Generative Zero-Shot Composed Image Retrieval

Lan Wang[1] Wei Ao[1] Vishnu Naresh Boddeti[1] Ser-Nam Lim[2]
[1] Michigan State University [2] University of Central Florida
{wanglan3, aowei, vishnu}@msu.edu sernam@ucf.edu

Abstract

Composed Image Retrieval (CIR) is a vision-language task utilizing queries comprising images and textual descriptions to achieve precise image retrieval. This task seeks to find images that are visually similar to a reference image while incorporating specific changes or features described textually (visual delta). CIR enables a more flexible and user-specific retrieval by bridging visual data with verbal instructions. This paper introduces a novel generative method that augments Composed Image Retrieval by Composed Image Generation (CIG) to provide pseudo-target images. CIG utilizes a textual inversion network to map reference images into semantic word space, which generates pseudo-target images in combination with textual descriptions. These images serve as additional visual information, significantly improving the accuracy and relevance of retrieved images when integrated into existing retrieval frameworks. Experiments conducted across multiple CIR datasets and several baseline methods demonstrate improvements in retrieval performance, which shows the potential of our approach as an effective add-on for existing composed image retrieval. Project Page: https://lan-lw.github.io/CIG/

1. Introduction

Composed Image Retrieval (CIR) takes the composed vision-language queries as input and searches for target images [3, 8, 17, 24, 30, 49]. By combining visual data and verbal instructions, CIR enables a more flexible, user-specific, and precise image retrieval. The composed query includes *a reference image* I_r and *a delta caption* T_r. The target image I_t should visually be like the reference image I_r but capture the content described by the caption T_r [4]. Compared to image-only or text-only retrieval, it is easier for the user to find an image (I_r) that depicts the necessary visual content and then describe the desired difference in natural language [49].

CIR datasets, like CIRR [30] and CIRCO [4], are made up of (I_r, T_r, I_t) triplets. Constructing CIR datasets, therefore, requires huge human labeling efforts, which hinders the widespread adoption and development of CIR algorithms. Existing methods can be categorized into two groups, namely *supervised* CIR and *zero-shot* CIR. Supervised CIR methods [2, 7, 9, 11, 19, 21, 24, 31, 32, 41, 46, 49, 52], are trained on CIR datasets. However, due to the lack of large-scale triplets datasets, supervised methods can only be trained and evaluated on small datasets, which limits their generalization capability. On the contrary, zero-shot methods [1, 4, 15, 26, 29, 38, 43, 44, 51] do not require triplets and train solely based on the reference image and delta caption.

The mainstream solutions like SEARLE [1, 4], Pic2Word [38], Context-I2W [45], LinCIR [15] encode the reference image as a pseudo token and further train a fusion encoder, $\Phi(I_r, T_r)$, that composes the reference image and delta caption in the language embedding space. There is a representation *gap* between the composed embeddings ($\Phi(I_r, T_r)$) and the target image embeddings ($\psi(I_t)$). The former is in the language space, while the latter is in the image space. Intuitively, if we train a fusion encoder ($\Psi(I_r, T_r)$) in the image space, we would improve the image retrieval performance because now both $\Psi(I_r, T_r)$ and $\psi(I_t)$ are in the image space. However, TIRG [49] reports inferior performance when integrating the text information into image embedding space. Some LLM-based methods, such as CIReVL [20], have significantly improved performance. However, their inefficiency remains an issue for retrieval. In this paper, we therefore propose a different approach towards this by asking the question: ***Can a pseudo target image help composed image retrieval?***

To answer the above question, we propose a Composed Image Generation (CIG) approach to assist the CIR task from image space. Our method is straightforward, utilizing the reference image and delta caption to generate pseudo-target images, which are then used to aid image retrieval, as shown in Figure 1. Our method does not require triplet datasets for pre-training, and it can serve as an add-on to any existing CIR method to enhance its performance. First, during the training phase, we only need image-caption pairs. Using pre-trained text inversion networks, we map the image to the token embedding space. The caption is then encoded to produce a composed embedding, which we use as

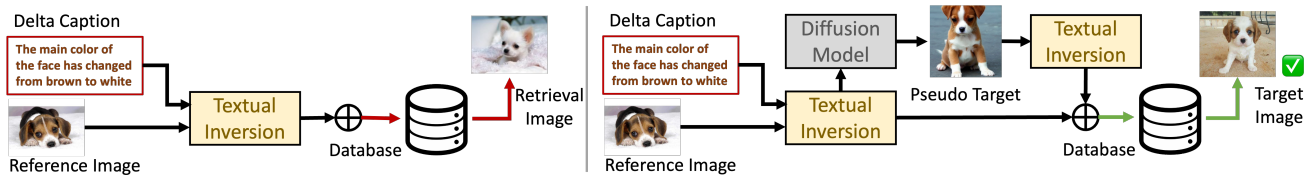

Figure 1. **Zero-Shot Composed Image Retrieval vs. Pseudo Target-Aided Composed Image Retrieval.** Conventional ZS-CIR methods map the image latent embedding into the token embedding space by textual inversion. The proposed Pseudo Target-Aided method provide additional information for composed embeddings from pseudo-target images.

a condition to train latent diffusion models [36] for reconstructing the same image, removing any need for triplets during training. Interestingly, we found that such a model, trained under self-supervision without even seeing a single ground truth target image, is capable of generating pseudo-target image during inference that greatly resembles the ground truth (see Figure 3 and 4 for example). During the retrieval stage, we map both the reference and pseudo target images to the token embedding space and encode them with the delta caption to generate a more informative composed textual embedding for target image retrieval. Our method can improve the performance of various approaches with only a minimal increase in time required.

Our contribution can be summarized as follows: 1) We explore an effective generative method for zero-shot compositional image retrieval, CIG, which can be combined with any CIR methods; 2) Our training process does not require any triplets, utilizing only image-caption pairs in a self-supervised training regime; 3) We conduct multiple experiments on different baselines and obtain significant improvements over different benchmarks; 4) CIG provides a new direction for the CIR task, directly generating a new image by users' instruction while staying faithful to the reference image.

2. Related Work

2.1. Composed Image Retrieval (CIR)

Composed image retrieval (CIR) enables users to input multimodal queries to search for images. CIR can be used for fashion [50] and e-commerce recommendation systems [2, 3, 27]. Compared to traditional image-only and text-only retrieval, users do not need to describe visual content and only need to provide the necessary difference caption. Existing works focus on two important problems: **(1)** how to fuse embeddings of a reference image and a delta caption, and **(2)** how to train CIR models without (reference image, delta caption, target image) triplet datasets. For **(1)**, naive fusion shows inferior performance due to the multi-modal gap between image and text. Recently, SEARLE [4], Pic2Word [38], LinCIR [15] treat the reference image as a pseudo token and plug the pseudo token into the caption sentence (referred to as pseudo caption), leading to SOTA performance. For **(2)**, the motivation stems from the fact that building CIR datasets requires huge human efforts. Zero-shot approaches that do not use CIR datasets during training arise in popularity as a result. Inspired by VLMs, the supervision is the contrastive loss between embeddings of the reference image and the pseudo caption. In this paper, we explore a new direction by generating pseudo images.

2.2. Vision-Language Model (VLM)

Vision-language models (VLMs) align image and text features in the same embedding space [35]. Transformers [47] are usually employed as image and text encoders in VLMs because they can effectively extract features from images [12] and texts [10]. VLMs are pretrained on large-scale datasets of (image, text) pairs and then finetuned for downstream tasks like classification, recognition, and localization. CLIP [35] (Contrastive Language-Image Pretraining) jointly trains image and text encoders to align corresponding pairs. VLMs show increasing representation learning ability under the *scaling law* [5, 13], but incur high computational costs for training. In addition to the above contrastive learning, PaLI-3 [6] encodes images to visual tokens that are concatenated with queries. PaLI-3 achieves SOTA performance on vision-language benchmarks while decreasing model size by $10\times$. To alleviate the training cost of VLMs, BLIP-2 [25] bootstraps off-the-shelf pretrained vision and language models and introduces a lightweight network to bridge the modality gap.

2.3. Diffusion Models

Diffusion models [16, 42] are a class of generative neural networks motivated by non-equilibrium thermodynamics. Diffusion models have been shown to generate high-quality images [33]. The original diffusion models operated in the pixel space, leading to prohibitively high computational costs. To address this problem, the latent diffusion model [36] was introduced that operates in the latent space by using variational autoencoder (VAE) [22]. Furthermore, to guide the image generation by the textual prompt, stable diffusion uses the cross attention [47] mechanism to integrate the text embedding of the prompt. Recent advancements in personalized text-to-image generation have focused on adapting diffusion models to synthesize images

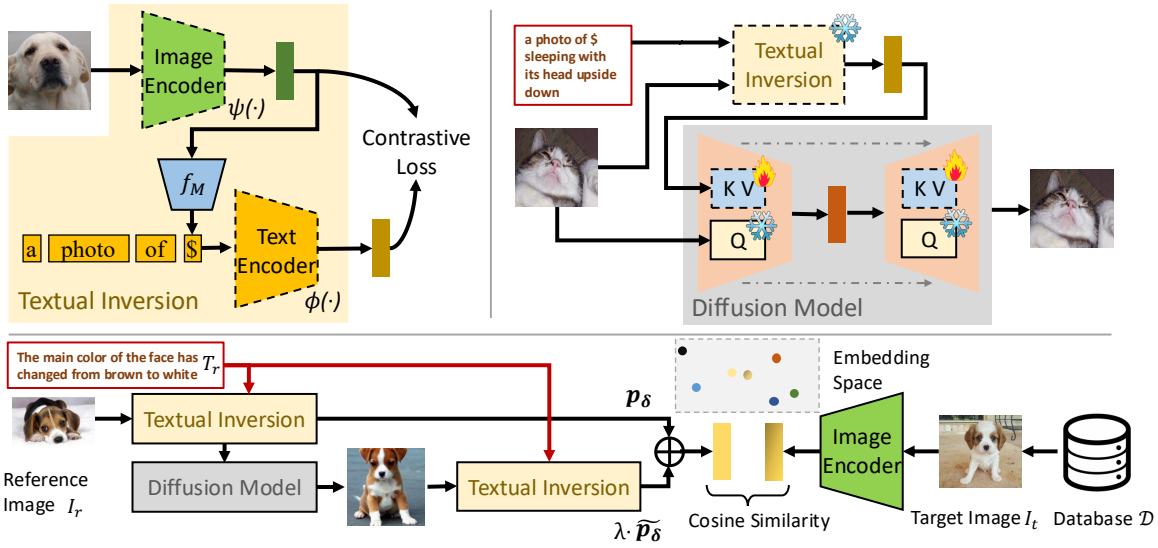

Figure 2. **Overview of the proposed Composed Image Generation (CIG).** During training, CIG model uses composed prompt embeddings as textual conditions, and learn image information from them. In inference stage, the reference image and delta caption form the composed prompt embedding, which CIG model utilizes to generate pseudo target images. These pseudo target images assist in improving ZS-CIR. Top: the training process, including textual inversion network pretraining (left) and CIG model pertaining (right); bottom: inference process for CIR.

of specific subjects or concepts [14, 23, 37]. These approaches improve subject and concept fidelity in diffusion-based generative models, inspiring subsequent work on controllable and personalized image synthesis. In this paper, we leverage a latent diffusion model to generate pseudo target images conditioned on composed textual embeddings. This modification enables the model to effectively synthesize images that integrate information from both the original reference image and the caption.

3. Methodology

In composed image retrieval, the user is querying an image database with a reference image I_r and delta caption T_r pair, as shown in Figure 1. The matched image in the database is called the target image I_t. As an example, consider that the user is querying a photo of *"dog"*. The reference image includes the necessary visual content of a dog, while a delta caption would describe the *difference* between the reference and target images, *e.g.*, *"The main color of the face has changed from brown to white"*.

3.1. Preliminaries

Image and Text Encoders. Following recent approaches like SEARLE [4], Pic2Word [38], LinCIR [15], we apply a pretrained VLM, specifically CLIP [35], by taking advantage of the multimodal representation learning ability of CLIP encoders. The CLIP model includes an image encoder $\psi(I)$ and a text encoder $\phi(T)$. The CLIP model was pretrained on 400 million (image, text) pairs, so the image and text embeddings yielded by its encoders are aligned in a common image-language feature space. In our pipeline, the visual feature of a reference image is extracted by the image encoder, which is then integrated into a delta caption text. Finally, the text encoder extracts the composed feature from the caption (see the upper left panel of Figure 2). Both image and text encoders are frozen during training due to the large number of parameters ($>$ 100 million). Instead, *composed* image-language models learn a fusion encoder $\Phi(I_r, T_r)$ that maps the reference image and delta caption, (I_r, T_r), to a composed feature in an (image-)text embedding space, *i.e.* $\Phi(I_r, T_r) \mapsto x \in \mathbb{R}^d$ where d is the predefined dimension of the embedding space. The features of the reference images are obtained by an image encoder, $\psi(I_t) \mapsto y \in \mathbb{R}^d$.

During training, the cosine similarity between x and y, $\frac{xy}{\|x\|\|y\|}$, is maximized. During inference, the target image is retrieved based on $I_t = \text{argmax}_{I \in \mathcal{D}} \frac{\Phi(I_r,T_r)\psi(I)}{\|\Phi(I_r,T_r)\|\|\psi(I)\|}$. There are two directions to learn the fusion encoder: **(1)** Language $\mapsto$ Image, $\Psi(I_r, T_r)$: inject language information into image embeddings, and **(2)** Image $\mapsto$ Language, $\Phi(I_r, T_r)$: inject image information into language embeddings. **(1)** TIRG [49] edits the image embeddings by integrating the language information using a gated residual connection. **(2)** CBIR [2] directly fuses embeddings of the reference image and the delta caption and fine-tunes the text encoder. SEARLE [4] uses a CLIP image encoder to gener-

ate a pseudo token given a reference image and then plugs in the pseudo token into the delta caption. Then, a CLIP text encoder is applied to map the edited delta caption to an embedding. In addition to frozen image and text encoders, Pic2Word [38] introduces a small trainable mapping network to learn the pseudo token given the image embedding yielded by the image encoder. LinCIR [15] shares similar ideas as SEARLE and pic2Word but only trains the text encoder using self-supervised learning.

Diffusion Models. Latent Diffusion Models (LDM) [36] works in the latent space of powerful pretrained autoencoders, which reduce the computational complexity of high-dimensional data. The model includes three components: the encoding to latent space, the diffusion process within this space, and the final decoding to reconstruct the data. Encoder $\mathcal{E}$ compresses images x to latent $z = \mathcal{E}(x)$, and Decoder $\mathcal{D}$ reconstructs it back to image $\mathcal{D}(z) \approx x$. A time-conditional UNet $\varepsilon_\theta(o, t)$ is trained to remove the noise using the objective:

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\varepsilon_\theta\sim\mathcal{N}(0,1),t}[||\varepsilon - \varepsilon_\theta(z_t, t)||_2^2]. \quad (1)$$

Similar to other generative models, LDM can be implemented with a conditional denoising autoencoder:

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\varepsilon_\theta\sim\mathcal{N}(0,1),t}[||\varepsilon - \varepsilon_\theta(z_t, t, \tau_\theta(y))||_2^2], \quad (2)$$

where y is the conditional input such as text, τ_θ is a domain specific encoder that projects y to an intermediate representation.

3.2. Composed Image Retrieval through Composed Image Generation

Figure 2 shows the framework of the proposed CIG. Given an image-caption dataset with image-caption pairs $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{t}_i)\}_{i=1}^N$, where $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{t} \in \mathcal{T}$ are images and captions, we first map the images to text embedding space and construct composed prompt embeddings. Then, we finetune a latent diffusion model to reconstruct the image using composed prompt as textual condition, after which the pretrained diffusion model is used to generate pseudo-target images in the inference stage. It is noteworthy that the finetuning is in fact self-supervised, seeking only to reconstruct the original image and does not require any triplets, but interestingly induced an ability to generate pseudo target during inference. Finally, those generated images are utilized to boost composed image retrieval.

Textual Inversion Network. The textual inversion network f_M learns a projection from the image latent embedding space to the token embedding space. Instead of learning from scratch, we employed a pre-trained textual inversion network from mainstream zero-shot CIR methods [4, 15, 38] since they are either pre-trained with a large number of image-text pairs with contrastive loss or augmented with carefully designed text context. Those pre-trained textual inversion networks have a good ability to extract information for reference images, which is important for composed image generation. Following methods [4, 15], we use the text embedding space of the CLIP text encoder as the target embedding space. The pretrained CLIP image encoder $\psi(\cdot)$ is first adopted to extract the image feature, then the pre-trained textual inversion networks project the CLIP features to word embeddings: $\boldsymbol{s} = f_M(\psi(\boldsymbol{x}))$.

Composed Image Generation. To integrate the pseudo token embedding with text, we construct the template T "a photo of S^* that $\{$*caption*$\}$" and extract its features using the CLIP text encoder. Then, the composed prompt embedding is obtained: $\boldsymbol{p} = \phi(T)$. To make the caption a better contextual description, we apply a keyword masking strategy to mask a keyword token from each caption. The "keyword" is defined as consecutive adjectives and nouns, which follows the settings of [15].

We finetune Latent Diffusion Models to adopt the composed prompt embedding as a textual condition for reconstruction. Unlike pure text prompts, our composed prompt embedding contains image information, which in effect induces the model to generate target images under the instruction of the delta caption and maintain the information contained in the reference images. A time-conditional UNet is finetuned for CIG objective:

$$L_{CIG} = \mathbb{E}_{\mathcal{E}(\boldsymbol{x}),\varepsilon_\theta\sim\mathcal{N}(0,1),t}[||\varepsilon - \varepsilon_\theta(\boldsymbol{z_t}, t, \boldsymbol{p})||_2^2], \quad (3)$$

where $z = \mathcal{E}(\boldsymbol{x})$ is images latent compressed by encoder $\mathcal{E}$ from LDMs and t is the timestep.

In inference stage, we construct a new template T_δ "a photo of S^* that $\{$*delta caption*$\}$" for CIG task. The LDMs take the composed embedding $\boldsymbol{p_\delta} = \phi(T_\delta)$ of delta caption T_r and reference image I_r as input, generating pseudo target image $\tilde{\boldsymbol{x}}$.

Composed Image Retrieval. To benefit CIR task, we combine pseudo target $\tilde{\boldsymbol{x}}$ with the reference image I_r and delta caption T_r in text embedding space. We conduct a similar way as generating the composed prompt embedding $\boldsymbol{p_\delta}$. We construct a new template $T_{\tilde{\delta}}$ for pseudo target images as "a photo of S^* that $\{$*delta caption*$\}$" using word embeddings $\tilde{\boldsymbol{s}} = f_M(\psi(\tilde{\boldsymbol{x}}))$, generating the composed text embedding of the pseudo image and delta caption, $\tilde{\boldsymbol{p_\delta}}$. Finally, the retrieved target is defined as:

$$I_t = \underset{I}{\operatorname{argmax}} \frac{(\boldsymbol{p_\delta} + \lambda \cdot \tilde{\boldsymbol{p_\delta}})\psi(I)}{\|(\boldsymbol{p_\delta} + \lambda \cdot \tilde{\boldsymbol{p_\delta}})\|\|\psi(I)\|}, \quad (4)$$

where λ is a trade-off hyperparameter to control the weight of the composed text embedding of the pseudo image and delta caption.

4. Experiments

4.1. Experimental Settings

Implementation Details. We employ the Stable Diffusion [36] V1-5 and SDXL [34] as our text-to-image models. We

Table 1. **Quantitative results on CIRR test set.** Results of Pic2Word[38], SEARLE[4], LinCIR [15] and those baselines + CIG (highlight in gray), using different CLIP backbones and diffusion models as shown. CIG significantly improves the performance of different baselines.

Method	Backbone	Recall@K				RecallSubset@K		
		K = 1	K = 5	K = 10	K = 50	K = 1	K = 2	K = 3
Image-only	ViT-B/32	6.89	22.99	33.68	59.23	21.04	41.04	60.31
Text-only		21.81	45.22	57.42	81.01	62.24	81.13	90.70
Image + Text		11.71	35.06	48.94	77.49	32.77	56.89	74.96
SEARLE		24.00	53.42	66.82	89.78	54.89	76.60	88.19
SEARLE + CIG		25.33	54.82	68.05	90.43	**57.86**	78.22	89.25
SEARLE + CIG-XL		24.75	54.36	67.81	90.58	56.24	77.18	89.01
SEARLE + CIG-XL turbo		**25.54**	**55.01**	**68.24**	**90.72**	57.52	**78.36**	**89.35**
Pic2Word	ViT-L/14	23.90	51.70	65.30	87.80	53.76	74.46	87.07
SEARLE		24.24	52.48	66.29	88.84	53.76	75.01	88.19
LinCIR		25.04	53.25	66.68	-	57.11	77.37	88.89
Pic2Word + CIG		24.63	52.75	65.28	86.51	56.96	77.01	88.82
SEARLE + CIG		25.71	54.51	67.23	88.94	56.60	76.77	88.99
SEARLE + CIG-XL		25.08	54.41	67.66	89.57	56.07	77.08	88.63
SEARLE + CIG-XL turbo		**26.72**	**55.52**	**68.10**	**89.59**	57.95	77.81	**89.45**
LinCIR + CIG		25.64	54.77	67.59	89.04	**58.12**	**78.34**	89.37
LinCIR + CIG-XL		25.06	53.69	66.99	89.01	55.78	76.63	88.41
LinCIR + CIG-XL turbo		26.17	54.94	67.64	89.23	58.0	77.86	89.34
LinCIR	ViT-G/14	35.25	64.72	76.05	-	63.35	82.22	91.98
LinCIR + CIG		**36.05**	**66.31**	**76.96**	**93.81**	64.94	83.18	91.93
LinCIR + CIG-XL		34.43	64.51	76.12	93.54	62.24	81.35	91.28
LinCIR + CIG-XL turbo		35.47	66.0	76.89	93.57	**65.13**	**83.25**	**92.24**

adopt the textual inversion networks from [4] for Stable Diffusion V1-5 and [15] for SDXL. In evaluation, we also use SDXL turbo [39], a real-time diffusion model. For fair comparison, in inference stage, we use the baselines' text inversion network and only provide the generated pseudo target images as additional information. To finetune the text-to-image model, we use 595K filtered image-text pairs from CC3M [40] provided by [28] as our training dataset. It contains 595K images-caption pairs that were selected to obtain a more balanced concept coverage distribution. We use their synthetic BLIP captions and re-download the original-sized images, resulting in 490K images-caption pairs. We take the first consecutive adjectives and noun as keywords and remove them from the caption. During training, we only update the key and value projection layers in the cross attention layer. Our model is trained with a learning rate of 1e-6 using a batch size of 16 on 4 $\times$ A6000 GPUs.

Datasets and Baselines. We use the CIRR [30], CIRCO [4], Fashion-IQ [50] and GeneCIS [48] datasets for CIR task. Following the original benchmarks, we use Recall@k as the metric on the CIRR, GeneCIS, and Fashion-IQ and the mean average precision (mAP@k) for the CIRCO dataset. We compare with recent ZS-CIR methods: Pic2Word [38], SEARLE [4] and LinCIR [15], with different backbones including ViT-B, ViT-L and ViT-G CLIP. In addition, we also compared with LLM based CIR method CIReVL [20]. The variants "image-only", "text-only" and "image+text" denote performing retrieval with CLIP using only the reference image, delta caption, as well as averaging their embeddings respectively.

4.2. ZS-CIR Benchmark Comparisons

We evaluate three different versions of CIG models: CIG, CIG-XL, CIG-XL turbo, which uses Stable Diffusion V1-5, SDXL, SDXL turbo as base models separately.

CIRR. Table 1 reports the result for CIRR test set. As previously observed, in CIRR, the reference images are almost not related to the target images [4, 20, 38]. Therefore, our method has a significant advantage on this dataset, as generating pseudo target images helps bridge the gap with the actual target images. Experimental results also confirm this. There is an improvement in $Recall$ and $Recall_{Subset}$ across different baselines. The proposed CIG got the most significant improvement for SEARLE, especially with ViT/L14 backbones, CIG-XL turbo increases the top 1 recall from 24.24 to 26.72. Overall, on the CIRR dataset, our method shows a stronger improvement over SEARLE compared to other baselines in terms of recall.

Fashion-IQ. Table 2 provides the result for Fashion-IQ validation set. Compared to other benchmarks, the improvement on the Fashion IQ dataset is slightly less significant. This may be due to the more challenging nature of clothing related generation, making the benefits from generated images limited. Previous work has shown that enhancing

Table 2. **Quantitative results on Fashion-IQ validation set.** Results of Pic2Word[38], SEARLE[4], LinCIR [15] and those baselines + CIG, using different CLIP backbones and diffusion models as shown. CIG improves the performance over different dress types.

Method	Backbone	Shirt		Dress		Toptee		Average	
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Image-only	ViT-B/32	6.92	14.23	4.46	12.19	6.32	13.77	5.90	13.37
Text-only		19.87	34.99	15.42	35.05	20.81	40.49	18.70	36.84
Image + Text		13.44	26.25	13.83	30.88	17.08	31.67	14.78	29.60
SEARLE		24.44	41.61	18.54	39.51	**25.70**	46.46	22.89	42.53
SEARLE + CIG-XL		24.58	41.41	17.74	39.86	25.65	46.35	22.99	42.54
SEARLE + CIG-XL turbo		**24.73**	**41.46**	**18.94**	**39.66**	25.50	**46.66**	**23.06**	**42.59**
Pic2Word	ViT-L/14	26.20	43.60	20.00	40.20	27.90	47.40	24.70	43.70
SEARLE		26.89	45.58	20.48	43.13	29.32	49.97	25.56	46.23
LinCIR		**29.10**	46.81	20.92	42.44	28.81	50.18	26.28	46.49
Pic2Word + CIG		25.71	43.62	20.67	42.64	27.44	48.19	24.60	44.82
Pic2Word + CIG-XL		25.86	43.38	20.38	42.19	28.45	48.29	24.90	44.62
Pic2Word + CIG-XL turbo		26.06	42.79	20.87	42.44	28.56	49.31	25.16	44.85
SEARLE + CIG-XL		26.99	45.44	21.12	43.38	29.27	50.38	25.79	46.40
SEARLE + CIG-XL turbo		26.99	45.44	20.62	43.58	29.37	**50.48**	25.66	46.50
LinCIR + CIG		28.70	47.06	21.12	43.88	29.73	50.23	26.52	47.05
LinCIR + CIG-XL		28.66	47.20	**21.27**	**43.98**	**29.83**	50.28	26.59	47.15
LinCIR + CIG-XL turbo		28.90	**47.25**	21.12	43.88	29.78	50.54	**26.60**	**47.22**
LinCIR	ViT-G/14	46.76	65.11	38.08	60.88	50.48	71.09	45.11	65.69
LinCIR + CIG		47.15	66.63	39.61	**61.28**	**50.69**	**71.65**	45.82	**66.52**
LinCIR + CIG-XL		**47.35**	**66.68**	**39.71**	60.93	50.69	71.39	**45.92**	66.34
LinCIR + CIG-XL turbo		47.30	66.44	39.56	61.08	50.54	71.55	45.80	66.35

Table 3. **Quantitative results on CIRCO test set.** Results of SEARLE[4] and LinCIR [15] and those baselines + CIG (highlight rows in gray), using different CLIP backbones and diffusion models as shown. CIG improves the performance .

Method	Backbone	mAP@k			
		k=5	k=10	k=25	k=50
Image-only	ViT-B/32	1.34	1.60	2.12	2.41
Text-only		2.56	2.67	2.98	3.18
Image + Text		2.65	3.25	4.14	4.54
SEARLE		9.35	9.94	11.13	11.84
SEARLE + CIG		10.19	10.6	11.83	12.47
SEARLE +CIG-XL		10.3	10.79	12.12	12.76
SEARLE +CIG-XL turbo		**10.45**	**11.02**	**12.34**	**13.0**
Pic2Word	ViT-L/14	8.72	9.51	10.64	11.29
SEARLE		11.68	12.73	14.33	15.12
LinCIR		12.59	13.58	15.00	15.86
SEARLE + CIG		12.13	13.02	14.63	15.41
SEARLE + CIG-XL		12.95	13.62	15.28	16.05
SEARLE + CIG-XL turbo		12.84	13.64	**15.32**	**16.17**
LinCIR + CIG -XL		**12.97**	13.64	15.14	16.01
LinCIR + CIG -XL turbo		12.84	**13.77**	15.25	16.12
LinCIR	ViT-G/14	19.71	21.01	23.13	24.18
LinCIR + CIG -XL		**20.64**	**21.90**	**24.04**	**25.20**
LinCIR + CIG -XL turbo		20.62	21.82	24.0	25.12

text context information significantly improves results on this dataset [4, 15]. Nonetheless, we still achieved further improvements across all baselines. Therefore, our method can be effectively combined with strong baselines.

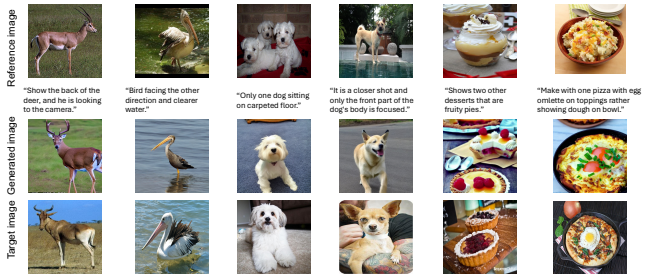

Figure 3. **Qualitative Evaluation on CIRR validation datasets.** CIG effectively make changes according to the caption while preserving the reference image features.

Figure 4. **Qualitative Evaluation on Fashion-IQ validation datasets.** CIG effectively modifies clothing while retaining original clothing features.

Table 4. **Quantitative results on GeneCIS test set.** * denotes our reproduced LinCIR model using their official repo. Our results are based on reproduced LinCIR model.

Method	Backbone	Focus Attribute			Change Attribute			Focus Object			Change Object			Avg
		R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1
Pic2Word	ViT-L/14	15.65	28.16	38.65	13.87	24.67	33.05	8.42	18.01	25.77	6.68	15.05	24.03	11.16
SEARLE	ViT-L/14	17.00	29.65	40.70	16.38	25.28	34.14	7.76	16.68	25.31	7.91	16.84	25.05	12.26
LinCIR	ViT-L/14	16.90	29.95	41.45	16.19	27.98	36.84	8.27	17.40	26.22	7.40	15.71	25.00	12.19
LinCIR*	ViT-L/14	16.60	29.65	40.35	16.19	27.98	36.84	8.21	17.91	25.56	7.96	15.61	25.05	12.24
LinCIR + CIG	ViT-L/14	16.95	29.40	40.43	16.05	28.55	37.17	8.42	17.45	26.73	8.57	15.36	24.85	**12.50**
LinCIR + CIG-XL	ViT-L/14	16.70	29.65	40.85	15.81	28.88	37.41	8.32	17.35	25.46	8.01	15.46	24.23	12.21
LinCIR + CIG-XL turbo	ViT-L/14	16.80	29.70	40.90	15.91	28.88	37.45	8.37	17.35	25.10	7.86	15.46	24.29	12.24
LinCIR	ViT-G/14	19.05	33.00	42.30	17.57	30.16	38.07	10.10	19.08	28.06	7.91	16.33	25.71	13.66
LinCIR*	ViT-G/14	18.95	32.95	42.70	17.90	30.11	38.97	9.80	18.88	27.76	7.70	16.33	25.97	13.59
LinCIR + CIG	ViT-G/14	19.05	32.85	42.40	17.8	30.26	39.16	10.61	19.23	27.40	7.91	16.94	25.36	**13.84**
LinCIR + CIG-XL	ViT-G/14	19.10	32.55	42.50	17.42	30.35	39.02	9.95	18.42	27.60	7.91	16.28	25.41	13.60
LinCIR + CIG-XL turbo	ViT-G/14	19.15	32.90	42.50	17.57	30.45	38.87	9.95	18.42	27.45	7.24	16.33	25.06	13.48

Table 5. **Quantitative comparison with LLM-based methods on CIRR and CIRCO test sets.** The backboone is ViT-L/14 using OpenCLIP weights [18]. The reported CIReVL results are reproduced from their official implementation.

Method	CIRR							CIRCO			
	Recall@K				RecallSubset@K			mAP@k			
	@1	@5	@10	@50	@1	@5	@10	k=5	k=10	k=25	k=50
CIReVL	27.18	56.94	**69.74**	89.54	60.75	**81.08**	91.11	22.40	23.05	25.14	26.23
CIReVL + CIG	**27.35**	**57.13**	69.57	**89.69**	**60.89**	80.92	**91.13**	**22.46**	**23.10**	**25.18**	**26.26**

CIRCO. Table 3 provides the results for the CIRCO test dataset. This dataset addresses the presence of false negatives in previous datasets. It is built based on real-world images and includes multiple ground truth for each reference image. Our method again shows improvements across different baselines on CIRCO. Notably, for SEARLE, our XL and XL Turbo models exhibit significant enhancements.

GeneCIS. The tasks in GeneCIS are different from previous datasets. It's a benchmark for conditional image similarity, applicable to four different retrieval tasks: focusing on an attribute or object, which involves finding images with a similar attribute or object, and changing an attribute or object, which involves altering a specific attribute or object while preserving other characteristics. This dataset is more challenging because the captions include just 1 - 2 words, which is not sufficient for generation. Table 4 shows the results on GeneCIS dataset. Our method produced improvements across all four different tasks. In object related task, our method provides more improvements. This is may be because object changes are easier for diffusion model understand than attributes.

4.3. Ablation Study and Performance Analysis

In this section, several experiments are conducted to provide a comprehensive understanding of CIG, including ablation studies, qualitative and quantitative evaluation on generated examples, and efficiency evaluations. More discussion can be found in Appendix.

Qualitative Evaluation on Composed Image Generation. We visualize the generated images across different benchmarks, using the Stable Diffusion V1-5 model. Figure 3 shows the results on the CIRR dataset, which mainly includes natural and daily scenes. From the results in the first three columns, our method effectively changes animals according to the caption while preserving the original image features. The example in the fourth column demonstrates a scenario where the reference image and target image have less in common in this dataset. Despite this, we successfully generated the target image, which features two fruity pies. In the generated images, we even retained the reference image's characteristic of having thick cream.

Figure 4 showcases the generated images on the FashionIQ dataset. On a clothing dataset that includes dresses, tops, and shirts, our method produces excellent results. It effectively modifies the images according to the captions, including changing patterns, sleeves, colors, and styles while retaining the original clothing features. The generated images closely resemble the target images. Additionally, in the last example, our generated image surpasses the target image by preserving the collar of the reference image, whereas the target image does not even retain the original clothing style. The visualization of CIRCO and GeneCIS can be found in Appendix.

Quantitative Evaluation Composed Image Generation. We further evaluated the quantitative performance of the generated images. Specifically, we calculated the cosine similarity between the generated images and the composed text embeddings, denoted as CLIP scores. We used text em-

Table 6. **Discussion on Generation Quantitative Evaluation and Efficiency Evaluation.**

(a) Composed Image Generation Quantitative Evaluation on CIRR validation set.

Method	target images	SEARLE + SD	SEARLE+ CIG
Clip Score	0.2555	0.2762	0.2810

(b) Additional inference time comparison.

Backbone	ViT-L/14		
Method	SD V1-5	SDXL	SDXL-turbo
Time(s)	2.45	2.56	0.14

(c) Inference time comparison on CIRR.

Backbone	ViT-L/14		
Method	SEARLE	CIReVL	CIG
Time(s)	0.026	1.87	0.17

Table 7. **Ablation Study on different fusion strategies.** Evaluations on CIRR and CIRCO datasets show that fusing on composed textual embedding is more beneficial.

(a) CIRR validation set (Recall@K and RecallSubset@K).

Method	**Recall@K**			**RecallSubset@K**		
	K=1	K=5	K=10	K=1	K=2	K=3
Text embedding	**26.17**	**54.75**	**68.14**	58.77	78.52	**89.45**
Pseudo token	25.64	54.63	67.57	**58.79**	**79.17**	89.33
Image level	24.32	54.48	67.09	53.70	75.17	86.61

(b) CIRCO validation set (mAP@K).

Method	**mAP@k**			
	K=5	K=10	K=25	K=50
Text embedding	**10.98**	**12.12**	**13.65**	**14.41**
Pseudo token	10.57	11.68	13.12	13.79
Image level	10.43	11.55	13.14	13.83

beddings from SEARLE [4] and ViT-L/14 as backbones. We compare the CLIP score with both target images and the images generated by original Stable Diffusion V1-5. Table 6a presents the results on the CIRR validation dataset, showing that our generated images are closer to the composed text embeddings than the target images and image generated by original Stable Diffusion model.

Efficiency Evaluation. Time efficiency is a crucial factor in retrieval tasks. With the use of diffusion models, there are natural concerns about the efficiency of our method. Therefore, we conducted an efficiency evaluation to show the additional inference time required by our method, as shown in the Table 6b. To evaluate the additional inference time, we use a single A6000 GPU with a batchsize of 16 and use SEARLE as the baseline. In the most demanding cases, such as with Stable Diffusion v1-5 and SDXL, the extra time per image is around 2 seconds, which is acceptable for current composed image retrieval tasks. Additionally, we propose an alternative solution using SDXL Turbo as our model. In this scenario, the additional inference time is only around 100 ms and does not impact the retrieval performance. We believe that once the training code for SDXL Turbo is released, both the generation quality and retrieval performance will further improve.

We additionally compared the average runtime per image across different methods. As shown in Table 6 (c), although CIG is slower than traditional methods, it is significantly faster than CIReVL, a popular LLM based method, balancing speed and performance. Table 5 shows that combining our method with CIReVL methods can further enhance performance, demonstrating that our approach is also effective when applied to LLM-based methods.

Ablation Study on different fusion strategies. We demonstrated the performance of integrating generated images in different settings, as shown in Table 7. We used SEARLE as the baseline and Stable Diffusion v1-5 as our base model. "Text embedding" represents our default setting, where the composed text embedding of caption and reference image is combined with the composed text embedding of caption and generated images. "Pseudo token" indicates that the generated image is first mapped to a pseudo token, which is then fused with the reference image's pseudo token. "Image level" means that the CLIP features are extracted from the generated image and directly fused with the composed text embedding, it primarily integrates image features, emphasizing spatial information. However, the CIR task relies more on global semantics. Therefore, fusion at the textual embedding level is more reasonable, as it better enhances semantic information. The results show that the first two fusion methods yield better performance on both CIR and CIRCO datasets.

5. Conclusion

In this paper, we explore a new direction to integrate image features and text features in composed image retrieval. Existing zero-shot CIR algorithms treat the reference image as a single *pseudo token* which is then plugged into the caption, yielding a pseudo caption. It is effective to describe the reference image as a single token in some cases, *e.g.* a single objective. However, it cannot maintain most visual content in the reference image, leading to a *multimodal gap*. To bridge the multimodal gap, we propose to generate a pseudo target image to preserve the visual content by integrating the reference image and the delta caption. To our knowledge, this paper is the first to answer the question both visually and experimentally: *Can the pseudo target image help composed image retrieval?* Our algorithm is simple and complementary to existing methods. We apply the latent diffusion model to generate pseudo target images by integrating the reference image and the delta caption. The training process does not require CIR datasets. Adequate visualization examples over different datasets show that the pseudo images are visually similar to the target images, which increases the cosine similarity in the embedding space, leading to better retrieval performance.

Acknowledgements. This work was supported by the Office of Naval Research (award #N00014-23-1-2417). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ONR.

References

[1] Lorenzo Agnolucci, Alberto Baldrati, Marco Bertini, and Alberto Del Bimbo. isearle: Improving textual inversion for zero-shot composed image retrieval. *arXiv preprint arXiv:2405.02951*, 2024. 1

[2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3

[3] 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*, 2022. 1, 2

[4] 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*, 2023. 1, 2, 3, 4, 5, 6, 8

[5] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 2

[6] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023. 2

[7] Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *Proceedings of the European Conference on Computer Vision*. Springer, 2020. 1

[8] 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*, 2020. 1

[9] Ginger Delmas, Rafael S. Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *International Conference on Learning Representations*, 2022. 1

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[11] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020. 1

[12] 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. 2

[13] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 2

[14] 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. *arXiv preprint arXiv:2208.01618*, 2022. 3

[15] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval. *arXiv preprint arXiv:2312.01998*, 2023. 1, 2, 3, 4, 5, 6

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[17] Mehrdad Hosseinzadeh and Yang Wang. Composed query image retrieval using locally bounded features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3596–3605, 2020. 1

[18] 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, 2021. If you use this software, please cite it as below. 7

[19] Surgan Jandial, Ayush Chopra, Pinkesh Badjatiya, Pranit Chawla, Mausoom Sarkar, and Balaji Krishnamurthy. Trace: Transform aggregate and compose visiolinguistic representations for image search with text feedback. *arXiv preprint arXiv:2009.01485*, 7, 2020. 1

[20] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*, 2023. 1, 5

[21] 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*, 2021. 1

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[23] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2023. 3

[24] 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*, pages 802–812, 2021. 1

[25] 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 *In-*

ternational conference on machine learning, pages 19730–19742. PMLR, 2023. 2

[26] Haoqiang Lin, Haokun Wen, Xuemeng Song, Meng Liu, Yupeng Hu, and Liqiang Nie. Fine-grained textual inversion network for zero-shot composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 240–250, 2024. 1

[27] Chang Liu, Peng Hou, Anxiang Zeng, and Han Yu. Transformer-empowered multi-modal item embedding for enhanced image search in e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 2

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 5

[29] Yikun Liu, Jiangchao Yao, Ya Zhang, Yanfeng Wang, and Weidi Xie. Zero-shot composed text-image retrieval. *arXiv preprint arXiv:2306.07272*, 2023. 1

[30] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021. 1, 5

[31] 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*, pages 2125–2134, 2021. 1

[32] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5753–5762, 2024. 1

[33] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2

[34] 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. *arXiv preprint arXiv:2307.01952*, 2023. 4

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 4

[37] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2023. 3

[38] 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*, 2023. 1, 2, 3, 4, 5, 6

[39] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023. 5

[40] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 5

[41] Minchul Shin, Yoonjae Cho, Byungsoo Ko, and Geonmo Gu. Rtic: Residual learning for text and image composition using graph convolutional network. *arXiv preprint arXiv:2104.03015*, 2021. 1

[42] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[43] Shitong Sun, Fanghua Ye, and Shaogang Gong. Training-free zero-shot composed image retrieval with local concept reranking. *arXiv preprint arXiv:2312.08924*, 2023. 1

[44] Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26951–26962, 2024. 1

[45] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5180–5188, 2024. 1

[46] Yuxin Tian, Shawn Newsam, and Kofi Boakye. Image search with text feedback by additive attention compositional learning. *arXiv preprint arXiv:2203.03809*, 2022. 1

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[48] Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6862–6872, 2023. 5

[49] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 6439–6448, 2019. 1, 3

[50] 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*, pages 11307–11317, 2021. 2, 5

[51] 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 *ACM SIGIR*, 2024. 1

[52] Youngjae Yu, Seunghwan Lee, Yuncheol Choi, and Gunhee Kim. Curlingnet: Compositional learning between images and text for fashion iq data. *arXiv preprint arXiv:2003.12299*, 2020. 1