ConText-CIR: Learning from Concepts in Text for Composed Image Retrieval

Eric Xing[1] Pranavi Kolouju[2] Robert Pless[3] Abby Stylianou[2] Nathan Jacobs[1]
[1]Washington University in St. Louis
[2]Saint Louis University
[3]The George Washington University
{e.xing,jacobsn}@wustl.edu, {pranavi.kolouju,abby.stylianou}@slu.edu, pless@gwu.edu

Abstract

Composed image retrieval (CIR) is the task of retrieving a target image specified by a query image and a relative text that describes a semantic modification to the query image. Existing methods in CIR struggle to accurately represent the image and the text modification, resulting in subpar performance. To address this limitation, we introduce a CIR framework, ConText-CIR, trained with a Text Concept-Consistency loss that encourages the representations of noun phrases in the text modification to better attend to the relevant parts of the query image. To support training with this loss function, we also propose a synthetic data generation pipeline that creates training data from existing CIR datasets or unlabeled images. We show that these components together enable stronger performance on CIR tasks, setting a new state-of-the-art in composed image retrieval in both the supervised and zero-shot settings on multiple benchmark datasets, including CIRR and CIRCO. Source code, model checkpoints, and our new datasets are available at https://github.com/mvrl/ConText-CIR.

1. Introduction

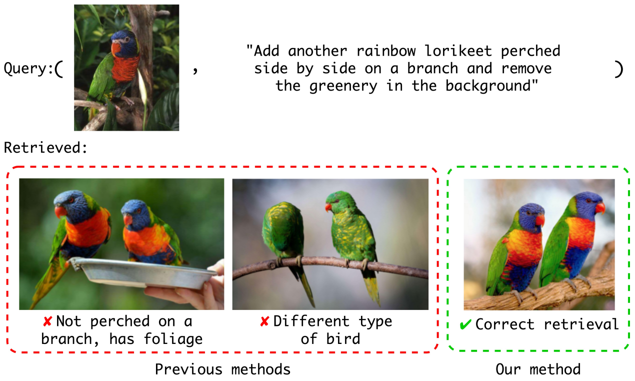

Query:(
,
"Add another rainbow lorikeet perched side by side on a branch and remove the greenery in the background"
)
Retrieved:
✘Not perched on a branch, has foliage
✘Different type of bird
✔Correct retrieval
Previous methods
Our method

Figure 1. A failure case of current composed image retrieval methods. Previous methods do not accurately capture the two conditions specified by the text.

Traditional image retrieval is a longstanding problem in computer vision [8] with a diverse set of applications, including visual search [48], image geolocalization [71], medical imaging [53], etc. There are two standard approaches to the image retrieval task: image-based [67, 73], where the input is an image and the task is finding images that are visually similar, and text-based [39, 56], where the input is natural language, and the task is finding images that match the text. Image embedding models make the image-based approach possible, and vision-language models such as CLIP [44], which have aligned image and text representations, have facilitated the text-based approach. However, both approaches have limitations: images are typically complex and contain a wide variety of objects, making it difficult to encode specific retrieval criteria in an image alone; conversely, it is difficult to specify complex visual information in text alone, as describing details of color patterns and the shapes of interest objects is often difficult and imprecise with text.

Recent work in multi-modal image retrieval aims to mitigate these inherent limitations of uni-modal queries by introducing the composed image retrieval (CIR) problem [24, 32, 61, 65]. CIR involves retrieving a target image specified by a reference image and a text modification to the reference image. Incorporating open-ended text as an additional search criterion allows the inclusion of diverse natural language concepts, modifiers, and search specifications with images to produce a flexible retrieval formulation. Current work in CIR, however, fails to model the interaction of longer, richer texts with images. Existing approaches and datasets focus on simple text modifiers, with a straightforward modification, most often to foreground objects. Real-world queries, however, are often complex, multi-attribute, and may express modifications to background objects. Figure 1 demonstrates the limitations of current CIR methods, where the query includes an image of a "rainbow lorikeet" and text saying to "add another lorikeet on the branch and to remove the greenery in the background". Current methods tend to retrieve results that are correlated with the concepts specified in the text modifier (adding a second lorikeet, but not changing the background, or adding a different type of bird), but do not completely capture the entire relationship between image query and text modification.

We introduce a novel CIR framework, ConText-CIR: Learning from **Con**cepts in **Text** for **C**omposed **I**mage **R**etrieval, that addresses the limitations of previous methods in handling complex, multi-attribute texts. Existing vision-language models like CLIP demonstrate reasonable alignment between text and image concepts when the text is simple. For multi-attribute texts, however, the correspondence between the text and relevant image features often becomes inconsistent. Our proposed concept-consistency (CC) loss addresses this issue by enforcing consistent cross-attention between specific noun phrases and corresponding image regions. Specifically, our loss function encourages the attention weight of a noun phrase in the context of a full sentence to match its weight when evaluated independently, thus reducing contextual interference in long text phrases.

Our framework is trained using a standard CIR formulation and may be easily extended to new datasets as methods and efforts for dataset generation develop. We aim to address the current limited performance of state-of-the-art CIR methods on queries with multi-attribute modifier text, increasing the applicability of CIR models on complex retrieval tasks and setting a new state-of-the-art on current benchmarks. Our specific contributions are the following:

- We introduce a new CIR framework, ConText-CIR, demonstrating state-of-the-art composed image retrieval performance on the CIRR and CIRCO benchmarks.
- We qualitatively show that our concept-consistency loss helps the underlying encoders to learn more specific object-centric representations.
- We introduce a synthetic data generation pipeline to generate multi-attribute text annotations for existing small CIR datasets or to generate new CIR datasets for new domains.

2. Background & Related Work

Our work sets a new state-of-the-art for composed image retrieval. In this section, we summarize the most relevant background and closely related work.

2.1. Composed Image Retrieval

Modern methods in composed image retrieval primarily use multimodal fusion methods to combine representations of the query image and relative text. This produced vision-language embedding may then be used to perform a database lookup to retrieve likely candidate images [3, 5, 12, 38, 57, 61]. The success of pre-trained vision language models [33, 45] has motivated a number of methods that utilize these performant models to extract powerful features to describe the query image and the modification text. Work in CIR has explored a variety of methods, including sophisticated attention-based mechanisms [9, 72], denoising methods [17], and simple interpolation-based approaches [23], to fuse these representations.

Powerful generative VLMs [11, 36] have also inspired training-free methods to perform CIR and composed video retrieval [58, 59] has been used as a proxy task to obtain powerful visuolingual understanding from large video datasets [2]. In another vein, textual inversion-based methods learn pseudoword representations of the image query [4, 47] in the text space of a pre-trained language model. These methods also benefit from learning greater image concept-text alignment by decomposing concepts in an image into multiple pseudo-words [16]. Other methods [63, 68] explicitly aim to learn multimodal alignments but do not learn the grounding of fine-grained text concepts with their corresponding image features. Some methods [32, 50] aim to learn finer-grained alignments between text and image by learning cross-modal masks or differences between features for fashion-centric retrieval. However, all of these methods suffer from poor concept-level alignment between precise text concepts and their corresponding image features due to a lack of structured conceptual guidance, inheriting only coarse-level alignment from contrastive pretraining [10, 51, 70].

2.1.1. Image-Text Datasets

Early work in this area focused on visual question answering regarding image composition. Data sets like NLVR [54], CLEVR [25], and GQA [21] focused on simple synthetic and real images with questions and ground-truth answers about shape, size, and object configurations. The simplicity of these datasets led to the development of NLVR2 [55], a larger dataset of natural image pairs with more complex compositional queries. Fashion-IQ [66], a fine-grained dataset for fashion-focused composed image retrieval, was one of the first true CIR datasets, providing image pairs with "relative" captions describing specific modifications to clothing, along with a baseline multimodal transformer approach for retrieval.

The CIRR dataset [38], derived from NLVR2, includes human annotations focused on modifications to target images, facilitating CIR tasks. Although the most commonly evaluated dataset, the CIRR dataset has some limitations that have been pointed out by others [4], including being limited to image pairs from NLVR2, having many captions that do not relate to the query at all, and only including annotations that describe a single modification to a foreground object. CIRCO [5] is a test-only dataset that consists of human-generated annotations for image pairs mined from the MS-COCO dataset [35] and notably contains more complex text annotations and multiple targets per annotation. While these annotations are high quality and more complex than the CIRR annotations, it does not have a training dataset and can only be used to evaluate the performance of models trained on other data. More recent datasets like LaSCo [30] produce larger training datasets of synthetic annotations by mining examples from larger existing annotated datasets like VQA2.0 [15]. The SynthTriplets18M uses InstructPix2Pix [6] to generate images based on automatically generated text prompts [17]. There are also domain-specific CIR datasets, such as the Birds-to-Words dataset [14] and the

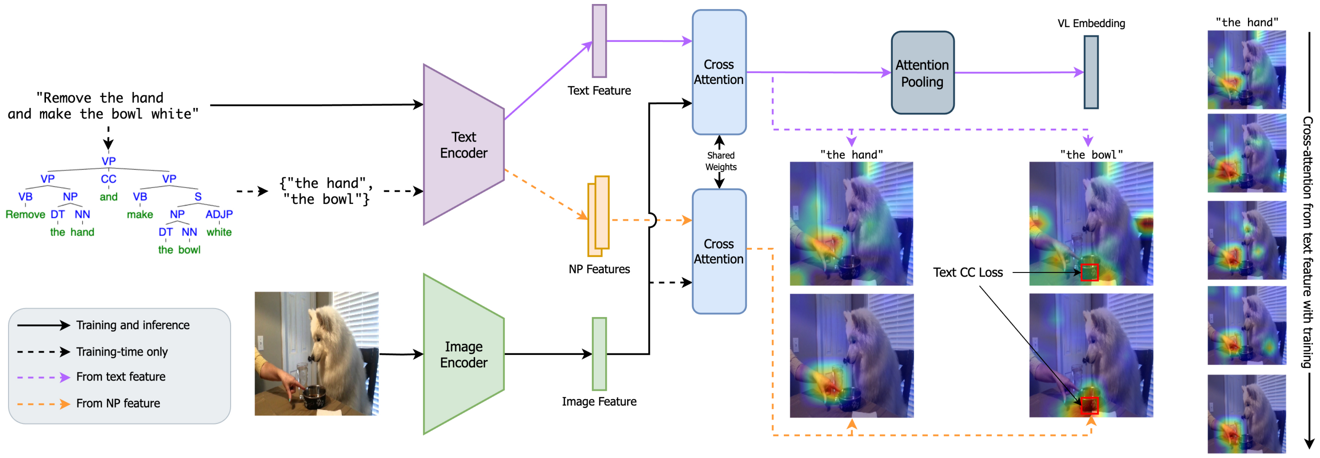


Figure 2. The overall architecture of our approach, ConText-CIR. The framework guides attention to the related image regions by penalizing large differences between attention maps resulting from concept-specific and whole-text representations for each noun phrase. During inference, our method operates efficiently using a simple cross-attention mechanism to combine image and text features. The right side of the figure shows that the cross-attention between the concept "the hand" *from the representation of the entire text* and the image query converges to the local region around the hand with very little spurious attention.

PatterCom remote sensing dataset [42], and video retrieval datasets [59, 60]. In general, existing CIR datasets are limited by the quality of their text annotations, motivating our work to leverage strong multimodal language models with refined prompts for generating high-quality CIR data.

2.2. Visual Grounding

Our proposed ConText-CIR framework enhances visual grounding, or the association between image regions and specific textual concepts, improving query-result alignment. Recent advancements in visual grounding span areas such as visual question answering (VQA), text-guided image editing, and text-guided segmentation. In VQA, Contrastive Region Guidance (CRG) strengthens visual grounding by comparing model outputs with and without text prompts to focus on relevant image areas without extra training [62]. In text-guided image editing, work has focused on enhancing localization via cross-attention mechanisms [19, 49, 64], addressing weak grounding due to loose text-to-attention maps. For Prompt2Prompt[19], cross-attention is refined to control pixel influence per token. LIME employs cross-attention regularization, penalizing irrelevant scores within the ROI, highlighting relevant tokens while downplaying unrelated ones [49]. In image segmentation, models like DINO [7] and SegmentAnything [27] show impressive results but lack *semantic* segmentation capabilities, leading to grounding efforts with natural language, such as Grounding DINO [37] and Grounded-SAM [46]. Other approaches like LSeg [31] train image encoders to align local embeddings to semantic text embeddings, while CMPC [20] enhances grounding by utilizing entity relationships in textual prompts.

3. Methodology

We propose ConText-CIR for training a composed image retrieval model, including a Text Concept-Consistency loss that improves the visual grounding of text concepts in queries.

3.1. Problem Setting

Composed image retrieval is formulated as finding a target image in a database of gallery images that best matches a query specified by a query image and relative text. In practice, this involves pre-indexing the database of gallery images with some embedding function f and retrieving candidates using a representation of the image-text query mapped into the same space.

Formally, we aim to learn a mapping from a query image, relative text tuple (I_q,T) to a target image I_T. We aim to learn a multimodal representation network $r = f(I_q,T)$ that produces a vector to query a database of images.

3.2. The ConText-CIR model

Our multimodal model consists of pair of encoders for image and text, ϕ_I, ϕ_T, a cross attention mechanism ϕ_A, and an attention pooling mechanism ϕ_P. Figure 2 shows a diagram of our pipeline. The final representation of a query pair consisting of an image and modification text is created using cross-attention to fuse the representations of text and image,

$$CrossAttn(I,T) = \phi_A(\phi_I(I),\phi_T(T)),$$

followed by attention pooling to obtain the final vision-language feature r:

$$r = f(I,T) = \phi_P(CrossAttn(I,T)).$$

This model is trained with two loss functions: a contrastive loss between query image-language representations and target image representations and a novel loss that explicitly encourages the cross-attention between text concepts to attend to their relevant image tokens. We describe each in the following sections.

3.3. Text Concept-Consistency Loss

The novel Text Concept-Consistency (Text CC) loss function encourages stronger consistency between noun phrases in the text and the correct image region. Visual language models have a strong capacity to find relationships between simple text and image regions [34, 37], however, representations extracted from multi-concept texts by language models may produce representations that have poorer alignment between particular concepts and their relevant image regions, as global contrastive pretraining methods do not ensure fine-grained alignment. Due to the complex nature of text-image interaction in CIR (involving changing image attributes, removing subjects, etc.), we aim to learn strong concept-level alignments between text and image representations. Following work in linguistics [41] we use noun phrases (NP) to represent concepts in a text.

Before training, we extract at most l noun phrase (NP) constituents in each text T. For each noun phrase NP_i we consider the in-context cross-attention of the tokens:

$$CrossAttn(I,T_{NP_i})$$

between the tokens for the i-th noun-phrase in the context of the original text T, and the isolated cross-attention:

$$CrossAttn(I,NP_i)$$

between the tokens for the i-th noun phrase when represented *by itself* with the text encoder. With these components, we define a loss function to encourage the in-context cross-attention map to match the isolated cross-attention map generated from the concept-specific embedding, summed over all noun phrases extracted from the sentence:

$$\mathcal{L}_{cc} = \sum_{i=1}^{l} \text{ReLU}\big(CrossAttn(I,T_{NP_i}) \\ - CrossAttn(I,NP_i) - \epsilon\big)$$

A slack variable ϵ permits some difference in the magnitude of cross-attention values between in-context and isolated noun-phrase embeddings and is set as a hyperparameter. Intuitively, learning from cross-attentions produced from concept-only embeddings will improve the grounding of concepts to images as the concept-specific embeddings cannot be confused with attributes from other concepts or suppressed by other concepts in the text, leading to focused grounding.

We also enforce the contrastive loss between representations of the query with text modification $r^q = f(I_q, T)$ and the target image $r_n^t = f(I_t, \text{“ ”})$ where I_q, I_t are the query and target images respectively. The target image is encoded with the empty text so that the query vision-language embedding and target image embedding lie in the same space. Following MagicLens [72], we also include the representation of the query image as an additional negative example, $r_n^{q-} = f(I_q, \text{“ ”})$ encoding the query image with the empty text. So, the contrastive loss is defined as follows for the n-th (I_q,T,I_t) training triplet:

$$\mathcal{L}_{cont} = -\log \frac{e^{\text{sim}(r_n^q, r_n^t)/\tau}}{e^{\text{sim}(r_n^q, r_n^{q-})/\tau} + \sum_{i=1}^{N} e^{\text{sim}(r_n^q, r_i^t)/\tau}}$$

where $\text{sim}(r_a, r_b)$ denotes the cosine similarity $\frac{r_a \cdot r_b}{||r_a|| ||r_b||}$, N is the training batch size, and τ is a temperature hyperparameter to scale the distribution of logits. The total loss is given as follows, where λ scales the weight given to each loss:

$$\mathcal{L}_{tot} = \mathcal{L}_{cont} + \lambda \mathcal{L}_{cc}.$$

3.4. ConText-CIR Inference

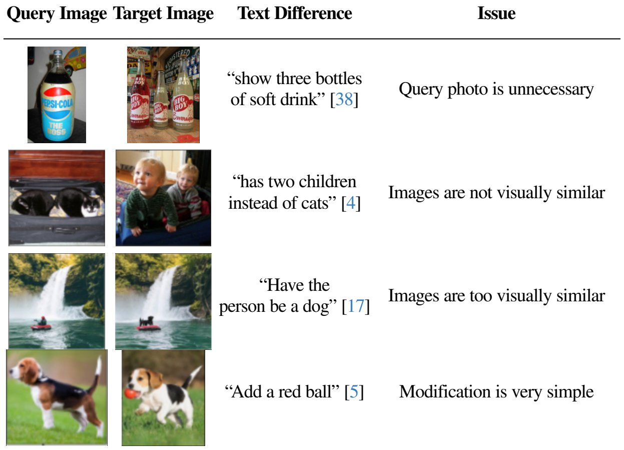
To perform image retrieval with respect to a gallery of candidate images, each candidate image C_i is encoded with a blank text vector as: $r_i = f(C_i, \text{“ ”})$. Given a query text pair I_q, T, each candidate image C_i is scored based on the similarity between r_i and $f(I_q, T)$; we report results in this paper on top-k results for various values of k.

3.5. Automated Data Generation Pipeline

CIR datasets can be divided into two categories: manually generated and automatically generated. Manually generated datasets like CIRR [38] ask human annotators to describe the differences between image pairs (though the pairs are often extracted from existing datasets). Automatically generated datasets tend to leverage existing labeled data to mine CIR queries (e.g., LaSCo [30]), or use image generation tools to produce images that match specific modification text, as in the case of the SynthTriplets18M dataset [17]. There are a variety of problems with these existing datasets regardless of the method of generation, including queries where the text on its own is sufficient to find the target image and issues with the degree of image similarity in the queries. Across existing datasets, the modifications are also typically simple, focusing on a single change to a foreground object. We show examples of these issues in Figure 3.

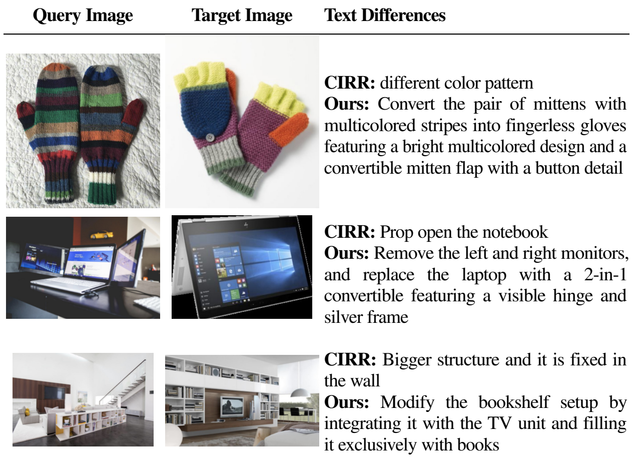
To address these issues, and facilitate the generation of new CIR datasets with realistically complicated text modifications, we introduce a novel pipeline good4cir [28] leveraging a large language model (specifically OpenAI's GPT-4o) to produce CIR triplets.[1] In this pipeline, we assume that there is an existing list of related images–this can come from either an existing CIR dataset with low quality annotations, or from a new domain

[1] The MagicLens paper presents a similar idea of using the PaLM2 LLM to create a CIR dataset, and reports results on a generated training dataset consisting of 36.7 million CIR triplets [72]. As of November 2024, this dataset is unavailable and no code has been shared to replicate it [1].



Query Image	Target Image	Text Difference	Issue
		"show three bottles of soft drink" [38]	Query photo is unnecessary
		"has two children instead of cats" [4]	Images are not visually similar
		"Have the person be a dog" [17]	Images are too visually similar
		"Add a red ball" [5]	Modification is very simple

Figure 3. Qualitative issues with existing CIR datasets.



Query Image	Target Image	Text Differences
		CIRR: different color pattern **Ours:** Convert the pair of mittens with multicolored stripes into fingerless gloves featuring a bright multicolored design and a convertible mitten flap with a button detail
		CIRR: Prop open the notebook **Ours:** Remove the left and right monitors, and replace the laptop with a 2-in-1 convertible featuring a visible hinge and silver frame
		CIRR: Bigger structure and it is fixed in the wall **Ours:** Modify the bookshelf setup by integrating it with the TV unit and filling it exclusively with books

Figure 4. Examples of original captions and rewritten CIRR captions.

with image pairs. We then decompose the CIR triplet generation task into four smaller, targeted tasks to mitigate hallucination and encourage the generation of fine-grained descriptors [13].

First task, we generate a list of objects found in the query image, and have the LLM describe each object with a list of descriptors. Next, we pass this list of objects and associated descriptors to the LLM along with the target image, and ask the LLM to generate a list of objects and descriptors for the target image according to the following criteria:

- If a new object is introduced in the target image, generate a set of descriptors for the given object.
- If there is an object in the target image that matches the descriptors of an object from the query image, adopt the exact set of descriptors to ensure consistency.
- If there is a similar object in the query and target images, but the descriptors provided in the list for the query image do not match the appearance of the object in target image, generate a new set of descriptors for the given object.

Next, we provide the LLM with both lists and have it generate a set of difference captions, each describing a single removal, addition, or modification of an object from the query image to the target image. For a given pair of images, it is possible to have a large number of modifications, each of which can be used to construct a CIR (query image, text modifier, target image) triplet.

The final step of our pipeline is to create additional triplets by concatenating multiple text modifiers into increasingly complex queries. We show examples of generated triplets in Figure 4 and provide additional details and examples in the Appendix.

4. Experimental Setup

ConText-CIR achieves significant performance improvement relative to the state-of-the-art across a broad range of standard evaluation protocols, all without increasing inference-time complexity or using a greater magnitude of training data.

4.1. Training Data

We train on a variety of datasets, including CIRR [38] and LaSCo [30], as well as two datasets generated using our pipeline. The first generated dataset is a rewritten version of the CIRR dataset. CIRR is one of the most commonly used CIR datasets for both training and evaluation, however, as demonstrated in Figures 3 and 4, the existing CIRR dataset contains relatively short and minimal captions that do not translate to domains where more detailed text queries are advantageous. Thus, we generate a rewritten CIRR dataset, CIRR$_R$, that applies the synthetic data pipeline to generate longer and more linguistically rich captions for the existing CIRR image pairs.

We additionally generate a new dataset, Hotel-CIR. This dataset is sourced from visually similar image pairs in the TraffickCam [52] hotel image database. This is a compelling domain, as the scenes are dense, with a large number of objects, and high degrees of visual similarity between images that don't necessarily come from the same class. The Appendix includes additional information regarding the CIRR$_R$ and Hotel-CIR datasets, including the exact prompts used to generate them and dataset statistics. We share these datasets to motivate further work in building CIR methods that are capable of supporting real-world queries.

We perform ablations on the inclusion of different datasets in the training process, and ultimately train a CIR model on a combined dataset of CIRR, CIRR$_R$, LaSCo, and Hotel-CIR, which we refer to as the Aggregated dataset.

4.2. Baseline Methods

We evaluate our method against state-of-the-art CIR methods that pre-index a database of gallery images with vector embeddings. For fairness, we compare to models initialized with CLIP ViT-B/L or OpenCLIP ViT-H when available. We evaluate against a number of textual inversion-based methods [4, 16, 47], late multi-modal fusion-based meth-

			Recall@K				Recall$_{subset}$@K		
Backbone	Method	Additional Data	K=1	K=5	K=10	K=50	K=1	$K=2$	$K=3$
$\leq$ ViT-B	ARTEMIS* [12]	FashionIQ	16.96	46.10	61.31	87.73	39.99	62.20	75.67
	CIRPLANT [38]	FashionIQ	19.55	52.55	68.39	92.38	39.20	63.03	79.49
	LF-BLIP [3]	FashionIQ	20.89	48.07	61.16	83.71	50.22	73.16	86.82
	Combiner [3]*	FashionIQ	33.59	65.35	77.35	95.21	62.39	81.81	92.02
	CLIP4CIR [5]*	FashionIQ	44.82	77.04	86.65	97.90	73.16	88.84	95.59
	CASE [30]	LaSCo	48.68	79.98	88.51	97.49	76.39	90.12	95.86
	CASE [30]	LaSCo+COCO	49.35	80.02	88.75	97.47	76.48	90.37	95.71
	ConText-CIR (ours)	Aggregated	**49.83**	**81.54**	**89.76**	**98.95**	**76.64**	**90.69**	**96.31**
ViT-L	CompoDiff [17]	ST18M+LAION2B	22.35	54.36	73.41	91.77	62.55	81.44	90.21
	CoVR-BLIP [59]	WebVid-CoVR	49.69	78.60	86.77	94.31	75.01	88.12	93.16
	ConText-CIR (ours)	Aggregated	**52.65**	**83.27**	**89.51**	**98.87**	**80.32**	**92.13**	**96.08**
ViT-H	ConText-CIR (ours)	Aggregated	**55.24**	**84.85**	**90.75**	**98.82**	**82.96**	**93.12**	**97.04**
ViT-G	CompoDiff [17]	ST18M+LAION2B	32.39	57.61	77.25	94.61	67.88	85.29	94.07
	CoVR-2 [58]	CC-CoIR	50.63	81.04	89.35	98.15	76.53	90.43	96.00
	CoVR-2 [58]	WV-CC-CoVIR	50.43	81.08	88.89	98.05	76.75	90.34	95.78

Table 1. Composed image retrieval results on the CIRR test set. Higher is better for all metrics, best results are shown in bold. The asterisks (*) in the $\leq$ ViT-B section denote that a ResNet-based backbone was used.

ods [3, 5, 12, 23, 38, 72], early fusion-based methods [30], composed video retrieval-based methods that can perform CIR [58, 59], and training-free methods [26, 57, 69].

We evaluate our framework on the CIRR [38] and CIRCO [4] test sets. CIRR also defines a subset retrieval metric, Recall$_{subset}$@K, where the model performs retrieval among a focused small subset of images for each query. The CIRCO [4] dataset addresses certain limitations of CIRR by increasing the gallery index size and by providing multiple ground-truth targets per image-text query. CIRCO sources images from the COCO 2017 unlabeled image set [35]. As there are multiple ground-truths per image-text query, we report mean average precision (mAP@K).

Additionally, we evaluate the performance of models trained on the Aggregated dataset on the CIRCO test set and models trained with LaSCo and Hotel-CIR on the CIRR test set, a CIR task setting defined in the literature as *zero-shot* CIR. Additional zero-shot results on FashionIQ [66] and ImageNet-R [18] are provided in the supplemental materials.

4.3. Implementation Details

We use Stanza [43] to perform constituency parsing over each modification text, extracting at most $l=10$ noun phrases from each parse tree. The noun phrases are extracted with a breadth-first search, so the branch-level noun phrases are prioritized to be kept over the leaf-level noun phrases. We utilize the official image and text encoders from CLIP ViT-B and ViT-L [44] and OpenCLIP ViT-H [22]. We follow the attention pooling formulation of Set Transformer [29]. We use the AdamW optimizer [40] with a cosine annealed learning rate cycling between `2e-5` and `2e-7` and a weight decay of `1e-2`.

4.4. Main Results

Table 1 gives both Recall@K and Recall$_{subset}$@K metrics on the CIRR test set. When isolating each backbone, we observe that our largest model demonstrates the highest retrieval performance across both Recall@K and Recall$_{subset}$@K metrics. Within each category of encoder size, ConText-CIR outperforms baseline methods across all metrics, especially for models using ViT-L and $\geq$ViT-H methods. Notably, ConText-CIR improves R@1 from 50.43 to 55.24 and R@1$_{subset}$ from 76.75 to 82.96. Unlike the second-best method CoVR-2 [58], ConText-CIR does not use ViT-G, a backbone with 60% more parameters than ViT-H, and does not pretrain on a dataset of 4.9 million samples. For the ViT-L and $\leq$ViT-B encoder sizes we also outperform baseline methods on all benchmarks.

4.5. Zero-shot Composed Image Retrieval

Table 2 gives zero-shot (excluding CIRR from training data) composed image retrieval metrics on CIRR. ConText-CIR demonstrates powerful zero-shot CIR performance across all encoder sizes, improving state-of-the-art R@1 by 4.78 for ViT-B, 5.38 for ViT-L, and 12.88 for ViT-H. Notably, our method with the ViT-H backbone also outperforms all methods using the much larger ViT-G. We observe that our method is able to outperform ViT-G-based CoVR-2 [58] without pertaining on 4.9 million samples.

We also report zero-shot composed image retrieval performance on CIRCO [4] in Table 3. Our model outperforms all baseline methods when compared with the same backbone architecture, and outperforms all methods based on ViT-G, even as the ViT-G backbone has almost 400M additional parameters.

			Recall@K				Recall$_{subset}$@K		
Backbone	Method	Training Data	K=1	K=5	K=10	K=50	K=1	$K=2$	$K=3$
ViT-B	SEARLE-OTI [4]	ImageNet1K	24.27	53.25	66.10	88.84	54.10	75.81	87.33
	SEARLE [4]	ImageNet1K	24.00	53.42	66.82	89.78	54.89	76.60	88.19
	LDRE [69]	-	25.69	55.13	69.04	89.90	60.53	80.65	90.70
	CIReVL [26]	-	23.94	52.51	66.00	86.95	60.17	80.05	90.19
	MagicLens [72]	web-scraped (36.7M)	27.0	58.0	70.9	91.1	66.7	83.9	92.4
	Slerp+TAT [23]	CC3M+LLaVA-Align+Laion-2M	28.19	55.88	68.77	88.51	61.13	80.63	90.68
	CASE [30]	LaSCo	30.89	60.75	73.88	92.84	60.17	80.17	90.41
	CASE [30]	LaSCo+COCO	35.40	65.78	78.53	94.63	64.29	82.66	91.61
	ours	LaSCo+Hotel-CIR	**40.18**	**70.04**	**81.56**	**96.21**	**72.25**	**87.46**	**94.52**
ViT-L	CompoDiff [17]	ST18M+LAION2B	22.35	54.36	73.41	91.77	62.55	81.44	90.21
	Pic2Word [47]	CC3M	23.90	51.70	65.30	87.80	-	-	-
	SEARLE-XL-OTI [4]	CC3M	24.87	52.31	66.29	88.58	53.80	74.31	86.94
	SEARLE-XL [4]	CC3M	24.24	52.48	66.29	88.84	53.76	75.01	88.19
	CIReVL [26]	-	24.55	52.31	64.92	86.34	59.54	79.88	89.69
	LinCIR [16]	CC3M+SDP+COYO700M+OWT	25.04	53.25	66.68	-	57.11	77.37	88.89
	LDRE [69]	-	26.53	55.57	67.54	88.50	60.43	80.31	89.90
	Slerp+TAT [23]	CC3M+LLaVA-Align+Laion-2M	30.94	59.40	70.94	89.18	64.70	82.92	92.31
	MagicLens [72]	web-scraped (36.7M)	30.1	61.7	74.4	92.6	68.1	84.8	93.2
	CoVR-BLIP [59]	WebVid-CoVR	38.48	66.70	77.25	91.47	69.28	83.76	91.11
	ours	LaSCo+Hotel-CIR	**43.86**	**73.39**	**82.00**	**96.14**	**75.62**	**89.44**	**95.03**
ViT-H	LinCIR [16]	CC3M+SDP+COYO700M+OWT	33.83	63.52	75.35	-	62.43	81.47	92.12
	ours	LaSCo+Hotel-CIR	**46.71**	**75.64**	**84.26**	**96.85**	**76.41**	**90.64**	**95.62**
ViT-G	GRB+LCR [57]	-	30.92	56.99	68.58	85.28	66.67	78.68	82.60
	CompoDiff [17]	ST18M+LAION2B	32.39	57.61	77.25	94.61	67.88	85.29	94.07
	TFCIR [57]	-	32.82	61.13	71.76	85.28	66.63	78.58	82.68
	CIReVL [26]	-	34.65	64.29	75.06	91.66	67.95	84.87	93.21
	LinCIR [16]	CC3M+SDP+COYO700M+OWT	35.25	64.72	76.05	-	63.35	82.22	91.98
	LDRE [69]	-	36.15	66.39	77.25	93.95	68.82	85.66	93.76
	CoVR-2 [58]	CC-CoIR	43.35	73.78	83.66	96.07	75.25	88.89	95.23
	CoVR-2 [58]	WV-CC-CoVIR	43.74	73.61	83.95	96.10	72.84	87.52	94.39

Table 2. Zero-shot composed image retrieval results on the CIRR test set. Higher is better for all metrics, best results are shown in bold.

4.6. Ablation Studies

We evaluate ConText-CIR on CIRR trained using different components of our Aggregated dataset in Table 4. We observe that ConText-CIR performs better as more data is added. The significant performance increase from 45.25 to 48.54 R@1 from CIRR to CIRR+CIRR$_{\mathbf{R}}$ indicates that our data generation pipeline is highly effective at synthesizing data for training CIR models, producing large performance increases even without incorporating any additional imagery. Each subsequent large-scale dataset that does include new imagery (LaSCo and Hotel-CIR) gives an additional large performance boost.

Next, we analyze the elements of ConText-CIR's design. We experiment with freezing the image or the text encoder, not considering branch-level noun phrases, and removing the Text CC loss. Freezing the text encoder results in a much larger performance drop when compared to freezing the image encoder (performance drop from 54.28 to 49.52 in R@1), indicating that the pre-trained CLIP text space is less conducive for CIR than the pre-trained CLIP image space. By comparison, only considering leaf noun phrases (a noun phrase that cannot be broken down into smaller syntactic units within the noun phrase itself) and removing the query image as a hard negative only results in a small decrease in performance. Finally, we see that removing the Text CC loss has a significant negative impact on performance (from 55.24 to 48.92 in R@1). This supports the idea that guiding the model to fuse representations of text concepts to their relevant image regions leads to higher performance.

4.7. Qualitative Results

In Figure 5, we examine the cross attention of noun phrase representations to the image query for models trained with and without the Text CC loss. The attention maps are averaged across the cross-attention of the tokens *from the representation of the whole text* in each noun phrase to the image query.

We qualitatively observe that the model's ability to ground the representations of noun phrases to their relevant image features improves significantly with the Text CC loss. Notably, the model trained without the Text CC loss often diffuses attention onto extraneous objects, such as the toilet and mirror in the first row. Furthermore, when the CC loss isn't used, the model often targets attention to completely wrong tokens. For instance, the model trained without the Text CC loss focuses attention associated with the concept "the bathroom towels" onto the wastebasket. These observations demonstrate how the CC loss guides the

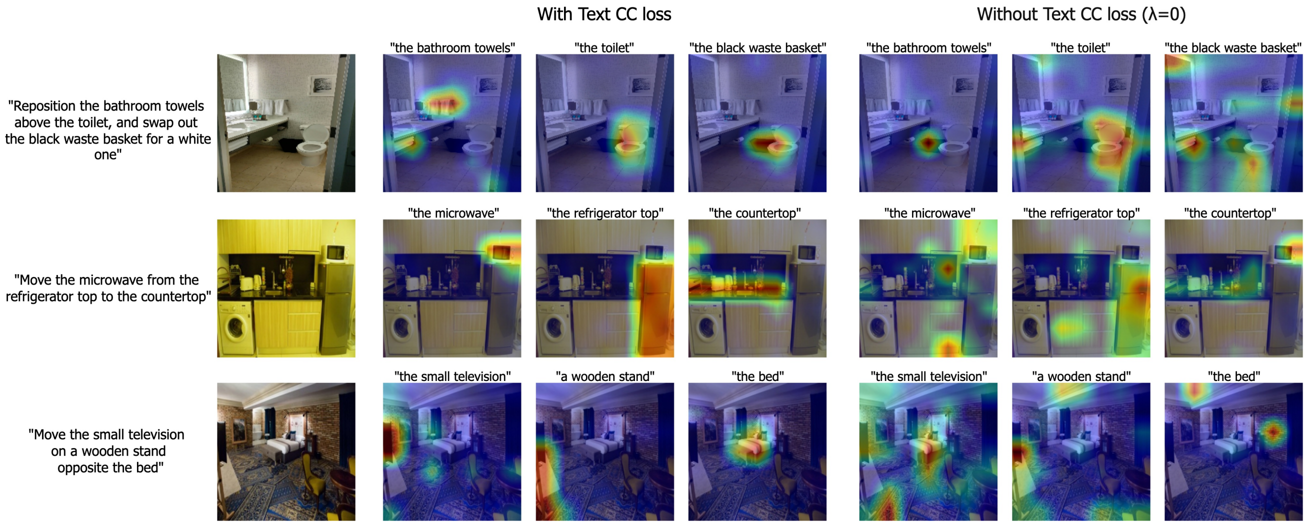


Figure 5. Noun-phrase-level cross attention maps for models trained with and without the Text Concept-Consistency loss.

		mAP@K			
Size	Method	K=5	K=10	K=25	K=50
B	SEARLE [4]	9.35	9.94	11.13	11.84
	CIReVL [26]	14.94	15.42	17.00	17.82
	LDRE [69]	17.96	18.32	20.21	21.11
	MagicLens [72]	23.1	23.8	25.8	26.7
	GRB+LCR [57]	25.38	26.93	29.82	30.74
	TFCIR [57]	26.52	28.25	31.23	31.99
	ours	**28.12**	**29.42**	**32.26**	**33.54**
L	SEARLE-XL [4]	11.68	12.73	14.33	15.12
	CompoDiff [17]	12.31	13.51	15.67	16.15
	LinCIR [16]	12.59	13.58	15.00	15.85
	CIReVL [26]	18.57	19.01	20.89	21.80
	CoVR-BLIP [59]	21.43	22.33	24.47	25.48
	LDRE [69]	23.35	24.03	26.44	27.50
	MagicLens [72]	29.6	30.8	33.4	34.4
	ours	**30.05**	**30.53**	**34.79**	**34.72**
H	ours	**31.74**	**32.64**	**35.05**	**36.42**
G	CompoDiff [17]	15.33	17.71	19.45	21.01
	LinCIR [16]	19.71	21.01	23.13	24.18
	CIReVL [26]	26.77	27.59	29.96	31.03
	CoVR-2 [58]	28.29	29.55	32.18	33.26
	LDRE [69]	31.12	32.24	34.95	36.03

Table 3. Zero-shot composed image retrieval results on the CIRCO test set. Higher is better for all metrics, best results are shown in bold.

model to learn to fuse concepts in text to their relevant image features *without computing concept features at inference time*.

	Recall@K			
Data	K=1	K=5	K=10	K=50
CIRR	45.25	77.52	86.88	97.24
CIRR+CIRR$_\mathbf{R}$	48.54	80.12	88.52	97.49
CIRR+CIRR$_\mathbf{R}$+LaSCo	53.12	83.78	89.48	98.67
Aggregated	55.24	84.85	90.75	98.82

Table 4. Ablation study over dataset components w/ ViT-H backbone.

	Recall@K			
Ablation	K=1	K=5	K=10	K=50
Freeze text encoder	49.52	78.46	86.33	94.25
Freeze image encoder	54.18	83.84	89.54	98.15
No Query Negative	54.52	83.95	89.57	98.12
Only leaf NPs	54.67	84.02	89.62	98.08
No Text CC ($\lambda=0$)	48.92	79.86	88.74	97.41
With Text CC ($\lambda=0.08$)	55.24	84.85	90.75	98.82

Table 5. Ablation study over model choices w/ ViT-H backbone.

5. Conclusions

We present ConText-CIR, a novel CIR method facilitating performant rapid retrieval over large image databases. Our Text Concept-Consistency loss improves the concept-level alignment of the text and image embeddings and produces qualitatively more focused attention maps. To train ConText-CIR, we also introduce a data generation strategy that can be used to rewrite the annotations in existing CIR datasets or to generate new CIR datasets. This method trained on these new datasets sets a new state-of-the-art in supervised and zero-shot composed image retrieval on multiple CIR benchmarks.

Acknowledgements

This material is based upon work done, in whole or in part, in coordination with the Department of Defense (DoD). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DoD and/or any agency or entity of the United States Government.

References

[1] https://github.com/google-deepmind/magiclens/issues/8. 4

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 2

[3] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21434–21442, 2022. 2, 6

[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 (ICCV)*, pages 15338–15347, 2023. 2, 5, 6, 7, 8

[5] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Composed image retrieval using contrastive learning and task-oriented clip-based features. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–24, 2023. 2, 5, 6

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 2

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3

[8] Wei Chen, Yu Liu, Weiping Wang, Erwin M. Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S. Lew. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7270–7292, 2023. 1

[9] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2998–3008, 2020. 2

[10] Yuxiao Chen, Jianbo Yuan, Yu Tian, Shijie Geng, Xinyu Li, Ding Zhou, Dimitris N. Metaxas, and Hongxia Yang. Revisiting multimodal representation in contrastive learning: From patch and token embeddings to finite discrete tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15095–15104, 2023. 2

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 2

[12] 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. 2, 6

[13] James Flemings, Wanrong Zhang, Bo Jiang, Zafar Takhirov, and Murali Annavaram. Characterizing context influence and hallucination in summarization, 2024. 5

[14] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. Neural naturalist: Generating fine-grained image comparisons. *arXiv preprint arXiv:1909.04101*, 2019. 2

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[16] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, , Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 5, 7, 8

[17] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *Transactions on Machine Learning Research*, 2024. Expert Certification. 2, 4, 5, 6, 7, 8

[18] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 6

[19] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022. 3

[20] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10488–10497, 2020. 3

[21] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 2

[22] 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. 6

[23] Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. arXiv preprint arXiv:2405.00571, 2024. 2, 6, 7

[24] Xintong Jiang, Yaxiong Wang, Yujiao Wu, Meng Wang, and Xueming Qian. Dual relation alignment for composed image retrieval, 2024. 1

[25] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 2

[26] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *International Conference on Learning Representations (ICLR)*, 2024. 6, 7, 8

[27] 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. *arXiv:2304.02643*, 2023. 3

[28] Pranavi Kolouju, Eric Xing, Robert Pless, Nathan Jacobs, and Abby Stylianou. good4cir: Generating detailed synthetic captions for composed image retrieval, 2025. 4

[29] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3744–3753, 2019. 6

[30] M. Levy, R. Ben-Ari, N. Darshan, and D. Lischinski. Data roaming and quality assessment for composed image retrieval. In *AAAI*, 2024. 2, 4, 5, 6, 7

[31] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 3

[32] Dafeng Li and Yingying Zhu. Visual-linguistic alignment and composition for image retrieval with text feedback. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 108–113, 2023. 1, 2

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[34] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 4

[35] 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 *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 2, 6

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 2

[37] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3, 4

[38] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134, 2021. 2, 4, 5, 6

[39] Zijun Long, Xuri Ge, Richard Mccreadie, and Joemon Jose. Cfir: Fast and effective long-text to image retrieval for large corpora, 2024. 1

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[41] G. Murphy. Noun phrase interpretation and conceptual combination. *Journal of Memory and Language*, 29(3), 1990. 4

[42] B. Psomas, I. Kakogeorgiou, N. Efthymiadis, G. Tolias, O. Chum, Y. Avrithis, and K. Karantzalos. Composed image retrieval for remote sensing. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 2024. 3

[43] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020. 6

[44] 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*, pages 8748–8763. PMLR, 2021. 1, 6

[45] 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

[46] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 3

[47] 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. *CVPR*, 2023. 2, 5, 7

[48] Devashish Shankar, Sujay Narumanchi, H A Ananya, Pramod Kompalli, and Krishnendu Chaudhury. Deep learning based large scale visual recommendation and search for e-commerce, 2017. 1

[49] Enis Simsar, Alessio Tonioni, Yongqin Xian, Thomas Hofmann, and Federico Tombari. Lime: Localized image editing via attention regularization in diffusion models, 2023. 3

[50] Chull Hwan Song, Taebaek Hwang, Jooyoung Yoon, Shunghyun Choi, and Yeong Hyeon Gu. Syncmask: Synchronized attentional masking for fashion-centric vision-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13948–13957, 2024. 2

[51] Kun Song, Huimin Ma, Bochao Zou, Huishuai Zhang, and Weiran Huang. Fd-align: Feature discrimination alignment for fine-tuning pre-trained models in few-shot learning. *NeurIPS*, 2023. 2

[52] Abby Stylianou, Hong Xuan, Maya Shende, Jonathan Brandt, Richard Souvenir, and Robert Pless. Hotels-50k: A global hotel recognition dataset. In *The AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 5

[53] Dhanya K. Sudhish, Latha R. Nair, and Shailesh S. Content-based image retrieval for medical diagnosis using fuzzy clustering and deep learning. *Biomedical Signal Processing and Control*, 88: 105620, 2024. 1

[54] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017. 2

[55] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. 1:6418–6428, 2019. 2

[56] Manal Sultan, Lia Jacobs, Abby Stylianou, and Robert Pless. Exploring clip for real world, text-based image retrieval. In *2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–6, 2023. 1

[57] Shitong Sun, Fanghua Ye, and Shaogang Gong. Training-free zero-shot composed image retrieval with local concept reranking. arXiv preprint arXiv:2312.08924, 2024. 2, 6, 7, 8

[58] L. Ventura, A. Yang, C. Schmid, and G. Varol. CoVR-2: Automatic Data Construction for Composed Video Retrieval . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(12):11409–11421, 2024. 2, 6, 7, 8

[59] L. Ventura, A. Yang, C. Schmid, and G. Varol. CoVR: Learning composed video retrieval from web video captions. *AAAI*, 2024. 2, 3, 6, 7, 8

[60] L. Ventura, A. Yang, C. Schmid, and G. Varol. CoVR-2: Automatic data construction for composed video retrieval. *IEEE TPAMI*, 2024. 3

[61] 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, 2

[62] David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. In *ECCV*, 2025. 3

[63] Yongquan Wan, Wenhai Wang, Guobing Zou, and Bofeng Zhang. Cross-modal feature alignment and fusion for composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 8384–8388, 2024. 2

[64] Kai Wang, Fei Yang, Shiqi Yang, Muhammad Atif Butt, and Joost van de Weijer. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. In *Advances in Neural Information Processing Systems*, pages 26291–26303. Curran Associates, Inc., 2023. 3

[65] 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*, page 915–923. ACM, 2023. 1

[66] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *CVPR*, 2021. 2, 6

[67] Hui Wu, Min Wang, Wengang Zhou, Zhenbo Lu, and Houqiang Li. Asymmetric Feature Fusion for Image Retrieval . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11082–11092, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 1

[68] Yahui Xu, Yi Bin, Jiwei Wei, Yang Yang, Guoqing Wang, and Heng Tao Shen. Align and retrieve: Composition and decomposition learning in image retrieval with text feedback. *IEEE Transactions on Multimedia*, 26:9936–9948, 2024. 2

[69] 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*, pages 80–90, 2024. 6, 7, 8

[70] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: fine-grained interactive language-image pre-training. *CoRR*, abs/2111.07783, 2021. 2

[71] Zelong Zeng, Zheng Wang, Fan Yang, and Shin'ichi Satoh. Geo-localization via ground-to-satellite cross-view image retrieval. *IEEE Transactions on Multimedia*, 25:2176–2188, 2023. 1

[72] K. Zhang, Y. Luan, H. Hu, K. Lee, S. Qiao, W. Chen, Y. Su, and M-W. Chang. MagicLens: Self-supervised image retrieval with open-ended instructions. In *PMLR*, 2024. 2, 4, 6, 7, 8

[73] Yidan Zhang, Ting Zhang, Dong Chen, Yujing Wang, Qi Chen, Xing Xie, Hao Sun, Weiwei Deng, Qi Zhang, Fan Yang, Mao Yang, Qingmin Liao, Jingdong Wang, and Baining Guo. Irgen: Generative modeling for image retrieval, 2024. 1