MagicLens: Self-Supervised Image Retrieval with Open-Ended Instructions

Kai Zhang [*1] **Yi Luan** [2] **Hexiang Hu** [2] **Kenton Lee** [2] **Siyuan Qiao** [2] **Wenhu Chen** [2] **Yu Su** [1] **Ming-Wei Chang** [2]

Abstract

Image retrieval, i.e., finding desired images given a reference image, inherently encompasses rich, multi-faceted search intents that are difficult to capture solely using image-based measures. Recent works leverage text instructions to allow users to more freely express their search intents. However, they primarily focus on image pairs that are visually similar and/or can be characterized by a small set of pre-defined relations. The core thesis of this paper is that text instructions can enable retrieving images with *richer relations beyond visual similarity*. To show this, we introduce MagicLens, a series of self-supervised image retrieval models that support open-ended instructions. MagicLens is built on a key novel insight: image pairs that naturally occur on the same web pages contain a wide range of *implicit* relations (e.g., `inside view of`), and we can bring those implicit relations *explicit* by synthesizing instructions via foundation models. Trained on 36.7M (*query image, instruction, target image*) triplets with rich semantic relations mined from the web, MagicLens achieves results comparable with or better than prior best on eight benchmarks of various image retrieval tasks, while maintaining high parameter efficiency with a significantly smaller model size. Additional human analyses on a 1.4M-image unseen corpus further demonstrate the diversity of search intents supported by MagicLens. Code and models are publicly available at the Project Website.

1. Introduction

Image retrieval is a long-established problem in computer vision (Datta et al., 2008; Gordo et al., 2016) with a wide range of real-world applications, such as visual search, object localization, and re-identification. However, since its inception, this task has suffered from ambiguous definitions due to the complex and rich content encapsulated in images. Similar images may differ in key aspects, and different images can share commonalities. In image search scenarios, users frequently present multiple search intents for a single query image, indicating that mere image relevance is insufficient for precise search results. For instance, when searching with an image of the `Burj Al Arab hotel in Dubai` (see Figure 1), a user might seek other attractions in Dubai or an interior view, each relating differently to the query image. Therefore, incorporating text instructions that articulate search intents is essential and indispensable for enhancing retrieval accuracy. Ideally, models should accurately capture and interpret diverse real-world search intents as conveyed by *open-ended text instructions*.

These open-ended search instructions, span a wide range of topics and concepts, and reflect the diverse ways users

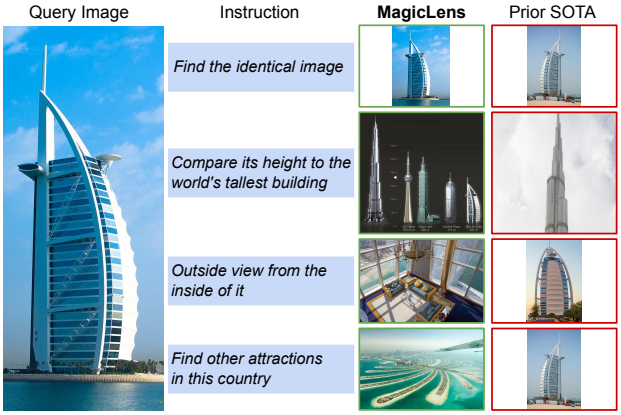

Figure 1. Top-1 retrieved images using MagicLens and the prior state-of-the-art (SOTA) method (Gu et al., 2024) from a retrieval pool with 1.4M images. The prior SOTA method, while accepting text instructions, primarily retrieves images based on visual similarity to the query image, ignoring the nuances of the text instructions. In contrast, MagicLens excels at retrieving both visually similar images and those that align with the deeper meaning and context of the text instructions — even when the images do not resemble the query. For example, if given a query image of the Burj Al Arab and the instruction "Find other attractions in this country", it can successfully locate images of the Palm Islands in Dubai.

[*]Work done at Google DeepMind. [1]The Ohio State University [2]Google DeepMind. Correspondence to: Kai Zhang <zhang.13253@osu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

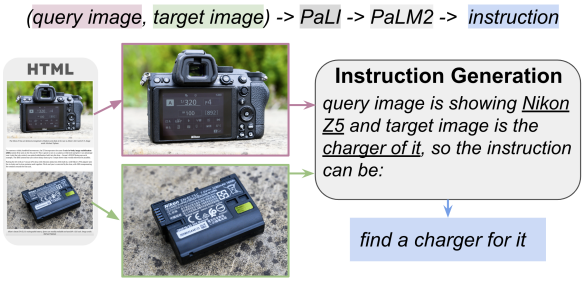

Figure 2. Data construction overview. We collect ***naturally occurring image pairs*** from the same web pages and use PaLI+PaLM2 to generate instructions connecting the two images.

interact with visual content, requiring the retrieval system to grasp not only the visual features of an image but also the nuanced semantic relation between the query image and desired results as expressed in the instructions. Existing models, however, are optimized towards one or a few restricted domains (Vo et al., 2019; Wu et al., 2021; Liu et al., 2021; Baldrati et al., 2023), where the types of visual similarities are manually defined as *a prior*. They either adjust the model architecture and training recipe to utilize image-caption data (Chen & Lai, 2023; Saito et al., 2023; Baldrati et al., 2023; Gu et al., 2024), or rely on synthetic data constructed from pre-defined instruction templates (Brooks et al., 2023; Gu et al., 2023). As a result, neither of these research directions can effectively model open-ended instructions, evidenced by Figure 1.

In this paper, we present MagicLens, a series of self-supervised image retrieval models trained on a wide range of (`query image`, `instruction`, `target image`) triplets that reflect *naturally occurring* semantic relations, mined from web pages and curated with state-of-the-art (SOTA) foundation models. Specifically, we extract image pairs that *naturally occur* on the same web page to form positive pairs that carry abundant but natural semantic relations. We then apply both large multimodal models (LMMs; Chen et al. (2023b;a)) and large language models (LLMs; Anil et al. (2023)) to refine the description of such open-ended semantic relation, into open-ended instruction. Figure 2 shows an overview of the data construction pipeline. For example, a camera review website[1] presenting the image of a `Nikon Camera` and the image of a `Nikon Charger` would offer an interesting and non-trivial relation *"charger of a product"*, which would then be curated by the LMM+LLM pipeline, and produce a final instruction `find a charger for it`. This process produces open-ended instructions that depict diverse semantic relations beyond mere visual similarity, resulting in a large-scale training dataset with 36.7M high-quality triplets over a wide distribution.

With the constructed dataset, we train dual-encoder models called MagicLens, which retrieve images given a query consisting of an image with an instruction. Our models achieve results comparable with or better than prior SOTA methods on eight benchmarks, including various multimodality-to-image and image-to-image retrieval tasks. In addition, MagicLens can retain or even significantly improve the text-to-image retrieval performance of the underlying single-modality encoders. With a 50 times smaller model size than prior SOTA methods, MagicLens outperforms them on multiple benchmarks: CIRCO (Baldrati et al., 2023), Domain Transfer ImageNet (Saito et al., 2023), and GeneCIS (Vaze et al., 2023). To further examine our models' capabilities in a more realistic scenario, we construct the largest retrieval pool to date with 1.4 million unseen images and perform retrieval given human-written search queries with diverse instructions. The human evaluation finds that MagicLens can successfully satisfy complex and beyond visual search intents, whereas the prior SOTA fails to do so.

Our contributions are threefold:

- We bring a novel insight for image retrieval: naturally occurring image pairs from the same web pages are strong self-supervised training signals. Based on this, we propose an effective pipeline, backed with LMMs and LLMs, to construct training data consisting of 36.7M triplets.
- We introduce MagicLens, a series of light-weight dual-encoders that jointly embed a pair of image and instruction, trained on the constructed dataset. Across multiple benchmarks, MagicLens outperforms previous SOTA methods but with a 50$\times$ smaller model size.
- We conduct an in-depth human evaluation and analysis on a 1.4M-scale retrieval pool, which is the largest to date. Remarkably high success rates show that MagicLens can well capture and satisfy diverse search intents, especially complex and beyond visual ones.

2. Related Work

Pre-Training Multimodal Encoders. Multimodal encoder pre-training (Faghri et al., 2017; Chen et al., 2021; Radford et al., 2021; Yu et al., 2022; Li et al., 2021; Kim et al., 2021; Wang et al., 2023; Li et al., 2022; 2023; Cherti et al., 2023) has witnessed great success in recent years. Pre-trained on web-scale image-caption data (Zhai et al., 2022; Schuhmann et al., 2022), these models align the representations of different modalities in a joint space, enabling zero-shot cross-modality retrieval. However, these works focus on encoding single modalities, without considering the composed representation of multiple modalities. Some later efforts (Hu et al., 2023a; Chen et al., 2023c; Wei et al., 2023) attempt to combine text and image embeddings via fine-tuning a small number of parameters on top of pre-trained single-modal encoders, without large-scale joint pre-training. Consequently, such an adaptation strategy shows inferior results on the

[1] https://amateurphotographer.com/review/nikon-z5-review/

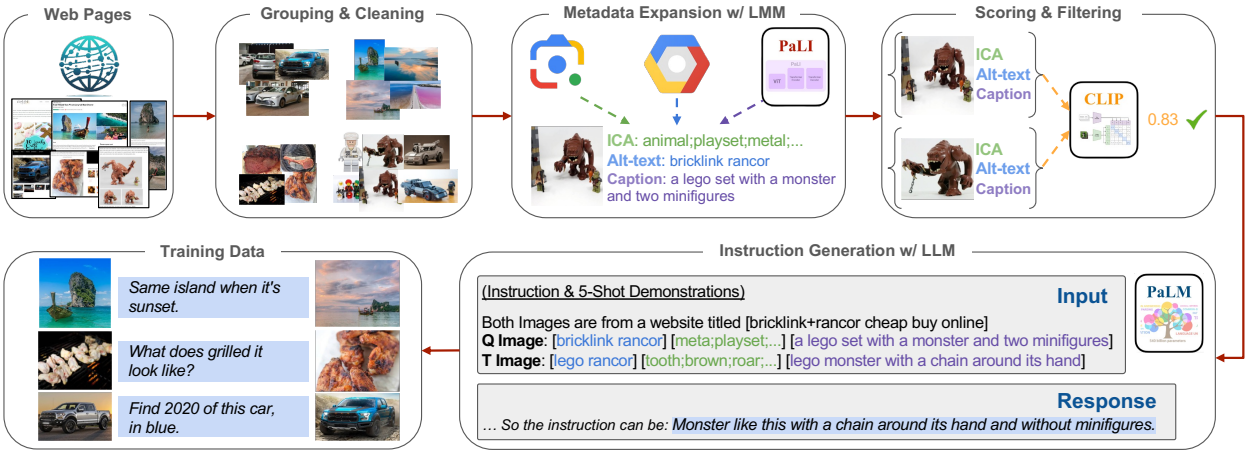

Figure 3. Data construction pipeline. We mine image pairs from the web via (1) grouping images from the same web page and cleaning them, (2) annotating metadata for each image with LMMs, and (3) scoring and filtering out unqualified image pairs. Eventually, we generate open-ended instructions using LLMs for the remaining image pairs.

task of our interest, emphasizing the importance of the MagicLens' self-supervised training.

Composed Image Retrieval. Composed image retrieval (CIR; Vo et al. (2019)) shares the same task form with us. However, all existing benchmarks (Liu et al., 2021; Baldrati et al., 2023; Wu et al., 2021) collect visually similar images first and then write instructions for image pairs. This limits the richness of image relations on these benchmarks and the models developed upon/for them. Recent works on zero-shot CIR (Saito et al., 2023; Baldrati et al., 2023; Gu et al., 2024) either design light-weight modality transformation or adjust training and model to use existing image-caption data. CIReVL (Karthik et al., 2024) uses LLM and LMM on-the-fly for CIR, limiting its efficiency. Please refer to Appendix B for more details of these methods. In terms of constructing training data, CompoDiff (Gu et al., 2023) synthesizes 18M triplets with LLMs and image generative models, following the same pipeline with Brooks et al. (2023). The key difference between their data and ours lies in the image quality and the image relations. As shown in Figure 2, our data comes from natural image pairs found on the same web pages. Thus, our data covers open-ended image relations over a wide distribution, including both visual and beyond-visual ones.

Retrieval with Instruction. Instruction tuning (Ouyang et al., 2022; Lou et al., 2024) enables models with strong cross-domain and zero-shot generalization capabilities in retrieving both textual (Su et al., 2023; Asai et al., 2023) and multimodal content (Wei et al., 2023). However, prior efforts focus on unifying different retrieval tasks with manually-written instructions as task prefixes of actual queries, on a hundred-scale basis. In contrast, our approach utilizes million-scale instructions that naturally express user's search intents.

3. MagicLens

3.1. Data Construction for Self-Supervised Training

Web documents contain multimodal contexts, featuring interleaved texts and images on pertinent subjects. Image pairs extracted from the same web page through co-occurrence frequently imply associations between images and specific relations. This encompasses a broad spectrum of image relations, ranging from visual similarity to more nuanced connections (e.g., Figure 2). Consequently, these naturally occurring image pairs serve as excellent self-supervised training signals for image retrieval models. Based on this insight, we propose a systematic data construction pipeline to mine image pairs from web pages and adopt LLMs to generate open-ended instructions that explicitly convey the image relations within each pair.

Mining Image Pairs from Web Pages. *(1) Grouping & Cleaning.* We collect all images with the same URL from Common Crawl[2] as a group of images from the same web page for potential pairing. Due to the inevitable noisy images introduced by simple grouping, we remove duplicated, low-resolution, and advertising images, as well as highly overlapped groups. This results in a large number of groups with more densely and intrinsically connected images.

(2) Metadata Expansion. To provide detailed textual information of images for later LLMs with massive metadata expansion, we annotate images with Alt-texts[3], image content annotation (ICA) labels[4], and captions. We discard images if their Alt-texts are unqualified. For ICA labels, we annotate entities for each image, such as general objects and activities. For image captions, we adopt a SOTA LMM —

[2]https://commoncrawl.org/
[3]https://en.wikipedia.org/wiki/Alt_attribute
[4]https://cloud.google.com/vision/docs/labels

PaLI (Chen et al., 2023a) to generate captions. Each type of metadata provides textual information about images from different perspectives. See Appendix A for more details.

(3) Scoring & Filtering. After obtaining groups of images along with their extensive metadata, we pair up images within the same group and eliminate unqualified pairs using a combination of relevance measures. We use the CLIP image-to-image score to assess visual relevance and the text-to-text score for non-visual relevance. Image pairs that don't meet the criteria, such as those with low scores in both aspects, are excluded from consideration. To avoid the over-sampling of redundant images and duplicated relations, we set a maximum of three pairs for each group, thereby ensuring a more uniform distribution of images and relations in our training data (see Figure 5).

Open-Ended Instruction Generation. With the informative metadata of high-quality paired images, LLMs are able to well understand the image content (ICA and caption) and their background information (Alt-text). Using instruction (Chung et al., 2022), few-shot demonstrations (Brown et al., 2020), and chain-of-thought prompting (Wei et al., 2022) techniques, PaLM2 (Anil et al., 2023) generates open-ended instructions that precisely connect the paired images (*image*$_q$, *image*$_t$). Figure 3 illustrates generated instructions and Appendix 10 shows the detailed prompt and demonstrations. Eventually, we obtain 36.7M triplets (*image*$_q$, *text*, *image*$_t$) for self-supervised image retrieval training.

3.2. MagicLens Model

Model Design. As shown in Figure 4, we adopt a simple dual-encoder architecture with shared parameters and initialize the backbone vision and language encoders with CoCa (Yu et al., 2022) or CLIP (Radford et al., 2021). To enable deep modality integration, we introduce multiple self-attention layers and design a single multi-head attention pooler, compressing the multimodal inputs into a single embedding $\boldsymbol{r}$ for later matching. Additionally, since the retrieval target comprises only an image without accompanying text, we employ an empty text string "" to transform the target into a multimodal input. We denote $\boldsymbol{r_q}$ as the embedding for the multimodal query (*image*$_q$, *text*) and $\boldsymbol{r_t}$ as the embedding for the target (*image*$_t$, ""). Considering the efficiency, we propose MagicLens-B and MagicLens-L, initialized with the base and large checkpoints, respectively.

Model Training. We use a simple contrastive loss to train MagicLens. Our model is updated by contrasting the paired query-target against other targets in one training batch. In particular, as the query image itself (*image*$_q$) can be a challenging hard negative for the multimodal query (*image*$_q$, *text*), we combine the query image itself and an empty text to encode (*image*$_q$, "") to get $\boldsymbol{r_t^{'}}$ as an additional query negative example. To scale up the number of negative examples, for each query image, we use all query negatives and other target negatives in the same batch. Formally, for the i-th training example, the loss function $\mathcal{L}_i$ is defined as,

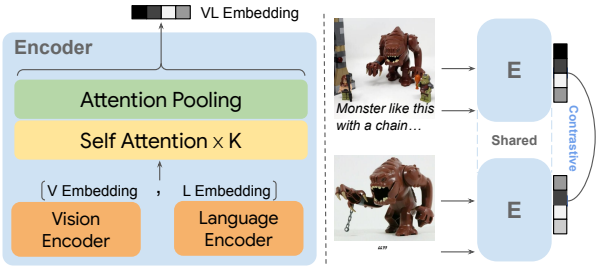

Figure 4. Model architecture and training of MagicLens Encoder (E), which takes the vision and language embeddings and feeds them as a sequence to self-attention layers for modality integration.

$$-\log \frac{e^{\text{sim}(\boldsymbol{r_q^i}, \boldsymbol{r_t^i})/\tau}}{\sum_{j=1}^{N}(e^{\text{sim}(\boldsymbol{r_q^i}, \boldsymbol{r_t^j})/\tau} + e^{\text{sim}(\boldsymbol{r_q^i}, \boldsymbol{r_t^{j'}})/\tau})},$$

where sim$(,)$ indicates a cosine similarity function $\frac{\boldsymbol{r_q}^T \boldsymbol{r_t}}{||\boldsymbol{r_q}|| \cdot ||\boldsymbol{r_t}||}$, N refers to the sampled batch size, and τ is a temperature hyperparameter for logit scaling. Please refer to Appendix A for more implementation details.

4. Experiments

4.1. Experiment Setup

Benchmarks and Metrics. To comprehensively evaluate MagicLens' multimodality-to-image retrieval ability, we consider three related tasks in a zero-shot, one-checkpoint setting: (1) composed image retrieval (CIR), (2) domain transfer retrieval, and (3) conditional image similarity. Each task has different yet limited sets of image relations. Table 2 shows the detailed statistics of five benchmarks.

Composed Image Retrieval. We consider one domain-specific and two open-domain benchmarks to evaluate the model's domain adaptability and its capability on real-world natural images, respectively. `FIQ` (Wu et al., 2021) is a fashion-domain benchmark with three disjoint retrieval sub-tasks: dress, shirt, and toptee. Following previous work (Saito et al., 2023; Baldrati et al., 2023; Gu et al., 2024), we evaluate on its validation set and report recall averaged over sub-tasks. `CIRR` (Liu et al., 2021) is the first dataset constructed on natural images (Suhr et al., 2019) with nine pre-defined relations between the query and the target images. It also includes a subset retrieval setting where models retrieve a target image from a dedicated small subset for each query. However, in addition to the limited size of the retrieval pool, it also suffers from false negative issues, as pointed out by Baldrati et al. (2023). We utilize recall (R and R$_s$) to evaluate standard retrieval and subset retrieval. In contrast, to better align with the real-world large-scale

Table 1. Performance comparison on five benchmarks of three multimodality-to-image retrieval tasks. The results of baselines are from the original papers. We mark the best results in bold and the second-best results underlined. ⋆CIReVL uses multiple model components including ChatGPT for retrieval, we report # parameters of components with known sizes. †PLI does not release code so we estimate.

Method	Backbone	# Total Params	Composed Image Retrieval				Domain Trans	Cond ImSim
			FIQ	CIRR		CIRCO	DTIN	GeneCIS
			R@10	R@1	R_s@1	mAP@5	R@10	R_s@1
PALAVRA (Cohen et al., 2022)	CLIP-B	176M	19.8	16.6	41.6	4.6	-	-
SEARLE (Baldrati et al., 2023)	CLIP-B	165M	22.9	24.0	54.9	9.4	-	-
CIReVL (Karthik et al., 2024)	CLIP-B	12.3B⋆	28.3	23.9	60.2	14.9	-	15.9
PLI (Chen & Lai, 2023)	BLIP-B	224M†	**35.9**	27.2	55.1	7.1	-	-
MagicLens-B	CLIP-B	166M	26.3	27.0	66.7	23.1	28.3	15.0
MagicLens-B	CoCa-B	267M	35.2	**31.6**	**69.3**	**30.8**	**46.8**	**17.4**
Pic2Word (Saito et al., 2023)	CLIP-L	429M	24.7	23.9	-	8.7	10.1	11.2
SEARLE (Baldrati et al., 2023)	CLIP-L	442M	25.6	24.2	53.8	11.7	-	12.3
Context-I2W (Tang et al., 2024)	CLIP-L	496M	27.8	25.6	-	-	12.9	-
CompoDiff (Gu et al., 2023)	CLIP-L	568M	36.0	18.2	57.4	12.6	-	14.9
CIReVL (Karthik et al., 2024)	CLIP-L	12.5B⋆	28.6	24.6	59.5	18.6	-	15.9
PLI (Chen & Lai, 2023)	CLIP-L	428M†	35.4	25.5	55.6	10.4	-	-
LinCIR (Gu et al., 2024)	CLIP-L	442M	26.3	25.0	57.1	12.6	-	12.2
MagicLens-L	CLIP-L	465M	30.7	30.1	68.1	29.6	41.5	16.3
MagicLens-L	CoCa-L	613M	**38.0**	**33.3**	**70.9**	**34.1**	**48.2**	**16.7**

Table 2. Statistics of five evaluation benchmarks. We average the number of queries over sub-tasks (e.g., FIQ), if available. The # Index represents the size of the retrieval pool that is shared amongst all queries and the # Subset Index is the average size of subsets, each of which is dedicated for a query.

	Open-Domain		# Query	# Index	# Subset Index
	Image?	Instr?			
FIQ	✗	✗	2,005	5,179	-
CIRR	✓	✗	4,148	2,316	8.3
CIRCO	✓	✗	800	123,403	-
DTIN	✓	✗	10,000	16,983	-
GeneCIS	✓	✗	2,008	-	13.8

retrieval, CIRCO annotates multiple ground truths for each query and has over 120K natural images (Lin et al., 2014) as the index set. Therefore, we regard CIRCO as our main benchmark. As each query has multiple targets, we adopt mean Average Precision (mAP) as the evaluation metric.

Domain Transfer Retrieval. Domain Transfer ImageNet (DTIN; Saito et al. (2023)) aims to retrieve an image from another domain with the same conceptual object shown in the query image. It is constructed from natural images in ImageNet (Deng et al., 2009) and images in other domains in ImageNet-R (Hendrycks et al., 2021). For example, given a domain keyword *"cartoon"* and a real horse image as a query, models are expected to retrieve a cartoon horse from the index set with images from multiple domains. It covers 4 domains, 10K objects, and over 16K images as the index set. Following prior works (Saito et al., 2023; Karthik et al., 2024), we report recall averaged over sub-tasks.

Conditional Image Similarity. GeneCIS (Vaze et al., 2023) is a keyword-conditioned image similarity measurement benchmark. It has four sub-tasks about changing or focusing the attribute or object in the given image. For each query image and keyword, models need to find the most similar images to the query image, conditioned on the given keyword, from a dedicated small subset with 13.8 images on average. For example, in the change-object sub-task with keyword *"car"* and an image, models need to find another image depicting a similar scene but with additional cars.

Baselines. We consider several baselines: (1) PALARVA (Cohen et al., 2022), (2) Pic2Word (Saito et al., 2023), (3) SEARLE (Baldrati et al., 2023), (4) ContextI2W (Tang et al., 2024), (5) LinCIR (Gu et al., 2024), (6) CIReVL (Karthik et al., 2024) (7) CompoDiff (Gu et al., 2023) and (8) PLI (Chen & Lai, 2023). Details of these methods are described in Appendix B.

4.2. Multimodality-to-Image Retrieval

Table 1 shows results over five benchmarks from three tasks, from which we have the following observations:

First, with the comparable model size, both CLIP- and CoCa-based MagicLens outperform previous state-of-the-art models across the four open-domain benchmarks by a large margin, especially CoCa-based MagicLens-L on the challenging CIRCO (mAP@5 from 12.6 to 34.1) and DTIN (R@10 from 12.9 to 48.2). This shows the strong capability of MagicLens. We leave full results on Appendix C and detailed parameter efficiency analysis on § 5.2.

Second, by comparing MagicLens-L to MagicLens-B, we find generally consistent performance improvements across five benchmarks. This demonstrates the constructed data is of high quality and can benefit larger models. Also, this observation shows the scalability thanks to the simple dual-encoder model architecture and contrastive loss.

Table 3. Zero-shot image-text retrieval results. Results are marked in bold if they are better than initialized checkpoints. ⋆CoCa we reproduced and used for MagicLens. †CoCa reported in the original paper.

Model	Flickr30K (1K test set)						MSCOCO (5K test set)					
	Image → Text			Text → Image			Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CoCa-B†	89.8	98.8	99.8	76.8	93.7	96.8	63.8	84.7	90.7	47.5	72.4	80.9
CoCa-B⋆	88.6	98.5	99.4	74.5	93.4	96.4	63.4	84.2	90.4	46.4	71.5	80.1
MagicLens-B	87.9	97.7	**99.5**	**76.2**	**93.7**	**96.5**	**64.8**	**85.5**	**91.2**	**48.9**	**73.9**	**82.5**
CoCa-L†	91.4	99.2	99.9	79.0	95.1	97.4	65.4	85.6	91.4	50.1	73.8	81.8
CoCa-L⋆	92.1	98.8	99.9	78.4	94.2	96.9	65.1	85.5	91.3	49.3	73.2	81.5
MagicLens-L	89.6	98.7	99.4	**79.7**	**95.0**	**97.4**	**67.7**	**87.6**	**92.7**	**53.1**	**77.4**	**84.9**

Table 4. Results on three image-to-image retrieval benchmarks. The results of baselines are from Lin et al. (2023) with separate checkpoints for each benchmark while MagicLens (ML) models are evaluated across benchmarks under one-checkpoint setting.

Method	TU-Berlin		Sketchy		QuickDraw	
	mAP	P@100	mAP@200	P@200	mAP	P@200
ViT (2021)	36.0	50.3	40.3	51.2	10.1	11.3
SOTA (2023)	56.9	63.7	52.5	62.4	14.5	21.6
ML-B-CLIP	45.9	57.9	49.3	60.6	10.1	14.0
ML-B-CoCa	61.7	72.1	70.5	77.2	13.9	19.9
ML-L-CLIP	62.9	73.1	68.2	75.8	15.1	20.4
ML-L-CoCa	**70.2**	**79.1**	**75.7**	**81.3**	**19.7**	**27.4**

4.3. Image-to-Image Retrieval

Although MagicLens models are trained for ($image_q$, *text*) → $image_t$ task format, they can naturally cover $image_q$ → $image_t$ tasks by providing a fixed text instruction for all $image_q$. As a case study, we consider zero-shot sketch based image retrieval (ZS-SBIR) task where models need to retrieve a natural image given a sketch of it. By simply using *"find a natural image of it"* for all query images, MagicLens can perform such a task.

Following the prior zero-shot SOTA methods (Liu et al., 2019; Lin et al., 2023) in this domain, we consider three benchmarks, namely TU-Berlin (Zhang et al., 2016), Sketchy (Yelamarthi et al., 2018), and QuickDraw (Dey et al., 2019). TU-Berlin has 30 classes, 2,400 sketch queries, and 27,989 natural images as index set; Sketchy consists of 21 classes unseen in ImageNet-1K and 12,694 queries over an index set with 12,553 natural images; QuickDraw has 30 classes, 92,291 queries, and a 54,146-sized index set. For each dataset, we report mAP and precision metrics used in the prior SOTA work (Lin et al., 2023).

Notably, unlike previous zero-shot methods that use separate checkpoints trained on each dataset and evaluated on the above holdout test set, we use the same checkpoints for evaluation on all benchmarks. Results are reported in Table 4 and we can find that our models outperform prior SOTA methods by a significant margin, despite our adherence to a single checkpoint setting. This demonstrates the strong generalization capability of MagicLens models and the diversity of tasks that they can cover.

4.4. Text-to-Image Retrieval

Since MagicLens models are built upon vision and language encoders, these backbone encoders after training can still be reused for *image* → *text* and *text* → *image* retrieval tasks. Therefore, we evaluate MagicLens' backbone encoders on Flickr30k (Plummer et al., 2015) and MSCOCO (Chen et al., 2015), using the same dataset splits and evaluation metrics as prior works (Radford et al., 2021; Yu et al., 2022).

Table 3 shows the comparison between the original encoders and the ones updated after MagicLens training. For *text* → *image* task, we can observe consistent and non-trivial improvements across all metrics on both datasets. For *image* → *text* task, we observe marginal drops. These observations show that our training recipe can enhance the backbone encoders for text-to-image retrieval. We can draw the same conclusion with CLIP, which is detailed in Table 17. The improvements might stem from the fact that our multimodality-to-image training task necessitates deep understanding of text instruction, thus improving backbone language encoders. These text-to-image results, along with the results on image-to-image and multimodality-to-image tasks, show that MagicLens can well handle various forms of image retrieval tasks, all with strong performances.

5. Analysis

5.1. Data Analysis

Comparison to Existing Training Data. Previous data construction efforts, including CompoDiff (Gu et al., 2023) and InstructPix2Pix (IP2P; Brooks et al. (2023)), use synthesized image pairs and essentially template-based instructions to train image retrieval models. Given the data availability and the fact that CompoDiff adopts a creation pipeline similar to IP2P, we use IP2P data as our baseline to explore the effects of different training data on downstream models. We compare a CoCa-based MagicLens-B model trained on all IP2P data (1M) with one trained on our downsampled, same-sized data, using the same training recipe. Table 5 shows that MagicLens + Ours achieves performance advantages over its variant trained with IP2P data (Mag-

Table 5. Performance comparison of CoCa-based MagicLens-B trained with 1M IP2P data against 1M our data. We report averaged R@10 & R@50 on FIQ and averaged R@1 & R_s@1 CIRR for comparisons with CompoDiff (Gu et al., 2023).

	FIQ (R_{AVG})	CIRR (R_{AVG})	CIRCO (mAP@5)	DTIN (R@10)	GeneCIS (R@1)
CompoDiff + IP2P	27.2	27.4	-	-	-
MagicLens + IP2P	29.8	33.7	13.6	30.2	14.5
MagicLens + Ours	**43.7**	**48.2**	**29.7**	**43.7**	**15.8**

(a) IP2P (b) Our training data

Figure 5. Word distributions of IP2P data and our data.

icLens + IP2P) on all five benchmarks. This proves that our data with natural images and template-free instructions can enable stronger image retrieval models. Please refer to Appendix C for detailed comparisons.

In addition, we compare these two models with IP2P-trained CompoDiff, which is a retrieval model designed for using synthesized images. Despite its specific design, MagicLens + IP2P still outperforms the CompoDiff + IP2P. Also, it achieves better results on CIRCO, DTIN, and GeneCIS than prior comparable-sized SOTA baselines. These show the advantage of our training recipe, as our model can achieve decent results even when trained on sub-optimal data.

To provide more insights, we visualize the words of instructions in IP2P and in our data separately in Figure 5. As we can see, IP2P data has a large number of instruction keywords like *"turn"* and *"make"* due to its template-based nature. Also, it has many coarse-grained keywords such as *"photograph"* and *"painting"*. In contrast, because of the controlled sampling from one web page described in § 3.1, our data has more diverse and equally distributed keywords, covering fine-grained labels like *"brand"*.

Data Scaling. To investigate the effect of our data scale on models, we train CoCa-based MagicLens-B on randomly sampled sets of 0.2M, 0.5M, 1M, 5M, 10M, 25M, and the entire 36.7M triplets. Results on five benchmarks and their average performance are illustrated in Figure 6. As the data size increases, MagicLens shows enhanced average performance, especially before the 10M mark. This indicates the effectiveness of scaling data.

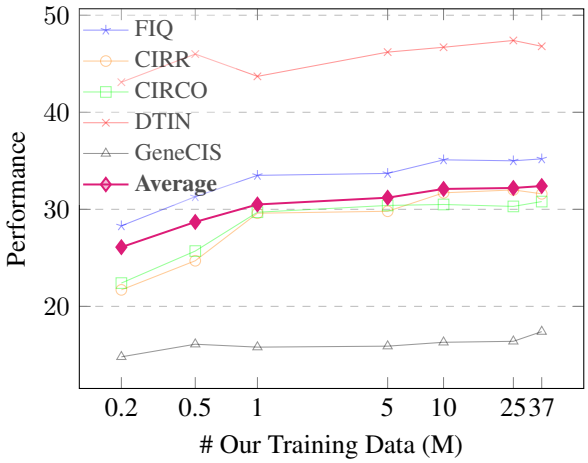

Figure 6. Performance of CoCa-based MagicLens-B when trained with different sizes of our data.

Table 6. Results of CoCa-based MagicLens-B trained with template-based and template-free instructions, at 1M scale.

Instruction	FIQ R@10	CIRR R@1	CIRCO mAP@5	DTIN R@10	GeneCIS R@1
Template-based	33.4	23.4	25.1	23.1	14.6
Template-free	**33.5**	**29.6**	**29.7**	**43.7**	**15.8**

Impacts of Instructions during Training. Instructions used in previous works (Brooks et al., 2023; Gu et al., 2023) are rooted from templates while our instructions are template-free. To investigate the effects of different instructions on downstream models, we also synthesize template-based instructions for naturally occurring image pairs collected in § 3.1. Specifically, due to the massive informative metadata of each image, we utilize LLMs to determine key metadata to fill pre-defined sentence structures. For our template-free instructions, LLMs are specifically guided to generate diverse and coherent instructions without adhering to any fixed template. We show concrete examples of different instructions in Figure 10 in the Appendix.

Table 6 compares the performance of two CoCa-based MagicLens-B models. Both of them are trained on 1M triplets, using the same image pairs but different instructions mentioned above. Template-free instructions clearly result in a stronger model, as evidenced by consistently better results on all benchmarks compared to the other model. This demonstrates that naturally expressed and diverse instructions can better stimulate the model to understand image relations and follow instructions.

5.2. Model Analysis

Model Size vs. Performance. Previous SOTA methods (Gu et al., 2024; 2023) consider using both larger vision and language encoders (Cherti et al., 2023) or using LMMs and LLMs on-the-fly (Karthik et al., 2024) for performance benefits. However, we argue that the model sizes and the

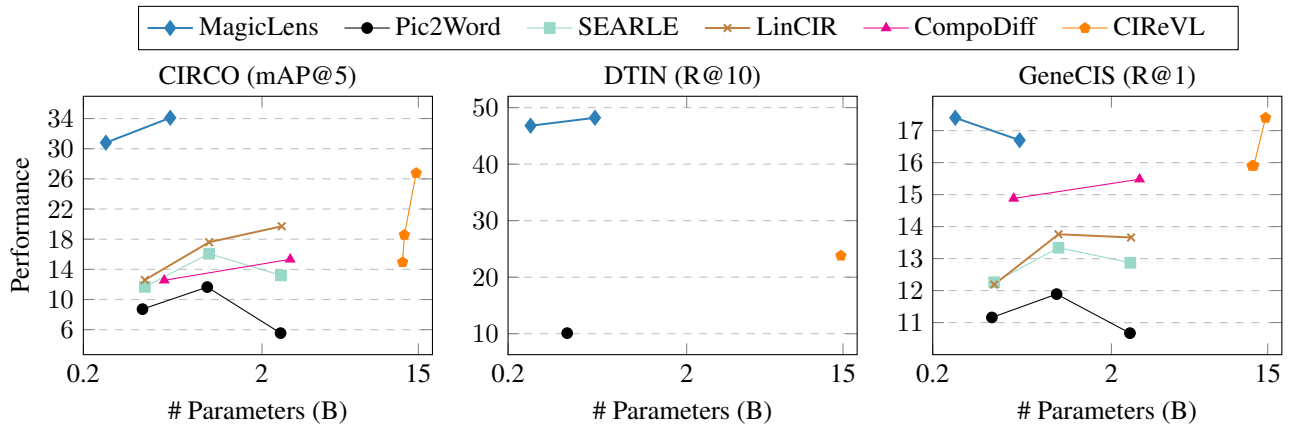

Figure 7. Model Size vs. Performance. MagicLens-B outperforms the SOTA CIReVL on three tasks even with 50× smaller # Parameters.

Table 7. Ablation study on CoCa-based MagicLens-B taking query images as negative samples during training (Qry Neg).

	FIQ R@10	CIRR R@1	CIRCO mAP@5	DTIN R@10	GeneCIS R@1
MagicLens	35.2	31.6	30.8	46.8	17.4
w/o Qry Neg	33.2	1.6	11.9	14.1	14.5

correlated efficiencies should also be taken into consideration for real-world deployments. In Figure 7, we visualize the relationship between model size and performance of various models on GeneCIS, CIRCO, and DTIN benchmarks. The results on GeneCIS and CIRCO are from Gu et al. (2024; 2023) and using CLIP-Large, OpenCLIP-Huge, and OpenCLIP-Giant backbones (Radford et al., 2021; Cherti et al., 2023). Results of CIReVL (Karthik et al., 2024) on DTIN are not fully reported by the authors. We omit the size of ChatGPT used and only count parameters of CIReVL's other model components (e.g., BLIP2-FLANT5-XXL + OpenCLIP-Giant).

Despite the 50× smaller size of CoCa-based MagicLens-B (267M) compared to other baselines (e.g., CIReVL with 14.6B), it achieves better performance on these benchmarks, with a significant advantage on the DTIN. This observation demonstrates the high parameter efficiency introduced by the parameter-sharing design in our model and the strong advantage of our data in enabling strong yet small models. Detailed results are in Table 14, 15, and 16 in Appendix.

Ablation on Contrastive Loss. Compared to standard contrastive loss, we introduce query images as hard negative examples during training. To investigate the impact of this design, we train CoCa-based MagicLens-B without these hard negatives and report the results in Table 7. As we can see, without query negatives, the performance of MagicLens drops across all benchmarks, significantly on the CIRR, CIRCO, and DTIN benchmarks. Also, we find in many cases, this model prefers to rank the query image itself higher than other images during retrieval, regardless of the given instructions. This indicates that differentiating closely similar images is crucial in improving the model's instruction understanding capabilities. Importantly, although using query negatives seems to limit MagicLens' ability to find the identical image, the first example in Figure 1 shows MagicLens can generalize to this instruction unseen during training and successfully retrieve the identical image.

Table 8. Results of MagicLens variants. CrossAttn indicates the model with cross-attention instead of self-attention for modality integration. FrozenEnc means the model with backbone vision and language encoders frozen during training.

	FIQ R@10	CIRR R@1	CIRCO mAP@5	DTIN R@10	GeneCIS R@1
MagicLens-B	**35.2**	**31.6**	**30.8**	**46.8**	**17.4**
w/ CrossAttn	31.0	28.3	27.0	41.4	16.2
w/ FrozenEnc	30.8	25.9	21.7	30.1	15.2
MagicLens-L	**38.0**	**33.3**	**34.1**	48.2	**16.7**
w/ CrossAttn	32.3	29.9	28.5	**52.5**	16.5
w/ FrozenEnc	32.5	26.5	23.0	29.4	15.5

Ablation on Model Architecture. We provide results of other model architectures we have explored in Table 8. In CrossAttn model arch, we explore various forms of cross attention, we report the best one which uses text embedding to attend concatenated image and text embeddings. However, even the best variant of this arch fails to reach the performance of self attention on most benchmarks.

We also explore the impact of freezing the backbone encoders initialized from CoCa (Yu et al., 2022) during training. The results of FrozenEnc are consistently worse than the fully-trained MagicLens. This proves that merely training additional layers on the top of single-modality encoders is not sufficient to deliver the strongest model.

5.3. Retrieval on 1.4M Open-Domain Image Corpus

To simulate image retrieval in a more realistic scenario, we hold out 1.4M unseen images as our index set, mak-

Table 9. One-on-one comparison (win rate) on a holdout index set with 1.4M images. Each setting has 50 queries with manually written instructions. The results are averaged over three evaluators.

Instruction Type	MagicLens-L Win	LinCIR Win	Tie
Simple Visual	**50.7**	41.3	8.0
Complex Visual	**61.3**	24.0	14.7
Beyond Visual	**80.0**	4.7	15.3

ing it the largest retrieval pool to date. We then collect 150 images and divide them into three disjoint groups with different types of manually written instructions: simple, complex, and beyond visual. Both simple and complex instructions are used in searching for visually similar images, but they differ in terms of complexity. Simple instructions describe only one visual difference (e.g. `same product with different color`) in the images given, whereas complex ones have multiple differences (e.g., `car` and `bag` examples in Figure 8). Beyond visual instructions aim to find images that share no visual similarities with the query images (e.g., `find other attractions`... in Figure 1).

Table 9 compares CoCa-based MagicLens-L with code-available previous-best model (LinCIR; Gu et al. (2024)), both with ViT-L backbones. For each query, one-on-one human evaluation is applied to the images retrieved by these two models to select the one that fully meets the instructions. If both or neither of the models succeed, the evaluators will mark them as a tie. We can observe LinCIR can handle simple instructions but suffers from complex instructions and almost completely fails on beyond visual instructions. In contrast, our method can satisfy diverse search intents expressed by all kinds of instructions, remarkably on the complex (61.3 vs. 24.0) and beyond visual (80 vs. 4.7) ones.

5.4. Qualitative Study

Figure 8 illustrates top-1 retrieval results on the holdout index set with 1.4M images. Even with complex instruction containing multiple conditions (`car` and `bag` examples), MagicLens is still able to accurately comprehend search intents and retrieve desired images. The `muffin` example showcases that MagicLens can understand the non-trivial temporal relation between images, thanks to the relation diversity introduced by naturally occurring image pairs. However, the image retrieved by MagicLens given the `3D anatomy` query may not be generally preferred since the instruction exemplifies the head. This suggests our model may return qualified yet imperfect examples when the instruction is not clearly expressed. Please refer to Figure 11 in Appendix D for more qualitative studies.

Figure 9 presents a visual case study on domain transfer retrieval using the DTIN benchmark. The text instruction presented in each domain is *"find this object in {domain}"*, where the same query image is used. All top-2 retrieved results are correct, highlighting the effectiveness of MagicLens in understanding conceptual image relations.

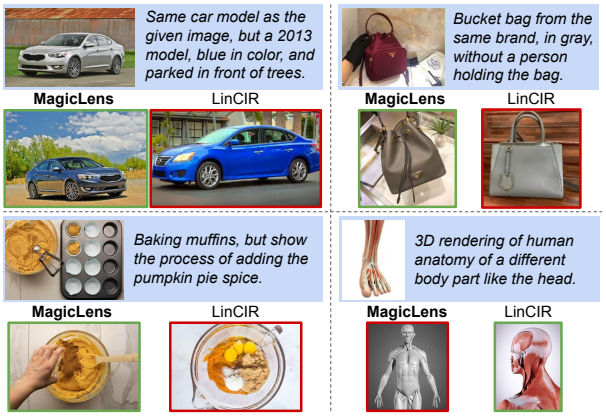

Figure 8. Top-1 retrieved images of CoCa-based MagicLens-L and LinCIR on the holdout index set with 1.4M images. Queries are with a blue background, while correct and incorrect retrieved images are marked with green and red outlines, respectively. LinCIR fails to retrieve correct results for `car`, `bag`, and `muffin` queries, even considering its top-5 results (see Figure 11 in Appendix).

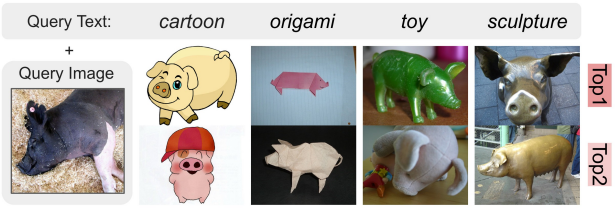

Figure 9. Top-2 retrieval results on four domains given the same query image on the DTIN benchmark.

6. Conclusion

We present MagicLens, a series of image retrieval models that follow open-ended text instructions. Despite being 50× smaller than prior SOTA methods, MagicLens achieves better results on multiple benchmarks including CIRCO, GeneCIS, and DTIN. Human evaluation on the 1.4M retrieval image pool shows that MagicLens can well satisfy diverse search intents expressed by open-ended instructions, especially complex and beyond visual ones. This indicates MagicLens' strong capability and potential for real-world search scenarios. Such retrieval models that support open-ended instructions can potentially benefit other vision-language tasks such as visual QA (Antol et al., 2015; Chen et al., 2023c), and enhance multimodal retrieval augmented models (Chen et al., 2022; Hu et al., 2023b). More importantly, we hope our recipe for constructing large-scale synthetic self-supervised training data can shed light on other research directions, such as multimodal retrieval, multimodal representation learning, and beyond.

Impact Statement

This work provides novel insights into self-supervised training by mining naturally occurring image pairs and develops image retrieval models that follow open-ended instructions to satisfy diverse search intents. It may enable a wide range of search scenarios and have potentials for real-world applications by providing users with more accurate search results. However, despite careful filtering out explicit and offensive images in the training data, MagicLens' ability to understand image relationships could still be misused for inappropriate image searches. Careful consideration and mitigation strategies are necessary to address these risks.

Acknowledgements

The authors would like to thank Jinhyuk Lee, William Cohen, Jonathan Berant, Kristina Toutanova, Boqing Gong, and other members from the Google DeepMind Seattle for their constructive feedback.

References

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., Ruder, S., Tay, Y., Xiao, K., Xu, Y., Zhang, Y., Abrego, G. H., Ahn, J., Austin, J., Barham, P., Botha, J., Bradbury, J., Brahma, S., Brooks, K., Catasta, M., Cheng, Y., Cherry, C., Choquette-Choo, C. A., Chowdhery, A., Crepy, C., Dave, S., Dehghani, M., Dev, S., Devlin, J., Díaz, M., Du, N., Dyer, E., Feinberg, V., Feng, F., Fienber, V., Freitag, M., Garcia, X., Gehrmann, S., Gonzalez, L., Gur-Ari, G., Hand, S., Hashemi, H., Hou, L., Howland, J., Hu, A., Hui, J., Hurwitz, J., Isard, M., Ittycheriah, A., Jagielski, M., Jia, W., Kenealy, K., Krikun, M., Kudugunta, S., Lan, C., Lee, K., Lee, B., Li, E., Li, M., Li, W., Li, Y., Li, J., Lim, H., Lin, H., Liu, Z., Liu, F., Maggioni, M., Mahendru, A., Maynez, J., Misra, V., Moussalem, M., Nado, Z., Nham, J., Ni, E., Nystrom, A., Parrish, A., Pellat, M., Polacek, M., Polozov, A., Pope, R., Qiao, S., Reif, E., Richter, B., Riley, P., Ros, A. C., Roy, A., Saeta, B., Samuel, R., Shelby, R., Slone, A., Smilkov, D., So, D. R., Sohn, D., Tokumine, S., Valter, D., Vasudevan, V., Vodrahalli, K., Wang, X., Wang, P., Wang, Z., Wang, T., Wieting, J., Wu, Y., Xu, K., Xu, Y., Xue, L., Yin, P., Yu, J., Zhang, Q., Zheng, S., Zheng, C., Zhou, W., Zhou, D., Petrov, S., and Wu, Y. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. VQA: visual question answering. In *Proceedings of ICCV*, 2015.

Asai, A., Schick, T., Lewis, P., Chen, X., Izacard, G., Riedel, S., Hajishirzi, H., and Yih, W.-t. Task-aware retrieval with instructions. In *Findings of ACL*, 2023.

Baldrati, A., Agnolucci, L., Bertini, M., and Del Bimbo, A. Zero-shot composed image retrieval with textual inversion. In *Proceedings of ICCV*, 2023.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of CVPR*, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Proceedings of NeurIPS*, 2020.

Chen, J. and Lai, H. Pretrain like your inference: Masked tuning improves zero-shot composed image retrieval. *arXiv preprint arXiv:2311.07622*, 2023.

Chen, J., Hu, H., Wu, H., Jiang, Y., and Wang, C. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of CVPR*, 2021.

Chen, W., Hu, H., Chen, X., Verga, P., and Cohen, W. W. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of EMNLP*, 2022.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., Ruiz, C. R., Goodman, S., Wang, X., Tay, Y., et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023a.

Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A. V., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B. K., Ruiz, C. R., Steiner, A. P., Angelova, A., Zhai, X., Houlsby, N., and Soricut, R. PaLI: A jointly-scaled multilingual language-image model. In *Proceedings of ICLR*, 2023b.

Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., and Chang, M. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of EMNLP*, 2023c.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of CVPR*, 2023.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Cohen, N., Gal, R., Meirom, E. A., Chechik, G., and Atzmon, Y. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *Proceedings of ECCV*, 2022.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 2008.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of CVPR*, 2009.

Dey, S., Riba, P., Dutta, A., Llados, J., and Song, Y.-Z. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proceedings of CVPR*, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR*, 2021.

Faghri, F., Fleet, D. J., Kiros, J. R., and Fidler, S. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

Gordo, A., Almazán, J., Revaud, J., and Larlus, D. Deep image retrieval: Learning global representations for image search. In *Proceedings of ECCV*, 2016.

Gu, G., Chun, S., Kim, W., Jun, H., Kang, Y., and Yun, S. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023.

Gu, G., Chun, S., Kim, W., , Kang, Y., and Yun, S. Language-only training of zero-shot composed image retrieval. In *Proceedings of CVPR*, 2024.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of ICCV*, 2021.

Hu, H., Luan, Y., Chen, Y., Khandelwal, U., Joshi, M., Lee, K., Toutanova, K., and Chang, M.-W. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of CVPR*, 2023a.

Hu, Z., Iscen, A., Sun, C., Wang, Z., Chang, K.-W., Sun, Y., Schmid, C., Ross, D. A., and Fathi, A. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of CVPR*, 2023b.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of ICML*, 2021.

Karthik, S., Roth, K., Mancini, M., and Akata, Z. Vision-by-language for training-free compositional image retrieval. In *Proceedings of ICLR*, 2024.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of ICML*, 2021.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. In *Proceedings of NeurIPS*, 2021.

Li, J., Li, D., Xiong, C., and Hoi, S. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of ICML*, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of ICML*, 2023.

Lin, F., Li, M., Li, D., Hospedales, T., Song, Y.-Z., and Qi, Y. Zero-shot everything sketch-based image retrieval, and in explainable style. In *Proceedings of CVPR*, 2023.

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. In *Proceedings of ECCV*, 2014.

Liu, Q., Xie, L., Wang, H., and Yuille, A. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *Proceedings of ICCV*, 2019.

Liu, Z., Rodriguez-Opazo, C., Teney, D., and Gould, S. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of ICCV*, 2021.

Lou, R., Zhang, K., and Yin, W. A comprehensive survey on instruction following. *arXiv preprint arXiv:2303.10475*, 2024.

OpenAI. Chatgpt, 2022. URL https://openai.com/blog/chatgpt.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Proceedings of NeurIPS*, 2022.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of ICCV*, 2015.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, 2021.

Saito, K., Sohn, K., Zhang, X., Li, C.-L., Lee, C.-Y., Saenko, K., and Pfister, T. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of CVPR*, 2023.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C. W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S. R., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Proceedings of NeurIPS*, 2022.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

Shazeer, N. and Stern, M. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of ICML*, 2018.

Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.-t., Smith, N. A., Zettlemoyer, L., and Yu, T. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of ACL*, 2023.

Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., and Artzi, Y. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of ACL*, 2019.

Tang, Y., Yu, J., Gai, K., Zhuang, J., Xiong, G., Hu, Y., and Wu, Q. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *Proceedings of AAAI*, 2024.

Vaze, S., Carion, N., and Misra, I. Genecis: A benchmark for general conditional image similarity. In *Proceedings of CVPR*, 2023.

Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.-J., Fei-Fei, L., and Hays, J. Composing text and image for image retrieval - an empirical odyssey. In *Proceedings of CVPR*, 2019.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., and Wei, F. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of CVPR*, 2023.

Wei, C., Chen, Y., Chen, H., Hu, H., Zhang, G., Fu, J., Ritter, A., and Chen, W. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of NeurIPS*, 2022.

Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., and Feris, R. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of CVPR*, 2021.

Yelamarthi, S. K., Reddy, S. K., Mishra, A., and Mittal, A. A zero-shot framework for sketch based image retrieval. In *Proceedings of ECCV*, 2018.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of CVPR*, 2022.

Zhang, H., Liu, S., Zhang, C., Ren, W., Wang, R., and Cao, X. Sketchnet: Sketch classification with web images. In *Proceedings of CVPR*, 2016.

Overview of Appendix

Our supplementary includes the following sections:

- **Section A: Implementation Details.**
- **Section B: Baselines.**
- **Section C: Full Results.**
- **Section D: More Qualitative Study.**

A. Implementation Details

Image Cleaning and Pairing. We use the Common Crawl and group images with identical URLs, considering them as images from the same websites. We regard two images as identical if their CLIP image embedding scores exceed 0.98 and remove them. If two groups share a high ratio of duplicated images (80%), we randomly remove one of those groups. The minimum resolution remained is 288x288, which matches the input size of CoCa models we used. For the concrete thresholds used for filtering, we have set 0.82 as the threshold for CLIP image-to-image similarity and 0.9 for text-to-text similarity over captions. Additionally, to ensure the uniqueness of the images, the target image must have a distinct ICA label with high text-image similarity to itself (0.32) and low similarity to the query image (0.18). Only image pairs that meet these requirements will be remained for the instruction generation stage.

Instruction Generation. We provide LLMs with massive metadata expansion including Alt-texts, image content annotation (ICA) labels, and image captions by using various tools and LMMs. Specifically, similar to Sharma et al. (2018), we analyze candidate Alt-texts with part-of-speech, sentiment, and pornography annotations of Google Natural Language APIs. We discard images if their Alt-texts only have rare tokens or if they are triggered by sentiment/pornography detectors For ICA labels, we utilize Google Vision APIs to annotate entities for each image such as general objects, locations, and activities. On average, each image has 25.2 fine-grained ICA labels. Also, we provide the instruction and two detailed demonstrations for instruction generation in Table 10.

Model. With the proposed data construction pipeline, we eventually collect $36,714,118$ triplets for pre-training. For model architecture, we design 4 randomly initialized self-attention layers on the top of vision and language encoders. Further, we utilize one attention pooling layer (Yu et al., 2022) for the final embedding. Following Jia et al. (2021); Yu et al. (2022), during the training of CoCa-based MagicLens, we set image resolution of 288×288 and patch size 18×18. For CLIP-based MagicLens, we set image resolution of 224×224 and use ViT-B16 and ViT-L14. For both CLIP and CoCa, we use contrastive image embedding and contrastive text embedding, which will be concatenated as a sequence with a fixed length of 2 in self-attention layers. The number of newly added self-attention layers is 4 and the τ is learnable and initialized with 0.07. We set the batch size as 2048 and trained our models for a maximum of 50,000 steps with Adafactor (Shazeer & Stern, 2018) and an early stopping mechanism. The learning rates are set differently for newly-introduced parameters and re-used CLIP or CoCa parameters, at 2e-5 and 2e-6, respectively. We train our base and large models on 64 and 128 TPUs, respectively. The training process lasts six hours for both models and the best checkpoints are selected based on the performance on the validation set of CIRR and CIRCO.

B. Baselines

We consider various baselines and detail them as follows: **(1) PALARVA** (Cohen et al., 2022) **(2) Pic2Word** (Saito et al., 2023), **(3) SEARLE** (Baldrati et al., 2023), **(4) ContextI2W** (Tang et al., 2024), and **(5) LinCIR** (Gu et al., 2024) train an additional mapping network to encode the given reference image as a pseudo word token. Then, it can be combined with the actual query text for text-to-image retrieval. These methods rely on image-caption pairs for mapping network training Further, LinCIR introduces text-only data for better mapping capability. **(6) CIReVL** (Karthik et al., 2024) is a training-free method by using BLIP-2 with FLANT5-XXL (Li et al., 2023) for query image caption generation, ChatGPT (OpenAI, 2022) for target image caption generation, and CLIP (Radford et al., 2021; Cherti et al., 2023) for the final text-to-image retrieval. Such a complex retrieval pipeline may limit their inference speed and potential practicalness in real-world scenarios. Inspired by diffusion models, **(7) CompoDiff** (Gu et al., 2023) regards query text as a condition to guide the image embedding generation and train the model with 18M synthesized data. **(8) PLI** (Chen & Lai, 2023) corrupts the image in image-caption data and regards the original image as a target to simulate the CIR task during the pre-training stage.

Table 10. Detailed prompt for query generation using PaLM2 (Anil et al., 2023).

Instruction	Based on the provided ALT_TEXT, TEXT_LABEL, and CAPTION of two different images, create an interesting text query which can be used with the source image to retrieve the target image. Note that the TEXT_LABEL and CAPTION are generated by models so they may not be 100% correct, especially when it's about very specific entities (e.g., a specific car type in some year), so selectively use the most likely correct information and generate the query. This query should include: 1) one general and unspecific similarity (same brand, similar toy, similar building, etc). 2) all differences that only the target image has. Remember the query should be concise, short, and not be able to directly retrieve the target image. The retrieval has to be done by combining the source image and text query.
Demonstrations	Both images are from the website [HOME - 1-of-1 Automotive Artworks] **Source Image**: ALT_TEXT [custom porsche cayman gt4 illustration framed]. TEXT_LABEL [Licence plate]. CAPTION [a drawing of a porsche gt4 rs coupe in a frame]. **Target Image**: ALT_TEXT [custom illustration of a 1972 porsche 911 blue]. TEXT_LABEL [Licence plate, Turquoise]. CAPTION [a framed print of a blue porsche 911 s all coupe]. **Think**: Both images are custom illustrations of Porsche cars as described in the alt_text. The source image is a Porsche Cayman GT4 while the target image is a 1972 Porsche 911 in blue. Therefore, the query should focus on the type of image (custom illustration of a Porsche car), but specify the different model and year (1972 Porsche 911) and color (blue). **Query**: [Porsche 911 in blue shown in the same illustrative way.] Both Images are from the website [Rapunzel Worksheet — Printable Worksheets and Activities for Teachers, Parents, Tutors and Homeschool Families] **Source Image**: ALT_TEXT [tangled rapunzel color pages printable]. TEXT_LABEL [Coloring book]. CAPTION [rapunzel in a boat with lanterns floating in the air coloring page]. **Target Image**: ALT_TEXT [cool rapunzel and flynn flower hair coloring page]. TEXT_LABEL [Coloring book, Floral design]. CAPTION [a black and white drawing of rapunzel tangled with long hair and flowers in her dress]. **Think**: Both images are coloring pages featuring Rapunzel as described in the alt_text. The source image shows Rapunzel in a boat with lanterns, while the target image shows Rapunzel with Flynn and flowers in her hair. Therefore, the query should focus on the type of image (Rapunzel coloring page), but specify the different scene (Rapunzel with Flynn and flowers in her hair, not in a boat, no lanterns). **Query**: [Same coloring page about Rapunzel but no boat or lantern, with more clear flowers in the character's hair] *. . . (Three more few-shot demonstrations)*

Table 11. Detailed performance of CoCa-based MagicLens-B trained with IP2P data and our constructed data, on the same 1M scale.

	FIQ	CIRR	CIRR	CIRCO	DTIN	GeneCIS
	R@10	R@1	R_s@1	mAP@5	R@10	R@1
MagicLens + IP2P (1M)	20.3	12.5	54.9	13.6	30.2	14.5
MagicLens + Ours (1M)	**33.5**	**29.6**	**66.9**	**29.7**	**43.7**	**15.8**

C. Full Results

C.1. Results on Five Multimodality-to-Image Benchmarks

Table 12, 13, and 14 show the full results on three CIR benchmarks (Wu et al., 2021; Liu et al., 2021; Baldrati et al., 2023). We report the performances of various models on `DT` and `GeneCIS` on Table 15 and Table 16, respectively. Some prior methods may use larger encoders (Gu et al., 2023; 2024) and develop a retrieval pipeline (Karthik et al., 2024) including LLMs (OpenAI, 2022) and LMMs (Li et al., 2023) for performance gains. Despite this, their results are still worse than that of MagicLens, supporting the parameter efficiency claimed in Figure 7.

C.2. Data Training Comparison

Table 11 compares CoCa-based MagicLens-B trained on IP2P data and on our data in detail, both at a 1M scale.

C.3. Text-to-Image Retrieval with CLIP-based MagicLens

We list the text-to-image retrieval results in Table 17 with the original CLIP and updated backbone CLIP encoders of MagicLens. The text-to-image retrieval performance is significantly boosted on both base and large models where image-to-text retrieval ability marginally drops. This aligns with the conclusion we draw on CoCa in § 4.4.

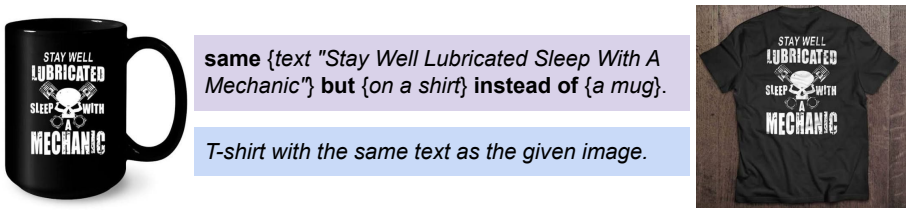

Figure 10. Examples of template-based and template-free instructions for the same image pair.

C.4. Examples of Template-based Instructions

We provide a concrete example of different instructions on the same image pair in Figure 10.

D. More Qualitative Study

We present detailed top-5 retrieval results of CoCa-based MagicLens-L and the code-available SOTA LinCIR (Gu et al., 2024) in Figure 11. 1) For the `bag` query, MagicLens can retrieve bags (the third and fourth images) from the same brand, even though they don't have shared visual clues (brand logo) with the query image. 2) Given the `house and gavel` query, our model successfully finds an interesting real-world scene and the perfect example in the top-2 results, but LinCIR fails to satisfy the query. This may stem from the limited representation abilities of a single pseudo token for an image with multiple objects. 3) The success on the `gazebo` example shows that MagicLens can understand simple numerical relations.

Table 12. Full results on the FIQ benchmark (Wu et al., 2021). CLIP-H and CLIP-G are OpenCLIP (Cherti et al., 2023) checkpoints. ⋆CIReVL uses multiple model components, we omit ChatGPT and report # parameters of other components (e.g., BLIP2-FLANT5-XXL + CLIP-G). †PLI does not release code so we estimate.

Method	# Total Params	Backbone Network	Dress		Shirt		Toptee		Overall	
			R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
PALAVRA (Cohen et al., 2022)	176M	CLIP-B	17.3	35.9	21.5	37.1	20.6	38.8	19.8	37.3
SEARLE (Baldrati et al., 2023)	165M	CLIP-B	18.5	39.5	24.4	41.6	25.7	46.5	22.9	42.5
CIReVL (Karthik et al., 2024)	12.3B⋆	CLIP-B	25.3	46.4	28.4	47.8	31.2	53.9	28.3	49.4
PLI (Chen & Lai, 2023)	224M	BLIP-B	28.6	50.8	38.1	57.8	40.9	62.7	**35.9**	**57.1**
MagicLens-B	166M	CLIP-B	21.5	41.3	27.3	48.8	30.2	52.3	26.3	47.4
MagicLens-B	267M	CoCa-B	29.0	48.9	36.5	55.5	40.2	61.9	35.2	55.4
Pic2Word (Saito et al., 2023)	429M	CLIP-L	20.0	40.2	26.2	43.6	27.9	47.4	24.7	43.7
SEARLE (Baldrati et al., 2023)	442M	CLIP-L	20.5	43.1	26.9	45.6	29.3	50.0	25.6	46.2
Context-I2W (Tang et al., 2024)	496M	CLIP-L	23.1	45.3	29.7	48.6	30.6	52.9	27.8	48.9
CompoDiff (Gu et al., 2023)	568M	CLIP-L	32.2	46.3	37.7	49.1	38.1	50.6	36.0	48.6
CIReVL (Karthik et al., 2024)	12.5B⋆	CLIP-L	24.8	44.8	29.5	47.4	31.4	53.7	28.6	48.6
PLI (Chen & Lai, 2023)	428M†	CLIP-L	28.1	51.1	38.6	58.5	39.4	62.7	35.4	57.4
LinCIR (Gu et al., 2024)	442M	CLIP-L	20.9	42.4	29.1	46.8	28.8	50.2	26.3	46.5
MagicLens-L	465M	CLIP-L	25.5	46.1	32.7	53.8	34.0	57.7	30.7	52.5
MagicLens-L	613M	CoCa-L	32.3	52.7	40.5	59.2	41.4	63.0	**38.0**	**58.2**
Pic2Word (Saito et al., 2023)	987M	CLIP-H	28.0	51.5	36.9	56.0	40.2	62.0	35.0	56.5
SEARLE (Baldrati et al., 2023)	1.0B	CLIP-H	28.5	51.1	36.5	55.5	38.8	60.9	34.6	55.8
LinCIR (Gu et al., 2024)	1.0B	CLIP-H	29.8	52.1	36.9	57.8	42.1	62.5	36.3	57.5
Pic2Word (Saito et al., 2023)	2.5B	CLIP-G	25.4	47.7	33.2	50.4	35.2	57.6	31.3	51.9
SEARLE (Baldrati et al., 2023)	2.6B	CLIP-G	28.2	50.3	36.5	55.4	39.8	61.5	34.8	55.7
CompoDiff (Gu et al., 2023)	2.9B	CLIP-G	37.8	49.1	41.3	55.2	44.3	56.4	39.0	51.7
CIReVL (Karthik et al., 2024)	14.6B⋆	CLIP-G	27.1	49.5	33.7	51.4	35.8	56.1	32.2	52.4
LinCIR (Gu et al., 2024)	2.6B	CLIP-G	38.1	60.9	46.8	65.1	50.5	71.1	**45.1**	**65.7**

Table 13. Full results on the CIRR benchmark (Liu et al., 2021). CLIP-H and CLIP-G are OpenCLIP (Cherti et al., 2023) checkpoints. ⋆CIReVL uses multiple model components, we omit ChatGPT and report # parameters of other components (e.g., BLIP2-FLANT5-XXL + CLIP-G). †PLI does not release code so we estimate.

Method	# Total Params	Backbone Network	Index Set				Subset		
			R@1	R@5	R@10	R@50	R@1	R@2	R@3
PALAVRA (Cohen et al., 2022)	176M	CLIP-B	16.6	43.5	58.5	84.0	41.6	65.3	80.9
SEARLE (Baldrati et al., 2023)	165M	CLIP-B	24.0	53.4	66.8	89.8	54.9	76.6	88.2
CIReVL (Karthik et al., 2024)	12.3B⋆	CLIP-B	23.9	52.5	66.0	87.0	60.2	80.1	90.2
PLI (Chen & Lai, 2023)	224M†	BLIP-B	27.2	58.9	71.4	91.3	55.1	77.4	89.1
MagicLens-B	166M	CLIP-B	27.0	58.0	70.9	91.1	66.7	83.9	92.4
MagicLens-B	267M	CoCa-B	**31.6**	**64.0**	**76.9**	**93.8**	**69.3**	**86.0**	**94.0**
Pic2Word (Saito et al., 2023)	429M	CLIP-L	23.9	51.7	65.3	87.8	-	-	-
SEARLE (Baldrati et al., 2023)	442M	CLIP-L	24.2	52.5	66.3	88.8	53.8	75.0	88.2
Context-I2W (Tang et al., 2024)	496M	CLIP-L	25.6	55.1	68.5	89.8	-	-	-
CompoDiff (Gu et al., 2023)	568M	CLIP-L	18.2	53.1	70.8	90.3	57.4	77.1	87.9
CIReVL (Karthik et al., 2024)	12.5B⋆	CLIP-L	24.6	52.3	64.9	86.3	59.5	79.9	89.7
PLI (Chen & Lai, 2023)	428M†	CLIP-L	25.5	54.6	67.6	88.7	55.6	77.5	89.5
LinCIR (Gu et al., 2024)	442M	CLIP-L	25.0	53.3	66.7	-	57.1	77.4	88.9
MagicLens-L	465M	CLIP-L	30.1	61.7	74.4	92.6	68.1	84.8	93.2
MagicLens-L	613M	CoCa-L	**33.3**	**67.0**	**77.9**	**94.4**	**70.9**	**87.3**	**94.5**
Pic2Word (Saito et al., 2023)	987M	CLIP-H	32.9	63.1	73.9	-	62.2	81.4	91.2
SEARLE (Baldrati et al., 2023)	1.0B	CLIP-H	34.0	64.0	75.3	-	64.6	83.2	92.8
LinCIR (Gu et al., 2024)	1.0B	CLIP-H	33.8	63.5	73.4	-	62.4	81.5	92.1
Pic2Word (Saito et al., 2023)	2.5B	CLIP-G	30.4	58.1	69.2	-	68.9	85.5	93.0
SEARLE (Baldrati et al., 2023)	2.6B	CLIP-G	34.8	64.1	75.1	-	68.7	84.7	93.2
CompoDiff (Gu et al., 2023)	2.9B	CLIP-G	26.7	55.1	74.5	92.0	64.5	82.4	91.8
CIReVL (Karthik et al., 2024)	14.6B⋆	CLIP-G	34.7	64.3	75.1	91.7	68.0	84.9	93.2
LinCIR (Gu et al., 2024)	2.6B	CLIP-G	**35.3**	64.7	76.1	-	63.4	82.2	92.0

Table 14. Full results on the CIRCO benchmark (Baldrati et al., 2023). CLIP-H and CLIP-G are OpenCLIP (Cherti et al., 2023) checkpoints. ⋆CIReVL uses multiple model components, we omit ChatGPT and report # parameters of other components (e.g., BLIP2-FLANT5-XXL + CLIP-G). †PLI does not release code so we estimate.

Method	# Total Params	Backbone Network	mAP@5	mAP@10	mAP@25	mAP@50
PALAVRA (Cohen et al., 2022)	176M	CLIP-B	4.6	5.3	6.3	6.8
SEARLE (Baldrati et al., 2023)	165M	CLIP-B	9.4	9.9	11.1	11.8
CIReVL (Karthik et al., 2024)	12.3B⋆	CLIP-B	14.9	15.4	17.0	17.8
PLI (Chen & Lai, 2023)	224M†	BLIP-B	7.1	8.0	9.2	9.7
MagicLens-B	166M	CLIP-B	23.1	23.8	25.8	26.7
MagicLens-B	267M	CoCa-B	**30.8**	**32.0**	**34.5**	**35.6**
Pic2Word (Saito et al., 2023)	429M	CLIP-L	8.7	9.5	10.6	11.3
SEARLE (Baldrati et al., 2023)	442M	CLIP-L	11.7	12.7	14.3	15.1
CompoDiff (Gu et al., 2023)	568M	CLIP-L	12.6	13.4	15.8	16.4
CIReVL (Karthik et al., 2024)	12.5B⋆	CLIP-L	18.6	19.0	20.9	21.8
PLI (Chen & Lai, 2023)	428M†	CLIP-L	10.4	11.6	13.0	13.7
LinCIR (Gu et al., 2024)	442M	CLIP-L	12.6	13.6	15.0	15.9
MagicLens-L	465M	CLIP-L	29.6	30.8	33.4	34.4
MagicLens-L	613M	CoCa-L	**34.1**	**35.4**	**38.1**	**39.2**
Pic2Word (Saito et al., 2023)	987M	CLIP-H	11.7	12.3	13.7	14.4
SEARLE (Baldrati et al., 2023)	1.0B	CLIP-H	16.1	16.9	18.8	19.7
LinCIR (Gu et al., 2024)	1.0B	CLIP-H	17.6	18.5	20.5	21.4
Pic2Word (Saito et al., 2023)	2.5B	CLIP-G	5.5	5.6	6.7	7.1
SEARLE (Baldrati et al., 2023)	2.6B	CLIP-G	13.2	13.9	15.3	16.0
CompoDiff (Gu et al., 2023)	2.9B	CLIP-G	15.3	17.7	19.4	-
CIReVL (Karthik et al., 2024)	14.6B⋆	CLIP-G	26.8	27.6	30.0	31.0
LinCIR (Gu et al., 2024)	2.6B	CLIP-G	19.7	21.0	23.1	24.2

Table 15. Full results on the DTIN benchmark (Saito et al., 2023). CLIP-G is a OpenCLIP (Cherti et al., 2023) checkpoint. ⋆CIReVL uses multiple model components, we omit ChatGPT and report # parameters of other components (e.g., BLIP2-FLANT5-XXL + CLIP-G).

Method	# Total Params	Backbone Network	Cartoon		Origami		Toy		Sculpture		Overall	
			R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Image-only (Saito et al., 2023)	304M	CLIP-L	0.3	4.5	0.2	1.8	0.6	5.7	0.3	4.0	0.4	4.0
Text-only (Saito et al., 2023)	124M	CLIP-L	0.2	1.1	0.8	3.7	0.8	2.4	0.4	2.0	0.5	2.3
Image+Text (Saito et al., 2023)	428M	CLIP-L	2.2	13.3	2.0	10.3	1.2	9.7	1.6	11.6	1.7	11.2
Pic2Word (Saito et al., 2023)	429M	CLIP-L	8.0	21.9	13.5	25.6	8.7	21.6	10.0	23.8	10.1	23.2
Context-I2W (Tang et al., 2024)	496M	CLIP-L	10.2	26.1	17.5	28.7	11.6	27.4	12.1	28.2	12.9	27.6
CIReVL (Karthik et al., 2024)	14.6B⋆	CLIP-G	19.2	42.8	30.2	41.3	22.2	43.1	23.4	45.0	23.8	43.0
MagicLens-B	166M	CLIP-B	49.4	67.0	13.8	26.3	25.8	43.4	24.3	41.3	28.3	44.5
MagicLens-B	267M	CoCa-B	65.8	73.3	29.3	38.6	46.7	57.7	45.3	57.1	46.8	56.7
MagicLens-L	465M	CLIP-L	62.6	72.2	21.5	33.4	43.8	58.4	38.0	54.2	41.5	54.5
MagicLens-L	613M	CoCa-L	60.1	69.6	36.0	44.7	45.2	56.9	51.4	59.6	**48.2**	**57.7**

Table 16. Full results on the GeneCIS benchmark (Vaze et al., 2023). ⋆CIReVL uses multiple model components, we omit ChatGPT and report # parameters of other components (e.g., BLIP2-FLANT5-XXL + CLIP-G).

Method	# Params	Backbone Network	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
CIReVL (2024)	12.3B⋆	CLIP-B	17.9	29.4	40.4	14.8	25.8	35.8	14.6	24.3	33.3	16.1	27.8	37.6	15.9
MagicLens-B	166M	CLIP-B	15.5	28.4	39.1	12.3	23.0	32.1	14.4	26.2	35.5	17.7	28.4	39.2	15.0
MagicLens-B	267M	CoCa-B	16.2	27.8	38.6	16.2	27.2	36.6	17.1	27.7	38.2	20.2	32.2	42.9	**17.4**
Pic2Word (2023)	429M	CLIP-L	15.7	28.2	38.7	13.9	24.7	33.1	8.4	18.0	25.8	6.7	15.1	24.0	11.2
SEARLE (2023)	442M	CLIP-L	17.0	29.7	40.7	16.4	25.3	34.1	8.0	16.9	25.6	7.9	16.8	24.8	12.3
CompoDiff (2023)	568M	CLIP-L	13.5	24.3	36.1	19.2	28.6	37.2	8.1	16.4	25.1	18.7	31.7	40.6	14.9
CIReVL (2024)	12.5B⋆	CLIP-L	19.5	31.8	42.0	14.4	26.0	35.2	12.3	21.8	30.5	17.2	28.9	37.6	15.9
LinCIR (2024)	442M	CLIP-L	16.9	30.0	41.5	16.2	28.0	36.8	8.3	17.4	26.2	7.4	15.7	25.0	12.2
MagicLens-L	465M	CLIP-L	16.1	28.2	39.0	15.6	27.5	36.3	16.3	26.2	35.5	17.1	29.5	39.7	16.3
MagicLens-L	613M	CoCa-L	16.6	28.7	39.3	16.0	27.5	36.5	15.7	27.6	37.3	18.7	31.7	40.2	**16.7**
Pic2Word (2023)	987M	CLIP-H	18.6	30.7	42.1	13.2	23.9	33.1	9.2	17.6	27.1	6.6	16.5	25.4	11.9
SEARLE (2023)	1.0B	CLIP-H	18.8	31.5	42.3	15.5	26.9	35.9	10.6	18.7	26.5	8.5	17.9	26.2	13.3
LinCIR (2024)	1.0B	CLIP-H	19.6	31.5	41.6	16.6	27.6	37.5	9.8	18.8	27.9	9.0	17.6	25.7	13.8
Pic2Word (2023)	2.5B	CLIP-G	12.5	23.4	33.7	11.7	21.9	30.9	9.9	19.3	27.4	8.6	18.2	26.1	10.7
SEARLE (2023)	2.6B	CLIP-G	16.3	29.4	40.7	16.2	27.3	35.5	10.8	18.2	27.9	8.3	15.6	25.8	12.9
CompoDiff (2023)	2.9B	CLIP-G	14.3	26.7	38.4	19.7	28.8	37.4	9.2	19.1	25.8	18.7	31.7	40.2	15.5
CIReVL (2024)	14.6B⋆	CLIP-G	20.5	34.0	44.5	16.1	28.6	39.4	14.7	25.2	33.0	18.1	31.2	41.0	**17.4**
LinCIR (2024)	2.6B	CLIP-G	19.1	33.0	42.3	17.6	30.2	38.1	10.1	19.1	28.1	7.9	16.3	25.7	13.7

Table 17. Zero-shot image-text retrieval results. Results are marked in bold if they are better than initialized checkpoints. ⋆CLIP and CoCa we reproduced and used for MagicLens. †CLIP and CoCa reported in the original paper.

Model	Flickr30K (1K test set)						MSCOCO (5K test set)					
	Image → Text			Text → Image			Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-B⋆	81.7	97.1	98.5	61.6	85.6	91.2	51.9	76.3	83.9	32.1	56.7	67.6
MagicLens-B	78.9	94.9	97.5	**67.8**	**88.8**	**93.4**	49.5	74.5	82.5	**40.1**	**65.4**	**75.1**
CoCa-B†	89.8	98.8	99.8	76.8	93.7	96.8	63.8	84.7	90.7	47.5	72.4	80.9
CoCa-B⋆	88.6	98.5	99.4	74.5	93.4	96.4	63.4	84.2	90.4	46.4	71.5	80.1
MagicLens-B	87.9	97.7	**99.5**	**76.2**	**93.7**	**96.5**	**64.8**	**85.5**	**91.2**	**48.9**	**73.9**	**82.5**
CLIP-L†	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	64.2	72.2
CLIP-L⋆	84.6	97.9	99.3	65.4	87.6	92.9	56.2	79.3	87.3	34.6	59.4	69.8
MagicLens-L	84.6	96.2	98.8	**72.5**	**91.5**	**95.2**	55.9	78.7	86.3	**44.3**	**69.4**	**78.3**
CoCa-L†	91.4	99.2	99.9	79.0	95.1	97.4	65.4	85.6	91.4	50.1	73.8	81.8
CoCa-L⋆	92.1	98.8	99.9	78.4	94.2	96.9	65.1	85.5	91.3	49.3	73.2	81.5
MagicLens-L	89.6	98.7	99.4	**79.7**	**95.0**	**97.4**	**67.7**	**87.6**	**92.7**	**53.1**	**77.4**	**84.9**

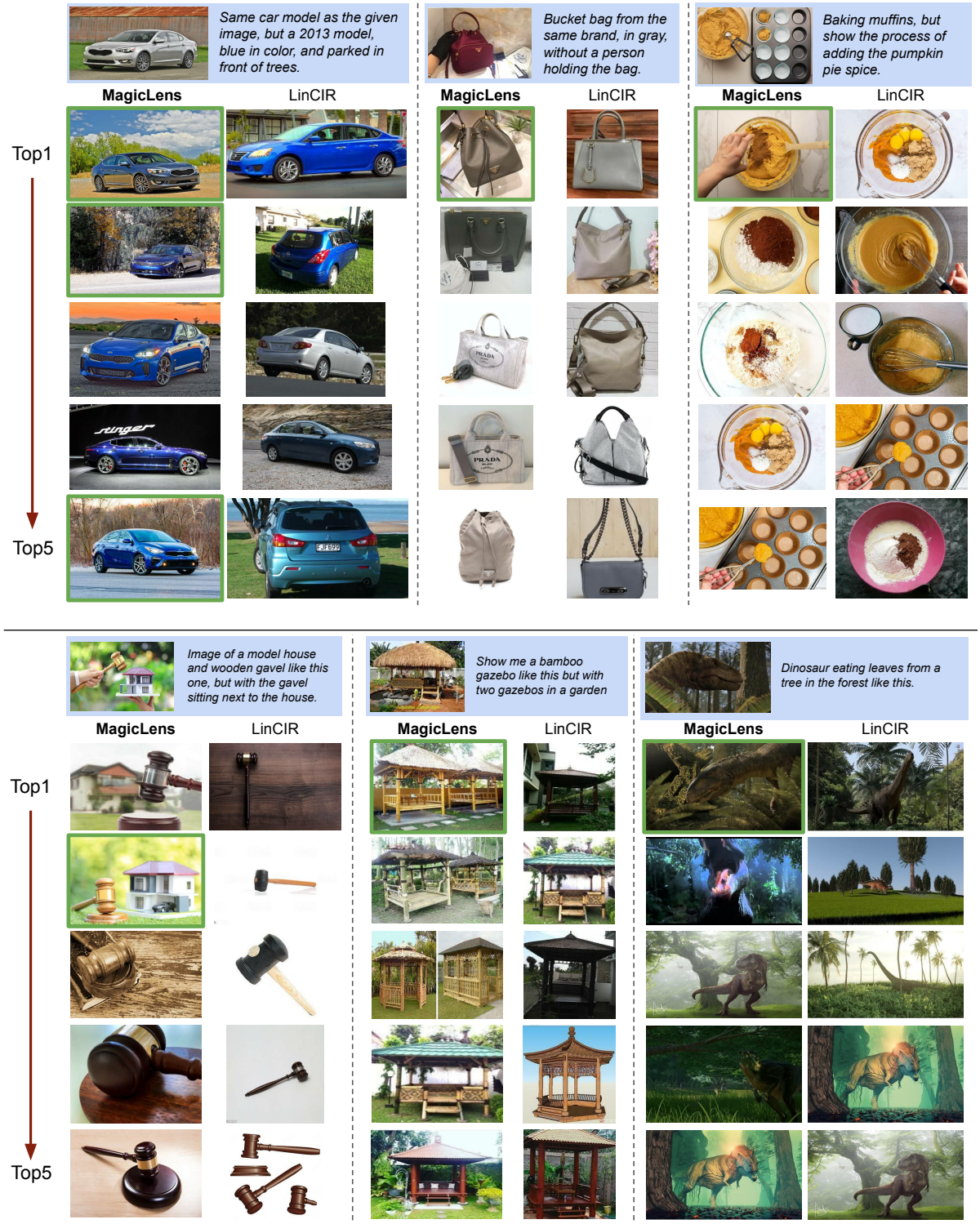

Figure 11. Top-5 retrieved images of CoCa-based MagicLens-L and LinCIR on the holdout index set with 1.4M images for queries shown in Figure 8 and more. Queries are with a blue background and only the most correct retrieved images are marked with green outlines.