with the largest improvement.

- **Two-stage training-free.** These methods caption the reference image with an image captioning model and then infer a composed query by integrating the manipulation text with an LLM. CIReVL (Karthik et al., 2024) uses BLIP-2 (Li et al., 2023) to generate an image description and an LLM (Brown et al., 2020) to generate the composed query. LDRE (Yang et al., 2024) ensembles multiple composed queries to enrich the semantics of the query composition. Since the LLM cannot see the reference image, these methods may lose critical visual details during query composition.

- **One-stage training-free.** These methods use an MLLM to generate the composed query directly from the reference image and the manipulation text, preserving the visual details relevant to the manipulation. OSrCIR (Tang et al., 2025b) employs chain-of-thought reasoning (Wei et al., 2022) to infer the implicit user intent. Paracosm (Wang et al., 2026) generates a pseudo target image via Qwen-Image-Edit (Wu et al., 2025) and synthesizes a counterpart for each gallery image to bridge the synthetic-to-real domain gap. While Paracosm achieves strong accuracy, imagining the target in pixel space incurs high computational overhead. A more comprehensive review is provided in Appendix E.

2.2 Key Observations

Despite the promising results of one-stage training-free methods in Sec. 2.1, they continue to exhibit certain limitations.

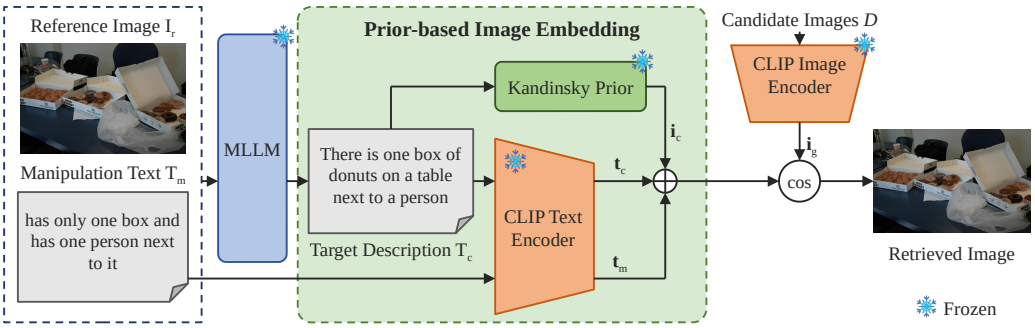


Figure 4: Overview of PPT. Given the reference image I_r and the manipulation text T_m, a frozen MLLM generates the target description, i.e., the textual composed query T_c. The Kandinsky prior maps T_c to the pseudo target image embedding $\mathbf{i}_c$ directly in the CLIP image embedding space, while the CLIP text encoder encodes T_c and T_m into $\mathbf{t}_c$ and $\mathbf{t}_m$. The three embeddings are aggregated into the enhanced target embedding, which ranks the candidate images $\mathcal{D}$ by its cosine similarity with their CLIP image embeddings $\mathbf{i}_g$. All models are frozen.

Redundant image details. Existing methods that imagine the target (Wang et al., 2026; 2025a; Li et al., 2025b) generate an explicit pseudo target image from the multimodal query and use its CLIP embedding for retrieval. However, rendering the target in pixels forces the generator to fill in details that the query does not specify. Fig. 3 compares, for the same diffusion prior sample, the embedding obtained by decoding it into pixels and re-encoding it with the one used directly by PPT: the decoded images contain invented or corrupted details, and over the 220 CIRCO validation queries the median rank of the ground-truth target is 49 for Kandinsky + CLIP image encoder versus 24 for PPT. Under the same pipeline, decoding and re-encoding also lowers mAP@5 on the CIRCO test set from 27.37 to 25.84 (Tab. 4). These observations suggest keeping the pseudo target in the embedding space, where it provides the target visual semantics without committing to redundant image details.

Inefficiency. Generating the pseudo target image and re-encoding it with the CLIP image encoder (Radford et al., 2021) is also costly at query time. As shown in Tab. 6, Paracosm takes 60 s per query on a single A100 GPU, whereas the whole PPT pipeline takes 3.2 s. A pseudo target in the CLIP image embedding space thus captures the target visual semantics with much higher efficiency.

3 Method

In this section, we introduce Diffusion Prior-driven Pseudo Target Extraction (PPT), a training-free framework for ZS-CIR. We first describe how PPT extracts the pseudo target embedding with a diffusion prior in Sec. 3.1, and then explain how this embedding enhances the composed query for retrieval in Sec. 3.2. A detailed illustration of PPT is provided in Fig. 4.

3.1 Diffusion Prior-driven Pseudo Target Extraction module

Existing methods (Wang et al., 2026; 2025a) imagine a pseudo target image in pixel space to provide rich visual information. However, as shown in Fig. 3 and Tab. 6, rendering the imagined target into pixels introduces redundant details and extra cost. We therefore extract the pseudo target embedding directly with a diffusion prior.

Specifically, we begin the pseudo target image embedding generation by generating the textual composed query T_c by applying MLLM $\mathcal{F}$ with the reference image I_r and the manipulation text T_m:

$$\mathrm{T}_c = \mathcal{F}(\mathrm{I}_r, \mathrm{T}_m), \tag{2}$$

Then, for the textual composed query T_c describing the target content, we obtain the pseudo target image embedding $\mathbf{i}_c$ with the diffusion prior ψ of Kandinsky (Razzhigaev et al., 2023), which maps a text to an image embedding in the CLIP space and thus depicts the target image content of the

composed query. The process is formulated as follows:

$$\mathbf{i}_c = \psi(\mathrm{T}_c), \tag{3}$$

On CIRCO, we condition the prior on T_m instead (Sec. 4.1). Since $\mathbf{i}_c$ is directly extracted in the embedding space, PPT requires neither the diffusion decoder nor the CLIP image encoder to obtain the pseudo target. The pseudo target image embedding $\mathbf{i}_c$ is then used to enhance the composed query for retrieval, as described in Sec. 3.2.

3.2 Pseudo Target-Enhanced Retrieval

Existing training-free methods (Karthik et al., 2024; Yang et al., 2024) retrieve the target image with the textual composed query alone, which relies on the target description generated by the language model. However, the textual composed query is limited in depicting the visual content of the target image, and the manipulation semantics may be weakened in the long target description, leading to suboptimal retrieval. To address this issue, we propose Pseudo Target-Enhanced Retrieval to aggregate the pseudo target image embedding with the textual embeddings, acquiring an enhanced target embedding.

Specifically, we begin the retrieval by encoding the textual composed query T_c and the manipulation text T_m with the CLIP (Radford et al., 2021) text encoder $\mathcal{E}_T$, obtaining the composed query embedding $\mathbf{t}_c$ and the manipulation text embedding $\mathbf{t}_m$:

$$\mathbf{t}_c = \mathcal{E}_T(\mathrm{T}_c), \quad \mathbf{t}_m = \mathcal{E}_T(\mathrm{T}_m), \tag{4}$$

where the composed query embedding $\mathbf{t}_c$ depicts the overall semantics of the target image, and the manipulation text embedding $\mathbf{t}_m$ enhances the manipulation semantics. Then, we use embedding aggregation to combine the composed query embedding $\mathbf{t}_c$, the manipulation text embedding $\mathbf{t}_m$ and the pseudo target image embedding $\mathbf{i}_c$, which provides the visual semantics of the target image. The process is formulated as follows:

$$\mathbf{q} = \lambda\,\mathbf{t}_c + (1-\lambda)\,\mathbf{t}_m + \mu\,\mathbf{i}_c, \tag{5}$$

where $\mathbf{q}$ denotes the enhanced target embedding, all three embeddings are L2-normalized, λ balances the two text embeddings, and μ controls the contribution of the pseudo target, i.e., the visual semantics. We set $\lambda = 0.3$ following Paracosm (Wang et al., 2026) and $\mu = 0.1$. After that, we utilize the CLIP image encoder $\mathcal{E}_I$ to obtain the gallery image embedding $\mathbf{i}_g = \mathcal{E}_I(\mathrm{I}_g)$ for each gallery image I_g, and assess the semantic relevance s by exploiting the similarity between the gallery image embedding $\mathbf{i}_g$ and the enhanced target embedding $\mathbf{q}$:

$$\mathrm{s} = \cos(\mathbf{q}, \mathbf{i}_g), \tag{6}$$

where $\cos(\cdot,\cdot)$ denotes the cosine similarity operation. Finally, the gallery images are ranked according to the semantic relevance s to obtain the retrieval results. Since the gallery image embeddings can be pre-computed offline, the CLIP image encoder is not involved in processing the query at inference time.

4 Experiments

We conduct extensive experiments on popular ZS-CIR datasets, including CIRCO (Baldrati et al., 2023), CIRR (Liu et al., 2021) for object/scene manipulation and FashionIQ (Wu et al., 2021) for attribute manipulation. In Sec. 4.1, we present the details of our experiment setup, encompassing the datasets, baselines and implementation details. After that, in Sec. 4.2, we report the main results. We conduct extensive ablation studies in Sec. 4.3.

4.1 Experimental setup

Datasets. We evaluate PPT on CIRCO and CIRR for object/scene manipulation and on FashionIQ for attribute manipulation. CIRCO is built from COCO (Lin et al., 2014) and provides multiple ground-truth targets per query. CIRR is the first CIR dataset of natural images. FashionIQ focuses on attribute manipulation of fashion items. Details are given in Appendix B.1.

Table 1: Comparison with other ZS-CIR methods on CIRCO and CIRR test sets for object/scene manipulation using ViT-L/14 and ViT-G/14 backbones. The best results are shown in bold and the second-best results are underlined. "-" denotes results not reported in the original papers. Paracosm synthesizes a counterpart image for every gallery image offline.

Backbone	Method	Training-free	CIRCO mAP@k				CIRR Recall@k				CIRR $\text{Recall}_{\text{subset}}$@k		
			k=5	k=10	k=25	k=50	k=1	k=5	k=10	k=50	k=1	k=2	k=3
ViT-L/14	Pic2Word (Saito et al., 2023)	✗	8.72	9.51	10.64	11.29	23.90	51.70	65.30	87.80	53.76	74.46	87.08
	SEARLE (Baldrati et al., 2023)	✗	11.68	12.73	14.33	15.12	24.24	52.48	66.29	88.84	53.76	75.01	88.19
	LinCIR (Gu et al., 2024b)	✗	12.59	13.58	15.00	15.85	25.04	53.25	66.68	-	57.11	77.37	88.89
	CIG + SEARLE (Wang et al., 2025a)	✗	12.84	13.64	15.32	16.17	26.72	55.52	68.10	**89.58**	57.95	77.81	89.45
	CIReVL (Karthik et al., 2024)	✓	18.57	19.01	20.89	21.80	24.55	52.31	64.92	86.34	59.54	79.88	89.69
	LDRE (Yang et al., 2024)	✓	23.35	24.03	26.44	27.50	26.53	55.57	67.54	-	60.43	80.31	89.90
	OSrCIR (Tang et al., 2025b)	✓	23.87	25.33	27.84	28.97	29.45	57.68	69.86	-	62.12	81.92	91.10
	STiTch (Li et al., 2026)	✓	25.55	26.27	28.81	29.99	28.87	57.97	69.90	-	**65.22**	**84.10**	**92.37**
	Paracosm (Wang et al., 2026)	✓	**30.24**	**31.51**	**34.29**	**35.42**	**31.95**	**61.56**	**72.96**	-	64.68	82.89	91.47
	PPT(ours)	✓	27.37	28.31	30.62	31.77	28.65	58.19	70.48	89.18	64.12	82.19	91.40
ViT-G/14	Pic2Word (Saito et al., 2023)	✗	5.54	5.59	6.68	7.12	30.41	58.12	69.23	-	68.92	85.45	93.04
	SEARLE (Baldrati et al., 2023)	✗	13.20	13.85	15.32	16.04	34.80	64.07	75.11	-	68.72	84.70	93.23
	LinCIR (Gu et al., 2024b)	✗	19.71	21.01	23.13	24.18	35.25	64.72	76.05	-	63.35	82.22	91.98
	CIG + LinCIR (Wang et al., 2025a)	✗	20.64	21.90	24.04	25.20	36.05	66.31	76.96	-	64.94	83.18	91.93
	CIReVL (Karthik et al., 2024)	✓	26.77	27.59	29.96	31.03	34.65	64.29	75.06	91.66	67.95	84.87	93.21
	LDRE (Yang et al., 2024)	✓	31.12	32.24	34.95	36.03	36.15	66.39	77.25	-	68.82	85.66	93.76
	OSrCIR (Tang et al., 2025b)	✓	30.47	31.14	35.03	36.59	37.26	67.25	77.33	-	69.22	85.28	93.55
	STiTch (Li et al., 2026)	✓	34.40	35.56	38.07	40.02	39.23	69.95	79.56	-	**73.56**	**89.50**	**95.86**
	Paracosm (Wang et al., 2026)	✓	**39.82**	**40.86**	**43.96**	**45.05**	**39.30**	**70.41**	**80.39**	-	70.82	86.92	94.46
	PPT(ours)	✓	38.14	39.57	42.58	43.78	38.55	68.05	78.34	**93.45**	71.33	87.16	94.48

Table 2: Comparison results between our PPT and other ZS-CIR methods on FashionIQ validation set for attribute manipulation across three categories (Shirt, Dress, Toptee). R@k refers to Recall@k. The best results are shown in bold and the second-best results are underlined.

Backbone	Method	Training-free	Shirt		Dress		Toptee		Average	
			R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
ViT-L/14	Pic2Word (Saito et al., 2023)	✗	26.20	43.60	20.00	40.20	27.90	47.40	24.70	43.70
	SEARLE (Baldrati et al., 2023)	✗	26.89	45.58	20.48	43.13	29.32	49.97	25.56	46.23
	LinCIR (Gu et al., 2024b)	✗	29.10	46.81	20.92	42.44	28.81	50.18	26.28	46.49
	CIG + LinCIR (Wang et al., 2025a)	✗	28.90	47.25	21.12	43.88	29.78	50.54	26.60	47.22
	CIReVL (Karthik et al., 2024)	✓	29.49	47.40	24.79	44.76	31.36	53.65	28.55	48.57
	LDRE (Yang et al., 2024)	✓	31.04	51.22	22.93	46.76	31.57	53.64	28.51	50.54
	OSrCIR (Tang et al., 2025b)	✓	33.17	52.03	**29.70**	**51.81**	**36.92**	**59.27**	**33.26**	**54.37**
	PPT (ours)	✓	**34.54**	**53.97**	27.07	46.85	**36.92**	59.00	32.84	53.28
ViT-G/14	Pic2Word (Saito et al., 2023)	✗	33.17	50.39	25.43	47.65	35.24	57.62	31.28	51.89
	SEARLE (Baldrati et al., 2023)	✗	36.46	55.35	28.16	50.32	39.83	61.45	34.82	55.71
	CIReVL (Karthik et al., 2024)	✓	33.71	51.42	27.07	49.53	35.80	56.14	32.19	52.36
	LDRE (Yang et al., 2024)	✓	35.94	**58.58**	26.11	51.12	35.42	56.67	32.49	55.46
	OSrCIR (Tang et al., 2025b)	✓	38.65	54.71	33.02	54.78	41.04	61.83	37.57	57.11
	STiTch (Li et al., 2026)	✓	**39.48**	56.59	**35.04**	**56.74**	**42.86**	**64.95**	**39.12**	**59.43**
	PPT (ours)	✓	36.56	54.86	31.43	53.64	41.31	64.15	36.43	57.65

Compared Methods. We compare PPT with state-of-the-art textual inversion and training-free ZS-CIR methods. The textual inversion methods are: 1) Pic2Word (Saito et al., 2023), which maps the reference image into a pseudo token in the language space; 2) SEARLE (Baldrati et al., 2023), which combines the pseudo token with a GPT-generated (Brown et al., 2020) caption; 3) LinCIR (Gu et al., 2024b), which masks the subjects in captions to improve training efficiency; and their CIG-enhanced variants (Wang et al., 2025a). The training-free methods are: 1) CIReVL (Karthik et al., 2024), a two-stage approach where a captioner describes the reference image and an LLM composes the query; 2) LDRE (Yang et al., 2024), which ensembles multiple composed queries; 3) OSrCIR (Tang et al., 2025b), a one-stage method where an MLLM generates the composed query via chain-of-thought reasoning; 4) STiTch (Li et al., 2026), a recent one-stage method; and 5) Paracosm (Wang et al., 2026), which generates a pseudo target image and synthesizes a counterpart for every gallery image.

Evaluation Metrics. Following STiTch (Li et al., 2026), we use Recall@k as the evaluation metric for CIRR and FashionIQ, and mean average precision (mAP@k) for CIRCO. Additionally, we also report $\text{Recall}_{\text{subset}}$@k for CIRR. Recall@k denotes the percentage of target images included in the top-k list. $\text{Recall}_{\text{subset}}$@k evaluates with the subset setting, where there are only a few selected images within the subset of the query in the database. For FashionIQ, we separately compute the Recalls for three categories (Dress, Shirt, and Toptee) from the FashionIQ validation split, and then calculate their average.

Implementation details. The default MLLM is Qwen2.5-VL-7B-Instruct with the reflective prompt of OSrCIR (Tang et al., 2025b). The diffusion prior is that of Kandinsky 2.1 for ViT-L/14 and that of Kandinsky 2.2 for ViT-G/14 (Razzhigaev et al., 2023), whose outputs lie in the image

Table 3: Ablation study on PPT for ZS-CIR. Baseline denotes the text-only variant without the pseudo target embedding ($\mu = 0$).

Backbone	Setting	CIRCO				CIRR			
		mAP@5	mAP@10	mAP@25	mAP@50	Recall@1	Recall@5	Recall@10	Recall@50
ViT-L/14	Baseline	24.90	25.91	28.47	29.55	27.66	56.63	69.23	88.89
	Baseline + PPT	27.37	28.31	30.62	31.77	28.65	58.19	70.48	89.18
ViT-G/14	Baseline	37.36	38.65	41.77	42.95	37.61	67.33	77.59	93.12
	Baseline + PPT	38.14	39.57	42.58	43.78	38.55	68.05	78.34	93.45

Table 4: Ablation study of the types of the pseudo target embedding $\mathbf{i}_c$.

Backbone	Types	CIRCO				CIRR			
		mAP@5	mAP@10	mAP@25	mAP@50	Recall@1	Recall@5	Recall@10	Recall@50
ViT-L/14	Kandinsky (No decoding)	27.37	28.31	30.62	31.77	28.65	58.19	70.48	89.18
	Kandinsky + CLIP image encoder	25.84	26.83	29.18	30.24	27.83	57.12	69.40	88.57
	SDXL-Turbo + CLIP image encoder	26.83	27.82	30.25	31.35	28.12	57.61	69.95	88.83
	Qwen-Image-Edit + CLIP image encoder	26.41	27.36	29.77	30.88	27.94	57.35	69.72	88.61
ViT-G/14	Kandinsky (No decoding)	38.14	39.57	42.58	43.78	38.55	68.05	78.34	93.45
	Kandinsky + CLIP image encoder	36.90	38.20	41.34	42.56	37.72	67.18	77.51	92.96
	SDXL-Turbo + CLIP image encoder	37.88	39.12	42.25	43.44	38.03	67.52	77.86	93.14
	Qwen-Image-Edit + CLIP image encoder	37.52	38.81	41.90	43.11	38.10	67.61	77.95	93.20

embedding spaces of CLIP ViT-L/14 and OpenCLIP ViT-bigG/14, respectively; it is sampled with $T = 25$ steps and a guidance scale of 4.0. For the ViT (Dosovitskiy et al., 2020) backbones, we use the official CLIP (Radford et al., 2021) weights for ViT-B/32 and ViT-L/14 and the OpenCLIP weights for ViT-G/14 (Ilharco et al., 2021; Cherti et al., 2023; Schuhmann et al., 2022). On CIRCO, following Paracosm (Wang et al., 2026), T_m concatenates the shared concept annotation with the relative caption, and the prior is conditioned on T_m, which performs better on the validation set; on the other datasets it is conditioned on T_c. All experiments are run on a single NVIDIA A100 GPU.

4.2 Main Results

CIRCO. Tab. 1 presents the results on CIRCO. PPT achieves the best mAP@k among all methods except Paracosm with both backbones. On ViT-L/14, PPT improves the textual inversion method CIG (Wang et al., 2025a) + SEARLE (Baldrati et al., 2023) by 15.03 mAP and the training-free method STiTch (Li et al., 2026) by 1.86 mAP, averaged over the four mAP@k metrics. Paracosm, which synthesizes a counterpart image for every gallery image offline, achieves higher mAP at about $19\times$ the latency of PPT (Sec. 4.4).

CIRR. Tab. 1 also reports the results on CIRR. On ViT-L/14, averaged over the six metrics reported by all methods (Recall@1/5/10 and $\text{Recall}_{\text{subset}}$@1/2/3), PPT reaches 65.84, 3.25 higher than CIG + SEARLE and comparable to STiTch (66.41), while Paracosm reaches 67.59. PPT thus remains competitive on both Recall and the challenging $\text{Recall}_{\text{subset}}$ metrics.

FashionIQ. Tab. 2 reports the results on the FashionIQ validation set, which requires capturing fine-grained fashion attributes. PPT achieves competitive performance on the three categories. On ViT-L/14, it achieves an average Recall (over R@10 and R@50) of 43.1, surpassing the 36.9 of the textual inversion method CIG + LinCIR.

4.3 Ablation Studies

In this section, we conduct ablation studies on PPT. Additional ablation studies are provided in Appendix C.2.

Effects of PPT. Tab. 3 compares PPT with the text-only baseline ($\mu = 0$). Adding the pseudo target embedding improves mAP@5 on CIRCO by 2.47 with ViT-L/14 and 0.78 with ViT-G/14, and

Table 5: Ablation study of the weight μ of the pseudo target image embedding $\mathbf{i}_c$ with the ViT-L/14 backbone ($\lambda = 0.3$). $\mu = 0$ reduces to the text-only baseline.

μ	CIRCO				CIRR			
	mAP@5	mAP@10	mAP@25	mAP@50	Recall@1	Recall@5	Recall@10	Recall@50
0	24.90	25.91	28.47	29.55	27.66	56.63	69.23	88.89
0.1	27.37	28.31	30.62	31.77	28.65	58.19	70.48	89.18

Table 6: Efficiency comparison on CIRR with the ViT-L/14 backbone. "LLM Calls" is the number of LLM invocations per query; "GPU Mem." is the peak GPU memory during inference. All inference times are measured by us on a single NVIDIA A100 GPU; the Recall@1 of Paracosm is taken from its paper (Wang et al., 2026).

Method	Type	LLM Calls ↓	Infer. ↓	GPU Mem. ↓	Recall@1 ↑
Paracosm (Wang et al., 2026) (with mental image)	Pixel	2	60 s	71.6 GB	**31.95**
PPT (ours)	Embedding	1	**3.2 s**	19.8 GB	28.65

Recall@1 on CIRR by 0.99 with ViT-L/14, showing that the pseudo target provides visual semantics complementary to the textual queries.

The Impact of Pseudo Target Embedding. Tab. 4 compares four sources of the pseudo target embedding: the Kandinsky prior embedding used directly (PPT), and the images generated by the Kandinsky decoder, SDXL-Turbo (Podell et al., 2023) and Qwen-Image-Edit (Wu et al., 2025) and re-encoded by the CLIP image encoder. Using the prior embedding directly performs best in all settings, indicating that rendering the pseudo target into pixels is unnecessary and even harmful.

Analysis of μ. Tab. 5 examines the weight μ of the pseudo target image embedding: $\mu = 0$ reduces PPT to the text-only baseline, and $\mu = 0.1$, used in all experiments, improves mAP@5 on CIRCO by 2.47 and Recall@1 on CIRR by 0.99.

4.4 Efficiency Analysis

In Tab. 6, we compare the inference cost of PPT with Paracosm (Wang et al., 2026), which imagines the target in pixel space. On a single NVIDIA A100 GPU, PPT takes 3.2 s per query, about $19\times$ faster than Paracosm (60 s), as it neither decodes nor re-encodes a pseudo target image. Paracosm achieves a higher Recall@1 (31.95 vs. 28.65) at the cost of this latency and of synthesizing the whole gallery offline (12.9 hours on CIRCO). Compared with the text-only baseline, the diffusion prior improves Recall@1 by 0.99 with only T prior denoising steps per query.

5 Conclusion

This study shows that a diffusion prior can imagine the pseudo target for ZS-CIR directly in the embedding space, avoiding the redundant details and the cost of rendering it in pixels. We identify these two issues of pixel-space imagination and present PPT, a training-free framework that extracts the pseudo target embedding with a diffusion prior and aggregates it with the textual queries. Experiments show that PPT consistently improves the text-only baseline and achieves a favorable trade-off between accuracy and efficiency: it is about $19\times$ faster than Paracosm, which attains higher accuracy by synthesizing images for both the queries and the gallery. A limitation of PPT is that its performance still depends on the quality of the textual composed query generated by the MLLM. A potential future direction is to generalize PPT to other zero-shot retrieval tasks.

AI USE STATEMENT

In this work, Large Language Models (LLMs) are only used for polishing the paper writing. They are not involved in research ideation, experimental design, data analysis, or the formulation of conclusions. All substantive intellectual contributions are made by the authors.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of our work. PPT is a training-free framework, and all components it relies on, including Qwen2.5-VL, the Kandinsky diffusion prior (Razzhigaev et al., 2023) and CLIP (Radford et al., 2021; Ilharco et al., 2021), are publicly available pretrained models. The formulation of PPT is described in Sec. 3, and the experimental setup, including datasets, compared methods, evaluation metrics and implementation details, is presented in Sec. 4.1. Further details of the datasets and the configurations of PPT are provided in Appendix B.1 and Appendix B.2. All experiments are conducted on the public benchmarks CIRCO, CIRR and FashionIQ. We will release the code upon publication.

REFERENCES

Lorenzo Agnolucci, Alberto Baldrati, Alberto Del Bimbo, and Marco Bertini. isearle: Improving textual inversion for zero-shot composed image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):10801–10817, 2025.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 1877–1901, 2022.

Junfeng An, Yingjian Li, Zheng Zhang, Yongyong Chen, and Guangming Lu. High-order correlation embedding for large-scale multi-modal hashing. In *Web and Big Data*, pp. 175–182, 2023. ISBN 978-3-031-25198-6.

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*, pp. 21466–21474, June 2022.

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*, pp. 15338–15347, October 2023.

Tong Bao, Che Liu, Derong Xu, Zhi Zheng, and Tong Xu. MLLM-I2W: Harnessing multimodal large language model for zero-shot composed image retrieval. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 1839–1849, January 2025.

Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about, 2022. URL `https://arxiv.org/abs/2112.09164`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387, 2025a.

Wei-Xin Chen, Yong-Yong Chen, and Shi-Chao Kan. Scene graph generation with large vision-language model and its applications. In *2025 IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2025b.

Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2998–3008, 2020.

Yinqi Chen, Yangting Zheng, Peiwen Li, Weijian Luo, Wenbin He, and Shuo Kang. Liberating the expressive capacity of deep hashing for image retrieval. *Information Processing & Management*, 62(6):104277, 2025c. ISSN 0306-4573.

Yiyang Chen, Zhedong Zheng, Wei Ji, Leigang Qu, and Tat-Seng Chua. Composed image retrieval with text feedback via multi-grained uncertainty regularization, 2024.

Zining Chen, Zhicheng Zhao, Fei Su, and Shijian Lu. Data-efficient generalization for zero-shot composed image retrieval. *Pattern Recognition*, 176:113187, 2026.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023.

Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, and Yuval Atzmon. This is my unicorn, fluffy: Personalizing frozen vision-language representations. In *European conference on computer vision*, pp. 558–577, 2022.

Jinrong Cui, Zhipeng He, Qiong Huang, Yulu Fu, Yuting Li, and Jie Wen. Structure-aware contrastive hashing for unsupervised cross-modal retrieval. *Neural Networks*, 174:106211, 2024. ISSN 0893-6080.

Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *International Conference on Learning Representations*, 2022.

Qili Deng, Shuai Wu, Jie Wen, and Yong Xu. Multi-level image representation for large-scale image-based instance retrieval. *CAAI Transactions on Intelligence Technology*, 3(1):33–39, 2018.

Xinfeng Dong, Dingwen Zhang, Longfei Han, Huaxiang Zhang, Li Liu, and Junwei Han. Clip-based knowledge projector for image–text matching. *Information Processing & Management*, 63(1): 104357, 2026. ISSN 0306-4573.

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. *arXiv preprint arXiv:2010.11929*, 2020.

Longye Du, Shuaiyu Deng, Ying Li, Jun Li, and Qi Tian. A survey on composed image retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.*, 21(7), 2025.

Yongchao Du, Min Wang, Wengang Zhou, Shuping Hui, and Houqiang Li. Image2sentence based asymmetrical zero-shot composed image retrieval. In *International Conference on Learning Representations*, 2024.

Zhiheng Fu, Zixu Li, Zhiwei Chen, Chunxiao Wang, Xuemeng Song, Yupeng Hu, and Liqiang Nie. PAIR: Complementarity-guided disentanglement for composed image retrieval. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/icassp49660.2025.10888153.

Zhixiao Fu, Xinyuan Chen, Jianfeng Dong, and Shouling Ji. Multi-order adversarial representation learning for composed query image retrieval. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 1685–1689, 2021. doi: 10.1109/icassp39728.2021.9414436.

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.

Peng Gao, Yujian Lee, Zailong Chen, Xubo Liu, Hui Zhang, Yiyang Hu, and Guquan Jing. NCL-CIR: Noise-aware contrastive learning for composed image retrieval. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/icassp49660.2025.10888719.

Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European conference on computer vision*, pp. 241–257. Springer, 2016.

Albert Gordo, Filip Radenovic, and Tamara Berg. Attention-based query expansion learning. In *European Conference on Computer Vision*, pp. 172–188, 2020.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion, 2024a.

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*, pp. 13225–13234, June 2024b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Qinlei Huang, Zhiwei Chen, Zixu Li, Chunxiao Wang, Xuemeng Song, Yupeng Hu, and Liqiang Nie. MEDIAN: Adaptive intermediate-grained aggregation network for composed image retrieval. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/icassp49660.2025.10890642.

T Huang. Image retrieval: Past, present, and future. In *International Symposium on Multimedia Infomation Processing*, 1997.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Kaihang Jiang, Wai Keung Wong, Jianyang Qin, Xiaozhao Fang, Jie Wen, Bingzhi Chen, and Hongbo Gao. Label prediction inherited hashing for cross-modal retrieval: Applying supervised hashing to unsupervised tasks. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6343–6352, 2025.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *International Conference on Learning Representations*, 2024.

Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1771–1779, May 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 802–812, June 2021.

Jiaxing Li, Wai Keung Wong, Lin Jiang, Kaihang Jiang, Xiaozhao Fang, Shengli Xie, and Jie Wen. Collaboratively semantic alignment and metric learning for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2311–2328, 2025a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 38th International Conference on Machine Learning*, 2022.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19730–19742, 23–29 Jul 2023.

Miaoge Li, Dongsheng Wang, Zening Sun, Jinsen Zhang, Wenhan Luo, and Jingcai Guo. Stitch: Semantic transition and transportation in collaboration for training-free zero-shot composed image retrieval, 2026. URL https://arxiv.org/abs/2605.21261.

You Li, Fan Ma, and Yi Yang. Imagine and seek: Improving composed image retrieval with an imagined proxy. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 3984–3993, June 2025b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Yicheng Liu, Jie Wen, Chengliang Liu, Xiaozhao Fang, Zuoyong Li, Yong Xu, and Zheng Zhang. Language-driven cross-modal classifier for zero-shot multi-label image recognition. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

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*, pp. 2125–2134, October 2021.

Yiming Ma, Hongkun Yang, Lionel Z. Wang, Bin Chen, Weizhi Xian, and Jianzhi Teng. Dear: Fine-grained vlm adaptation by decomposing attention head roles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31514–31523, June 2026.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

Bill Psomas, George Retsinas, Nikos Efthymiadis, Panagiotis Filntisis, Yannis Avrithis, Petros Maragos, Ondrej Chum, and Giorgos Tolias. Instance-level composed image retrieval. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Guozhi Qiu, Zhiwei Chen, Zixu Li, Qinlei Huang, Zhiheng Fu, Xuemeng Song, and Yupeng Hu. MELT: Improve composed image retrieval via the modification frequentation-rarity balance network. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 13007–13011, 2026. doi: 10.1109/icassp55912.2026.11463370.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 18–24 Jul 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R.G. Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 251–260, 2010. ISBN 9781605589336.

Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion, 2023.

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*, pp. 19305–19314, June 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *European Conference on Computer Vision*, pp. 139–156, 2024.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6418–6428, July 2019.

Shitong Sun, Fanghua Ye, and Shaogang Gong. Training-free zero-shot composed image retrieval with local concept reranking, 2024a.

Shitong Sun, Qilei Li, Shaogang Gong, Weitong Cai, Philip Torr, and Jindong Gu. Benchcir: Benchmarking robustness in composed image retrieval across modalities. *Pattern Recognition*, 179:113724, 2026.

Zelong Sun, Guoxing Yang, Zhiwu Lu, Hao Jiang, Guojie Zhu, and Zhao Cao. Image retrieval with composed query by multi-scale multi-modal fusion. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 5950–5954, 2024b. doi: 10.1109/icassp48485.2024.10446291.

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*, pp. 26951–26962, June 2024.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Gaopeng Gou, and Qi Wu. Missing target-relevant information prediction with world model for accurate zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24785–24795, June 2025a.

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 (CVPR)*, pp. 14400–14410, June 2025b.

Zijing Tian, Zhonghong Ou, Yifan Zhu, Shuai Lyu, Hanyu Zhang, Jinghua Xiao, and Meina Song. Multi-sea: Multi-stage semantic enhancement and aggregation for image–text retrieval. *Information Processing & Management*, 62(5):104165, 2025. ISSN 0306-4573.

Kenta Uesugi, Naoki Saito, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Triplet synthesis for enhancing composed image retrieval via counterfactual image generation. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/icassp49660.2025.10887762.

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*, pp. 6439–6448, June 2019.

Yongquan Wan, Guobing Zou, and Bofeng Zhang. Composed image retrieval: a survey on recent research and development. *Applied Intelligence*, 55(7), 2025.

Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 154–162, 2017.

Lan Wang, Wei Ao, Vishnu Naresh Boddeti, and Ser-Nam Lim. Generative zero-shot composed image retrieval. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29690–29700, 2025a.

Song Wang, Huan Zhao, Yunbo Wang, Jing Huang, and Keqin Li. Cross-modal image–text search via efficient discrete class alignment hashing. *Information Processing & Management*, 59(3):102886, 2022. ISSN 0306-4573.

Tong Wang, Yunhan Zhao, and Shu Kong. Generating a paracosm for training-free zero-shot composed image retrieval. In *European Conference on Computer Vision (ECCV)*, 2026.

Xin Wang, Haipeng Zhang, Mang Li, Zhaohui Xia, Yueguo Chen, Yu Zhang, and Chunyu Wei. Generative editing in the joint vision-language space for zero-shot composed image retrieval, 2025b.

Yifan Wang, Liyuan Liu, Chun Yuan, Minbo Li, and Jing Liu. Negative-sensitive framework with semantic enhancement for composed image retrieval. *IEEE Transactions on Multimedia*, 26: 7608–7621, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. Target-guided composed image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 915–923, 2023. ISBN 9798400701085.

Haokun Wen, Xuemeng Song, Xiaolin Chen, Yinwei Wei, Liqiang Nie, and Tat-Seng Chua. Simple but effective raw-data level multimodal fusion for composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 229–239, 2024. ISBN 9798400704314.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, et al. Qwen-image technical report, 2025.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11307–11317, June 2021.

Peiyu Xie, Jinxing Li, Guangming Lu, Yong Xu, and David Zhang. Caption assisted multimodal large language model for video moment retrieval. *IEEE Transactions on Image Processing*, 34: 6755–6766, 2025.

Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, and Changsheng Xu. Ldre: Llm-based divergent reasoning and ensemble for zero-shot composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 80–90, 2024. ISBN 9798400704314.

Guo Yu, Di Wang, Chengwei Yan, Feng Yan, Nan Luo, Yifeng Wang, and Quan Wang. I2id: Disentangling identity features via synchronized masking for zero-shot composed person retrieval. *Pattern Recognition*, 179:113654, 2026. ISSN 0031-3203.

Mingyu Zhang, Zixu Li, Zhiwei Chen, Zhiheng Fu, Xiaowei Zhu, Jiajia Nie, Yinwei Wei, and Yupeng Hu. HINT: Composed image retrieval with dual-path compositional contextualized network. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 13002–13006, 2026. doi: 10.1109/icassp55912.2026.11461012.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16795–16804, 2022.

Diffusion Prior-driven Pseudo Target Extraction for Training-Free Zero-Shot Composed Image Retrieval

Supplementary Material

OVERVIEW

This material provides supplementary details to the main paper, including the following sections:

- (A) **Method Details**
  - (A.1) Details of Diffusion Prior-driven Pseudo Target Extraction Module
- (B) **Details of Experimental Settings**
  - (B.1) Datasets and Benchmarks
  - (B.2) Reproduction Details of PPT
- (C) **More Experimental Results**
  - (C.1) Additional Main Results
  - (C.2) Additional Ablation Studies
- (D) **More Visualizations**
  - (D.1) Visualization of Diffusion Prior-driven Pseudo Target Extraction Module
  - (D.2) Qualitative Analysis
  - (D.3) Visualization of Failure Cases
- (E) **Related Work**
  - (E.1) Composed Image Retrieval
  - (E.2) Vision and Language Pretraining Models
  - (E.3) Diffusion Prior

Algorithm 1 Computing Process of PPT

Input: Reference image I_r, manipulation text T_m, reflective CoT prompt p_c, image-search database $\mathcal{D}$.
Parameters: Frozen MLLM $\mathcal{F}$, frozen diffusion prior ψ, frozen CLIP image encoder $\mathcal{E}_I$, text encoder $\mathcal{E}_T$; denoising steps T, aggregation weights λ and μ.
Output: Retrieved target image I_t.

1: Initialize pre-trained and frozen models $\mathcal{F}$, ψ, $\mathcal{E}_I$, $\mathcal{E}_T$.
2: Generate the textual composed query:
$$\mathrm{T}_c = \mathcal{F}(p_c \circ \mathrm{I}_r \circ \mathrm{T}_m)$$
3: Compute the text embeddings:
$$\mathbf{t}_c = \mathcal{E}_T(\mathrm{T}_c), \qquad \mathbf{t}_m = \mathcal{E}_T(\mathrm{T}_m)$$
4: Sample initial noise $\mathbf{z}^{(T)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
5: **for** $t = T, T-1, \ldots, 1$ **do**
6: Predict the clean embedding:
$$\hat{\mathbf{z}}_0 = \psi\Big(\mathbf{z}^{(t)}, t, \mathrm{T}_c\Big)$$
7: Update the sample: $\mathbf{z}^{(t-1)} = \mathrm{Step}(\hat{\mathbf{z}}_0, \mathbf{z}^{(t)}, t)$.
8: **end for**
9: Map back to the CLIP image embedding space:
$$\mathbf{i}_c = \mathrm{InvNorm}(\mathbf{z}^{(0)})$$
10: Aggregate the enhanced target embedding:
$$\mathbf{q} = \lambda\, \mathbf{t}_c + (1-\lambda)\, \mathbf{t}_m + \mu\, \mathbf{i}_c$$
11: **for** each gallery image I_g in $\mathcal{D}$ **do**
12: Compute the gallery image embedding:
$$\mathbf{i}_g = \mathcal{E}_I(\mathrm{I}_g)$$
13: Compute the semantic relevance: $\mathrm{s}_g = \cos(\mathbf{q}, \mathbf{i}_g)$.
14: **end for**
15: Retrieve target image: $\mathrm{I}_t = \arg\max_{\mathrm{I}_g \in \mathcal{D}} \mathrm{s}_g$.
16: **return** I_t

A Method Details

In this section, we provide a more detailed introduction to the technical details of PPT presented in Sec. 3.

A.1 Details of Diffusion Prior-driven Pseudo Target Extraction Module

As described in the main paper, PPT uses the diffusion prior of Kandinsky (Razzhigaev et al., 2023) to generate the pseudo target image embedding directly, instead of imagining a pseudo target image with redundant details. In the following, we detail the computing process of PPT.

The complete computing process of PPT is summarized in Alg. 1. All the networks involved are pre-trained and kept frozen, so that the whole pipeline remains training-free. Given a reference image I_r and a manipulation text T_m, the frozen MLLM $\mathcal{F}$ is first prompted with the reflective prompt p_c of OSrCIR (Tang et al., 2025b) to produce the textual composed query T_c, which describes the content of the intended target image (line 2). The frozen CLIP (Radford et al., 2021) text encoder $\mathcal{E}_T$ then encodes T_c and T_m into the composed query embedding $\mathbf{t}_c$ and the manipulation text embedding $\mathbf{t}_m$ (line 3), where the latter preserves the manipulation semantics that a single composed description tends to dilute.

The pseudo target embedding is obtained by the diffusion prior rather than by generating an image. Starting from Gaussian noise $\mathbf{z}^{(T)}$, the Kandinsky prior (Razzhigaev et al., 2023) ψ is queried for T denoising steps, and at every step it directly regresses the clean embedding $\hat{\mathbf{z}}_0$ conditioned on T_c (T_m on CIRCO) following unCLIP (Ramesh et al., 2022) (lines 4–8). Because the prior is trained on element-wise normalized visual embeddings, an inverse normalization maps the denoised result back to the original CLIP image embedding space, yielding the pseudo target image embedding $\mathbf{i}_c$ (line 9). This is the step at which PPT departs from existing MLLM-based pipelines: since the

Dataset	subset	# query	# database
CIRR (Liu et al., 2021)	–	4,148	2,315
CIRCO (Baldrati et al., 2023)	–	800	123,403
Fashion IQ (Wu et al., 2021)	Shirt	2,038	6,346
	Dress	2,017	3,817
	Toptee	1,961	5,373

Table 7: **Statistics of datasets for ZS-CIR.** We list query counts in the test/validation sets and the numbers of database images for each benchmark. Since PPT is a training-free ZS-CIR method, we only utilize the officially released evaluation splits: test sets for CIRR and CIRCO, and the validation set for Fashion IQ.

Item	Setting
Model	Qwen2.5-VL-7B-Instruct
# Parameters	8.3B (vision encoder 0.7B + LLM 7.6B)
Vision encoder	ViT with window attention, patch size 14
Precision	bfloat16
Decoding	greedy, temperature $= 0$
Max new tokens	512
Image resolution	dynamic, at most $1{,}003{,}520$ pixels
Prompt	reflective CoT prompt of OSrCIR (Tang et al., 2025b)

Table 8: **Configuration of the MLLM.** Settings of the MLLM used for generating the textual composed query T_c, kept identical across all benchmarks. The MLLM is used off the shelf without fine-tuning, and greedy decoding makes T_c deterministic. Each query requires a single MLLM call.

Table 9: Comparison with other ZS-CIR methods on CIRCO and CIRR test sets using the ViT-B/32 backbone. Results with ViT-L/14 and ViT-G/14 are reported in Tab. 1.

Method	Training-free	CIRCO				CIRR						
		mAP@k				Recall@k				$\text{Recall}_{\text{subset}}$@k		
		k=5	k=10	k=25	k=50	k=1	k=5	k=10	k=50	k=1	k=2	k=3
Pic2Word (Saito et al., 2023)	✗	5.92	6.37	7.31	7.89	21.34	48.61	62.07	86.12	50.73	72.35	86.08
SEARLE (Baldrati et al., 2023)	✗	9.35	9.94	11.13	11.84	24.00	53.42	66.82	89.78	54.89	76.60	88.19
LinCIR (Gu et al., 2024b)	✗	8.63	9.21	10.48	11.15	23.51	52.07	65.83	88.64	55.12	76.43	88.02
CIReVL (Karthik et al., 2024)	✓	14.94	15.42	17.00	17.82	23.94	52.51	66.00	86.95	60.17	80.05	90.19
LDRE (Yang et al., 2024)	✓	17.96	18.32	20.21	21.11	25.69	**55.13**	**69.04**	**89.90**	60.53	80.65	90.70
OSrCIR (Tang et al., 2025b)	✓	18.04	19.17	20.94	21.85	25.42	54.54	68.19	-	62.31	80.86	91.13
STiTch (Li et al., 2026)	✓	20.12	21.03	22.95	23.88	25.61	54.72	68.30	-	**63.84**	**82.47**	**91.96**
PPT(ours)	✓	**21.80**	**22.61**	**24.57**	**25.49**	**25.93**	54.98	68.72	89.41	63.15	81.73	91.42

prior of Kandinsky 2.1 (resp. 2.2) is trained in the image embedding space of CLIP ViT-L/14 (resp. OpenCLIP ViT-bigG/14), $\mathbf{i}_c$ already lies in the same space as the gallery image embeddings of the corresponding backbone, so neither a diffusion decoder nor a CLIP image encoder is required to obtain it.

The three embeddings are finally aggregated into the enhanced target embedding $\mathbf{q}$, with λ balancing the two text embeddings and μ controlling the contribution of the pseudo target (line 10). Each gallery image $\mathrm{I}_g \in \mathcal{D}$ is encoded by the frozen CLIP image encoder $\mathcal{E}_I$ and scored by its cosine similarity with $\mathbf{q}$ (lines 11–14), and the image with the highest score is returned as the retrieval result (lines 15–16). Note that the gallery image embeddings depend only on $\mathcal{D}$ and are therefore computed once offline, so the per-query cost of PPT consists of one MLLM call and T forward passes of the diffusion prior.

B Details of Experimental Settings

B.1 Datasets and Benchmarks

Following established settings, we evaluate our model on the CIRCO, CIRR and FashionIQ datasets. We summarize these datasets in Tab. 7.

CIRR. The Composed Image Retrieval on Real-life images (CIRR) (Liu et al., 2021) dataset constructs negative samples by mining visually similar images from NLVR2 (Suhr et al., 2019). To mitigate the false negative issue, CIRR organizes visually similar images into subsets. Since each

Table 10: Comparison results on the FashionIQ validation set using the ViT-B/32 backbone. Results with ViT-L/14 and ViT-G/14 are reported in Tab. 2.

Method	Training-free	Shirt		Dress		Toptee		Average	
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
SEARLE (Baldrati et al., 2023)	✗	24.44	41.61	18.54	39.51	25.70	46.46	22.89	42.53
SEARLE-OTI (Agnolucci et al., 2025)	✗	25.37	41.32	17.85	39.91	24.12	45.79	22.44	42.34
CIReVL (Karthik et al., 2024)	✓	28.36	47.84	25.29	46.36	31.21	53.85	28.29	49.35
LDRE (Yang et al., 2024)	✓	27.38	46.27	19.97	41.84	27.07	48.78	24.81	45.63
OSrCIR (Tang et al., 2025b)	✓	**31.16**	**51.13**	**29.35**	**50.37**	**36.51**	**58.71**	**32.34**	**53.40**
PPT (ours)	✓	32.43	51.86	27.61	48.92	35.87	58.14	31.97	52.97

Table 11: Comparison with Paracosm, which also imagines a pseudo target, on CIRCO and CIRR test sets. The accuracy of Paracosm is taken from its paper (Wang et al., 2026), where Recall@50 on CIRR is not reported. Latency is the average end-to-end time per query, measured by us on a single NVIDIA A100 GPU.

Backbone	Method	CIRCO				CIRR				Latency ↓
		mAP@5	mAP@10	mAP@25	mAP@50	Recall@1	Recall@5	Recall@10	Recall@50	
ViT-L/14	Paracosm (Wang et al., 2026)	**30.24**	**31.51**	**34.29**	**35.42**	**31.95**	**61.56**	**72.96**	–	60 s
	PPT (ours)	27.37	28.31	30.62	31.77	28.65	58.19	70.48	89.18	**3.2 s**
ViT-G/14	Paracosm (Wang et al., 2026)	**39.82**	**40.86**	**43.96**	**45.05**	**39.30**	**70.41**	**80.39**	–	60 s
	PPT (ours)	38.14	39.57	42.58	43.78	38.55	68.05	78.34	93.45	**3.2 s**

subset contains images that are visually close to the target, evaluating retrieval within this constrained subset places greater demands on the model's discriminative capability.

CIRCO. Composed Image Retrieval on Common Objects in context (CIRCO) (Baldrati et al., 2023) is based on open-domain real-world images (Lin et al., 2014). It is the first dataset for CIR with multiple ground-truth targets per query and fine-grained annotations.

FashionIQ. Fashion Interactive Queries (Fashion IQ) (Wu et al., 2021) contains fashion images across three categories (Shirt, Dress, Toptee) from Amazon.com. Following prior work (Saito et al., 2023; Baldrati et al., 2023; Karthik et al., 2024; Yang et al., 2024), we evaluate on the validation set, as the test set is not publicly released.

B.2 Reproduction Details of PPT

Qwen2.5-VL-Instruct. We use Qwen2.5-VL-7B-Instruct, an instruction-tuned multimodal large language model that takes images and text as input and generates text, to produce the textual composed query T_c. Its configuration is listed in Tab. 8. The model is used off the shelf without fine-tuning, and the prompt and the other basic settings follow OSrCIR (Tang et al., 2025b).

Diffusion prior models. We evaluate three generation models: Kandinsky (Razzhigaev et al., 2023), SDXL (Podell et al., 2023) and Qwen-Image-Edit (Wu et al., 2025). Kandinsky follows the unCLIP (Ramesh et al., 2022) design and contains a diffusion prior that maps a CLIP text embedding to a CLIP image embedding, so the output of its prior can be taken as $\mathbf{i}_c$ directly. SDXL-Turbo is a distilled variant of the latent diffusion model SDXL that generates an image in a single step, which we condition on the manipulation text, and Qwen-Image-Edit is an instruction-based editing model built on Qwen-Image; neither provides such a prior, so both have to generate a pixel-space image and re-encode it with a CLIP image encoder to obtain $\mathbf{i}_c$.

For both object/scene and attribute manipulation, the composed query is generated with the reflective prompt of OSrCIR (Tang et al., 2025b). The embedding aggregation is controlled by $\lambda = 0.3$ and $\mu = 0.1$ (Eq. 5).

Table 12: Results with GPT-4o as the MLLM on the CIRCO test set with the ViT-L/14 backbone. "+ PPT" adds the pseudo target embedding $\mathbf{i}_c$ to the composed query of each method.

MLLM	Method	mAP@5	mAP@10	mAP@25	mAP@50
GPT-4o	CIReVL (Karthik et al., 2024)	20.43	21.07	23.12	24.05
	CIReVL + PPT	22.86	23.61	25.74	26.71
	LDRE (Yang et al., 2024)	24.62	25.38	27.83	28.90
	LDRE + PPT	26.35	27.19	29.58	30.67
	OSrCIR (Tang et al., 2025b)	25.71	26.64	29.12	30.24
	OSrCIR + PPT	**28.46**	**29.37**	**31.80**	**32.91**
Qwen2.5-VL-7B	PPT (ours)	27.37	28.31	30.62	31.77

Table 13: Detailed statistics of PPT with the ViT-L/14 backbone over 5 random seeds of the diffusion prior. "Avg. #tokens" is the average length of the textual composed query T_c.

Dataset	Metric	Mean	Std Error	95% CI	Avg. #tokens
CIRCO	mAP@5	27.37	0.09	[27.12, 27.62]	41.6
CIRR	Recall@1	28.65	0.07	[28.46, 28.84]	36.2
FashionIQ	Avg. R@10	32.84	0.12	[32.51, 33.17]	22.8

C More Experimental Results

C.1 Additional Main Results

In this section, we provide additional results to supplement the main paper. Tab. 9 and Tab. 10 present the results with the ViT-B/32 backbone, and Tab. 11, Tab. 12 and Tab. 13 present further comparisons and statistics.

Comparison with Paracosm. We further compare PPT with Paracosm (Wang et al., 2026), a recent one-stage method that, like PPT, imagines a pseudo target, but does so in pixel space: it generates a pseudo target image with Qwen-Image-Edit (Wu et al., 2025) and re-encodes it with the CLIP image encoder. Paracosm additionally synthesizes a counterpart image for every gallery image offline to bridge the synthetic-to-real domain gap. As summarized in Tab. 11, Paracosm achieves higher accuracy than PPT on both benchmarks, e.g., 30.24 vs. 27.37 mAP@5 on CIRCO with ViT-L/14 and 39.82 vs. 38.14 with ViT-G/14, whereas PPT is about $19\times$ faster per query (3.2 s vs. 60 s) and requires no offline processing of the gallery. PPT therefore trades a moderate accuracy gap for a much lower inference cost.

Results on GPT-4o. To examine whether PPT generalizes across MLLMs, we replace the MLLM with GPT-4o and add the pseudo target embedding to the composed queries of several baselines. As summarized in Tab. 12, PPT improves the mAP of all these baselines.

Detailed statistical results. In addition to the primary metrics, Tab. 13 reports further statistics, including standard errors (Std Error), confidence intervals (CI), and average response lengths (Avg. #tokens), to illustrate the distribution and the reliability of the evaluation results.

Radar plot details. Fig. 1a visualizes the CIRCO test results in Tab. 1 for the training-free methods, i.e., CIReVL, LDRE, OSrCIR, STiTch and PPT. The chart has eight axes: the right half shows mAP@5, mAP@10, mAP@25 and mAP@50 with ViT-L/14, and the left half shows the same four metrics with ViT-G/14. Each axis is normalized independently, placing the lowest value among the plotted methods at 46% of the radius and the highest value on the outer ring; the numbers next to the vertices are the unnormalized mAP values of PPT. The textual inversion methods are omitted because their much lower scores would stretch the radial range and compress the training-free methods, and CIG is omitted because Tab. 1 reports a different CIG variant for each backbone. The exact numbers of all methods are listed in Tab. 1.

Table 14: Sensitivity to the number of denoising steps T of the diffusion prior with the ViT-L/14 backbone. "Prior time" is the per-query time of the prior on a single NVIDIA A100 GPU.

T	CIRCO				CIRR				Prior time ↓
	mAP@5	mAP@10	mAP@25	mAP@50	Recall@1	Recall@5	Recall@10	Recall@50	
5	26.48	27.40	29.81	30.93	28.02	57.61	69.95	88.87	24 ms
10	27.05	27.98	30.33	31.46	28.43	57.98	70.29	89.05	47 ms
25	**27.37**	**28.31**	**30.62**	**31.77**	**28.65**	**58.19**	**70.48**	**89.18**	115 ms
50	27.31	28.24	30.57	31.70	28.61	58.14	70.43	89.16	228 ms
100	27.29	28.22	30.55	31.69	28.58	58.12	70.41	89.15	455 ms

C.2 Additional Ablation Studies

In addition to the ablations in Sec. 4.3 of the main paper, we further study the backbones, the diffusion prior, the MLLM and the number of denoising steps.

Backbone up-scaling. Tab. 9 and Tab. 1 report the results of PPT with the ViT-B/32, ViT-L/14 and ViT-G/14 backbones. PPT performs better with stronger backbones, and on CIRCO its advantage over the training-free methods that do not synthesize images holds for all three backbones.

Diffusion prior design. PPT obtains the pseudo target embedding from a diffusion prior instead of from a generated image. To examine this choice, we change only the source of $\mathbf{i}_c$ while keeping the MLLM, the textual composed query T_c, the aggregation, λ and μ unchanged, and compare four variants in Tab. 4: the Kandinsky prior embedding used directly (PPT); the image decoded by Kandinsky from the same prior embedding and re-encoded by the CLIP image encoder; and the images generated by SDXL-Turbo (Podell et al., 2023) and Qwen-Image-Edit (Wu et al., 2025) and re-encoded by the CLIP image encoder. With both ViT-L/14 and ViT-G/14, using the prior embedding directly gives the best results on all metrics, and decoding the same prior embedding into pixels and re-encoding it lowers mAP@5 on CIRCO by 1.53 and 1.24, respectively. This suggests that the gain comes from avoiding the decode-and-re-encode round trip rather than from the choice of the generator.

More ablation study on MLLMs. To demonstrate the generalization ability of our method across different MLLMs, we include ablation studies conducted on GPT-4o. Specifically, in Tab. 12, we replace Qwen2.5-VL-Instruct with GPT-4o for the textual composed query generation. These results show that with a strong model, PPT still achieves superior performance. The improvement for CIRCO dataset can be attributed to the enhanced pseudo target embedding, allowing the model to acquire the visual target image content and enhance its performance.

Sensitivity analysis of steps. Tab. 14 reports the effect of the number of denoising steps $T \in \{5, 10, 25, 50, 100\}$. The performance increases from $T = 5$ to $T = 25$ and then saturates, while the time of the prior grows linearly with T; we therefore use $T = 25$.

D More Visualizations

D.1 Visualization of Diffusion Prior-driven Pseudo Target Extraction Module

Fig. 3 in the main paper compares the pseudo target embedding with and without the decoding process by visualizing the Grad-CAM maps of its similarity with the target image. The embedding extracted directly by the prior tends to focus on the regions referred to by the manipulation text, whereas the embedding of Kandinsky + CLIP image encoder is distracted by details invented or corrupted by the decoder.

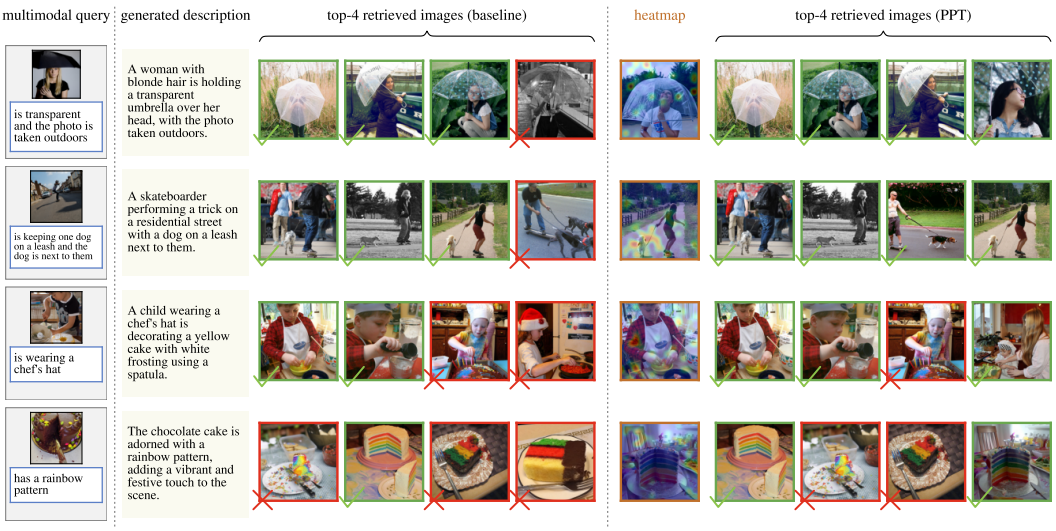


Figure 5: Qualitative comparison on the CIRCO validation set (ViT-G/14). For each multimodal query, the MLLM generates the target description in the second column. The baseline retrieves with the textual composed query alone ($\mathbf{t}_c$ and $\mathbf{t}_m$), while PPT additionally aggregates the pseudo target embedding $\mathbf{i}_c$ extracted by the diffusion prior. Green borders and ticks mark annotated ground-truth targets and red borders and crosses mark the other images; since CIRCO does not annotate every valid target, some images marked in red may still satisfy the query. The heatmap is the Grad-CAM map of the cosine similarity between $\mathbf{i}_c$ and the CLIP embedding of the annotated target image, overlaid on that image. The examples are drawn from the queries on which PPT retrieves more ground-truth images in its top-4 than the baseline.

D.2 Qualitative Analysis

As shown in Fig. 5, we compare PPT with the text-only baseline on object/scene manipulation. By adding the pseudo target embedding, PPT retrieves more ground-truth images in its top-4, and the Grad-CAM maps show that the pseudo target embedding attends to the content required by the query.

Fig. 6 shows examples of attribute manipulation from the Shirt and Dress categories of FashionIQ, where adding the pseudo target embedding brings the target image into the top-4.

D.3 Visualization of Failure Cases

PPT achieves training-free composed image retrieval. While our method demonstrates competitive improvements compared to existing methods, the quality of the textual composed query remains a critical factor. Fig. 7 depicts common failure cases of PPT, particularly with complex manipulation texts. Challenges include handling multiple objects and attributes.



Figure 6: Qualitative comparison on the FashionIQ validation set (Shirt and Dress, ViT-G/14). The layout follows Fig. 5: the baseline retrieves with the textual composed query alone ($\mathbf{t}_c$ and $\mathbf{t}_m$), while PPT additionally aggregates the pseudo target embedding $\mathbf{i}_c$. FashionIQ has a single target per query, marked by the green border and tick. The heatmap is the Grad-CAM map of the cosine similarity between $\mathbf{i}_c$ and the CLIP embedding of the target image, overlaid on that image. The examples are drawn from the queries whose target enters the top-4 with PPT but not with the baseline.

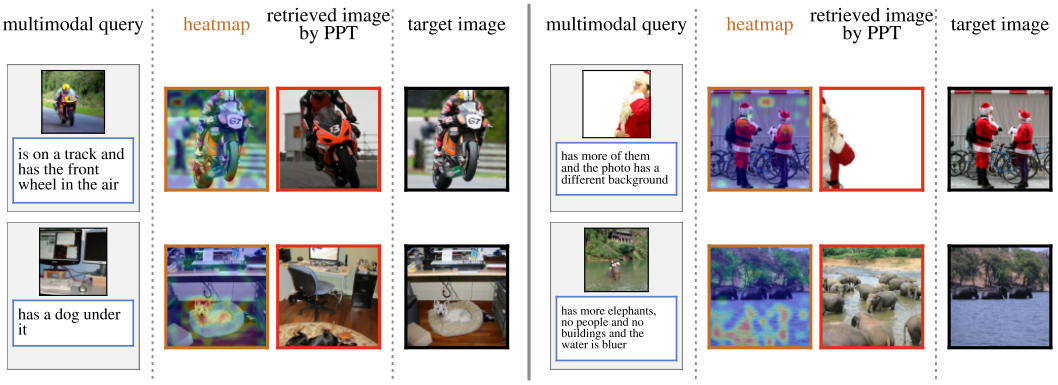


Figure 7: **Failure cases** on the CIRCO validation set (ViT-G/14). Each example shows the multimodal query, the Grad-CAM heatmap of the pseudo target embedding $\mathbf{i}_c$ on the target image, the top-1 image retrieved by PPT and the annotated target image. (1) The retrieved image shows a motorcyclist on a track with both wheels on the ground, missing the fine-grained pose "front wheel in the air". (2) The heatmap concentrates on the dog in the target image, yet the retrieved image shows a dog lying on the floor in front of the desk rather than under it, i.e., the spatial relation is not captured. (3) The retrieved image keeps a single Santa on a white background, ignoring both "more of them" and "a different background". (4) The retrieved image satisfies "more elephants, no people and no buildings" but not "the water is bluer", showing the difficulty of meeting several constraints at once.

E Related Work

E.1 Composed Image Retrieval

CIR (Vo et al., 2019; Cui et al., 2024; Sun et al., 2026) combines a reference image and manipulation text to retrieve the target image. Early supervised methods learn a composed query by fusing visual and textual features (Vo et al., 2019; Lee et al., 2021; Chen et al., 2024; 2020; Kim et al., 2021; Fu et al., 2021; Sun et al., 2024b). For example, TIRG (Vo et al., 2019) fuses CNN image features

and LSTM text features through a gating-and-residual module. Subsequent supervised works (Wen et al., 2023; Baldrati et al., 2022; Fu et al., 2025; Huang et al., 2025; Gao et al., 2025; Zhang et al., 2026; Qiu et al., 2026) strengthen this composition through attention-based explicit-and-implicit matching (Delmas et al., 2022), raw-data-level multimodal fusion (Wen et al., 2024), and negative-aware semantic enhancement (Wang et al., 2024). However, such methods rely on annotated CIR triplets, whose collection is labor-intensive and hard to scale (Uesugi et al., 2025; Psomas et al., 2025). Zero-Shot CIR (Baldrati et al., 2023; Saito et al., 2023; Chen et al., 2026; Yu et al., 2026; Wang et al., 2025a; 2026; Li et al., 2026; 2025b; Sun et al., 2024a; Suo et al., 2024; Du et al., 2024; Wang et al., 2025b) avoids such triplets by leveraging pretrained vision-language models. Textual inversion methods map the reference image into a pseudo-word token, an idea rooted in the personalization of frozen vision–language models (Gal et al., 2022; Cohen et al., 2022) and pioneered by Pic2Word (Saito et al., 2023) and SEARLE (Baldrati et al., 2023) and refined by iSEARLE (Agnolucci et al., 2025). LLM-based methods instead exploit the reasoning ability of (M)LLMs (Xie et al., 2025). CIReVL (Karthik et al., 2024) prompts an LLM with a generated caption, LDRE (Yang et al., 2024) adds query ensembling, and MLLM-I2W (Bao et al., 2025) fine-tunes a multimodal LLM to synthesize composed captions. Although these approaches improve query construction, they still retrieve with a purely textual query, which cannot fully depict the visual content of the target image. Recent one-stage methods such as Paracosm (Wang et al., 2026) instead imagine the target by generating a pseudo target image, which brings visual evidence into the query but introduces redundant image details as well as the cost of pixel decoding and image re-encoding. In contrast, PPT imagines the target directly in the CLIP image embedding space with a diffusion prior, and combines this pseudo target embedding with the textual composed query and the manipulation text.

E.2 Vision and Language Pretraining Models

Vision and Language Pretraining (VLP) models learn transferable cross-modal representations from large-scale image–text pairs. CLIP (Radford et al., 2021; Gal et al., 2022; Cohen et al., 2022) trains dual image and text encoders on web-scale image–text pairs with a contrastive objective, aligning the two modalities in a shared embedding space for zero-shot transfer (Zhou et al., 2022; Ma et al., 2026; Dong et al., 2026). ALIGN (Jia et al., 2021) scales this paradigm to a noisier and larger corpus, showing that a simple contrastive objective remains effective without curated supervision. BLIP (Li et al., 2022) introduces a caption-and-filter bootstrapping scheme to learn from noisy web data and unifies vision–language understanding and generation in a single model. Flamingo (Alayrac et al., 2022) bridges frozen visual encoders and large language models through cross-attention layers, enabling few-shot multimodal reasoning. Such captioning and reasoning abilities (Chen et al., 2025a) are exploited by recent ZS-CIR pipelines (Karthik et al., 2024) to translate visual content into the language space (Chen et al., 2025b). Despite their success, these models are supervised at the level of the global image–text embedding, which captures holistic semantics but discards the region-level correspondence required to localize the changes described by manipulation text (Shao et al., 2024; Liu et al., 2024; Kirillov et al.). In this work, we do not attempt to recover such region-level correspondence; instead, we exploit the shared embedding space of frozen CLIP encoders, in which a diffusion prior can synthesize the pseudo target image embedding directly from text, so that the textual queries and the pseudo target are compared with the gallery images without generating or re-encoding any image.

E.3 Diffusion Prior

A diffusion prior (Ho et al., 2020; Ramesh et al., 2022; Razzhigaev et al., 2023) is a diffusion model that denoises in an embedding space rather than in pixel space, learning to sample the representation of one modality conditioned on another. For example, unCLIP (Ramesh et al., 2022) places such a prior between the two CLIP encoders, generating a CLIP image embedding from a text embedding before a separate decoder renders it into pixels. Because the prior operates on a compact embedding instead of a high-resolution latent, it has since been adopted wherever a target representation must be synthesized rather than observed, including image variation and editing (Ramesh et al., 2022) and pixel-level decoding of frozen self-supervised representations (Bordes et al., 2022). This generative view has been carried into CIR: CompoDiff (Gu et al., 2024a) treats the composed query itself as a generation target, denoising a CLIP embedding conditioned on the reference image and the manipulation text over large-scale synthetic triplets, while PrediCIR (Tang et al., 2025a) predicts the

target visual content that is absent from the reference before mapping it into the word space. Different from CompoDiff, which trains its diffusion model on large-scale synthetic triplets, PPT requires no training: it uses an off-the-shelf text-conditioned prior (Razzhigaev et al., 2023) and conditions it on the textual composed query generated by an MLLM. Different from methods that decode the imagined target into pixels, PPT stops at the prior and uses the sampled embedding as the pseudo target directly, so that only T denoising steps over a compact vector are added at query time.