Missing Target-Relevant Information Prediction with World Model for Accurate Zero-Shot Composed Image Retrieval

Supplementary Material

Table of Contents

We provide more details and a discussion of the main paper.

A. More Ablation Study

Table 1 presents additional ablation analyses for our PrediCIR model. **In models '1-3', we assessed the impact of varying crop sizes for constructing source and target views.** Using different crop sizes, unlike the consistent size in model '1', results in significant performance degradation. This decline is attributed to discrepancies in position embeddings between the source and target views, which complicate the model's ability to predict features spatially aligned with the reference image. **In models '4-6', we explored the effects of different aspect ratios.** Altering the aspect ratios, whether increasing (model '5') or decreasing (model '6'), led to an average performance decline of 2.30% and 3.22%, respectively, underscoring the sensitivity of model performance to aspect ratio adjustments. **In models '7-11', we further evaluated the impact of alternative solutions for key modules.** The results demonstrate that omitting our dynamic cropping strategy (model '8') or excluding reference image features in gating, which solely predicts the entire target image (model '9'), resulted in average performance reductions of 4.12% and 3.50%, respectively. This confirms the critical role of our strategies in maintaining model efficacy. Additionally, attempting to predict multiple target views from a single source view (model '10') also led to an average performance decline of 2.40%, further validating the effectiveness of our targeted cropping strategy. Using a Faster R-CNN detector on CC3M for semantic-aware cropping (model '11'), which resulted in a 3.90% performance decrease on CIRR and FashionIQ. While this strategy preserves object boundaries, it limits the diversity of training samples, particularly for fine-grained attribute manipulations like FashionIQ (drops by 6.15%). In contrast, our simple but effective random cropping strategy ensures richer and more variable training samples, benefiting the predictive world model despite possible inappropriate bboxs, which aligns with prior findings (*e.g.,* MAE [13], I-JEPA [1, 18]).

		CIRR			Fashion-IQ	
	Methods	R1	R5	R10	R10	R50
Influence of different crop sizes for world view generation						
1.	Source:$(0.2, 0.25)$, Target:$(0.2, 0.25)$	27.2	57.0	70.2	30.1	52.3
2.	Source:$(0.15, 0.2)$, Target:$(0.2, 0.25)$	24.7	53.9	66.2	25.5	48.1
3.	Source:$(0.2, 0.25)$, Target:$(0.15, 0.2)$	25.3	54.8	67.1	26.8	49.4
Influence of aspect ratios						
4.	aspect ratios: $(0.75, 1.5)$	27.2	57.0	70.2	30.1	52.3
5.	aspect ratios: $(1.0, 1.5)$	25.7	55.0	67.5	27.2	49.9
6.	aspect ratios: $(0.75, 1.0)$	25.0	54.2	66.4	26.3	48.8
Influence of different crop strategies						
7.	single-blocks	27.2	57.0	70.2	30.1	52.3
8.	w/o dynamic crop strategy	23.8	54.1	66.8	25.5	46.0
9.	w/o source	24.5	54.7	67.0	25.9	47.2
10.	multi-blocks	25.6	55.2	67.4	27.2	49.4
11.	semantic-aware crop strategy	24.7	55.1	67.4	24.8	45.3

Table 1. More ablation study on CIRR and FashionIQ.

B. More Effectiveness and Efficiency Analysis

B.1. GeneCIS full results

In Table 2, we report the full table of GeneCIS results.

B.2. Comparison with fewer training samples.

In Table 2-7, we present more evidence supporting the efficacy and efficiency of our PrediCIR. With only 50% of the training data, PrediCIR matches and exceeds the performance of the state-of-the-art (SoTA) Context-I2W model, proving our method's superiority.

C. Visualization of Predictor Representations

In Figure 1, we leverage the RCDM framework to visualize more samples of our PrediCIR's predicted target image feature into pixel space (Please refer to Section G.1) for more details). The prediction effectively identifies the missing visual content in the reference images based on manipulation texts (*e.g.,* a Papa Smurf print, a dog not eating, a monkey in origami style, and a dog facing the camera). This pattern remains consistent, proving our predictor's ability

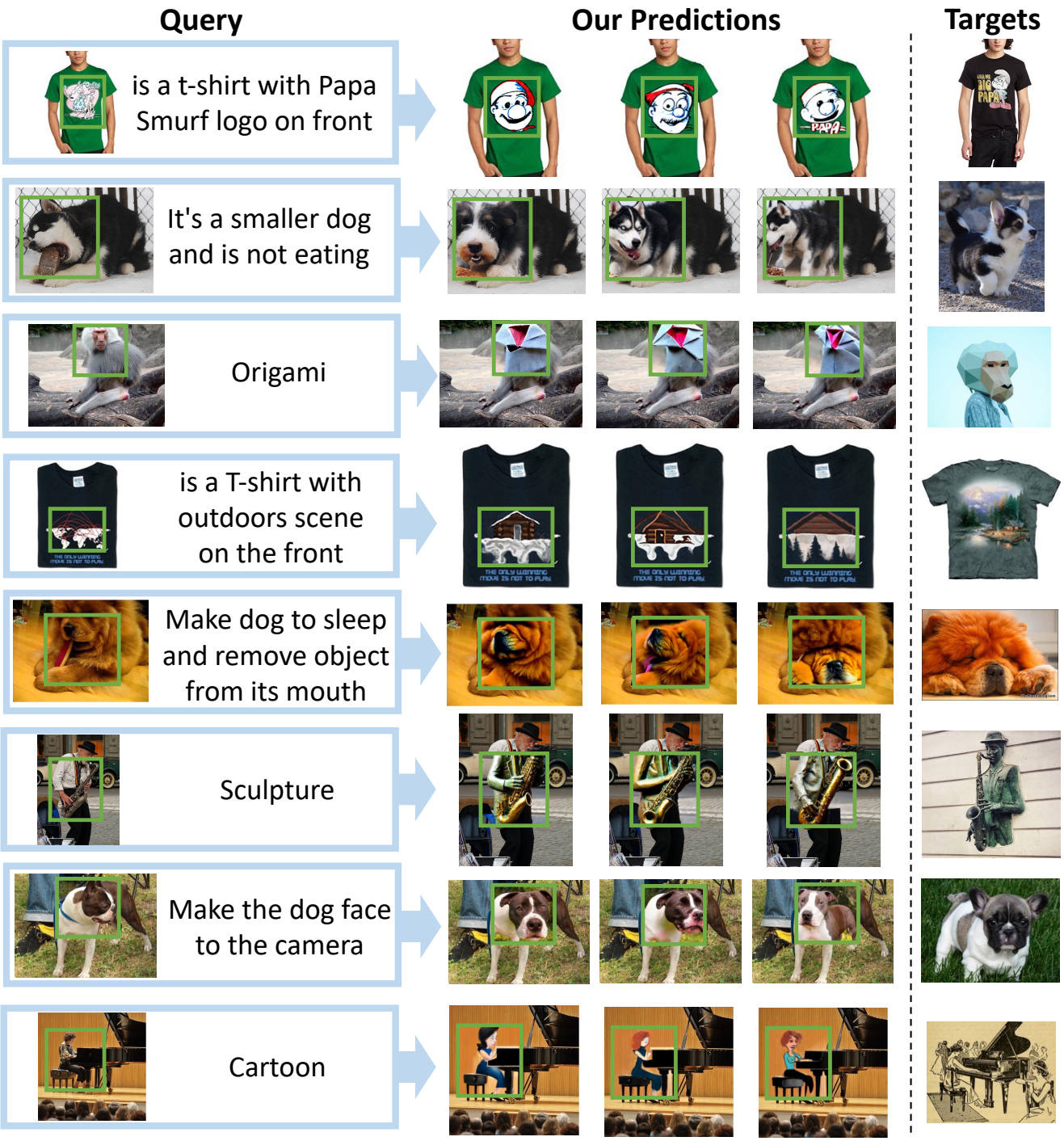

Figure 1. Visualization of our predictor representations. Green bounding boxes contain samples from a generative model decoding the output of our pretrained predictor.

GeneCIS →		Focus Attribute			Change Attribute			Focus Object			Change Object			Average
Backbone	Method	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
ViT-L/14	SEARLE	17.1	29.6	40.7	16.3	25.2	34.2	12.0	**22.2**	30.9	12.0	24.1	33.9	14.4
	LinCIR	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
	Context-I2W	17.2	30.5	41.7	16.4	28.3	**37.1**	8.7	17.9	26.9	7.7	16.0	25.4	12.7
	PrediCIR(50%)	17.7	31.4	42.2	17.8	29.3	35.4	10.7	18.4	29.0	12.5	20.6	29.8	14.7
	PrediCIR(100%)	**18.2**	**31.9**	**42.6**	**18.7**	**30.4**	35.4	**12.7**	19.0	**31.2**	**16.9**	**25.5**	**34.1**	**16.6**
ViT-G/14*	LinCIR	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
	CompoDfff	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
	PrediCIR	**19.3**	**33.2**	**42.7**	**19.9**	**30.7**	**38.9**	**12.8**	**19.4**	**32.3**	**18.9**	**32.2**	**40.6**	**18.7**

Table 2. **Comparison on GeneCIS Test Data.** PrediCIR is able to significantly outperform adaptive methods across all Fashion-IQ sub-benchmarks, with its inherent modularity allowing for further simply scaling to achieve additional large gains. (*) OpenCLIP weights [16].

			Dress		Shrit		TopTee		Average	
Backbones	Methods	Conferences	R10	R50	R10	R50	R10	R50	R10	R50
ViT-L/14	Pic2Word†	CVPR 2023	20.0	40.2	26.2	43.6	27.9	47.4	24.7	43.7
	SEARLE-XL†	ICCV 2023	20.3	43.2	27.4	45.7	29.3	50.2	25.7	46.3
	LinCIR†	CVPR 2024	20.9	42.4	29.1	46.8	28.8	50.2	26.3	46.5
	Context-I2W†	AAAI 2024	23.1	45.3	29.7	48.6	30.6	52.9	27.8	48.9
	PrediCIR(50%)	–	24.2	47.8	30.8	50.2	31.9	54.3	29.0	50.8
	PrediCIR(100%)	–	**25.4**	**49.5**	**31.8**	**52.0**	**33.1**	**55.4**	**30.1**	**52.3**
ViT-G/14	CompoDiff†	TMLR 2024	37.8	49.1	41.3	55.2	44.3	56.4	39.0	51.7
	LinCIR†	CVPR 2024	38.1	60.9	46.8	65.1	50.5	71.1	45.1	65.7
	PrediCIR	–	**39.2**	**61.8**	**47.1**	**67.0**	**52.5**	**72.8**	**46.3**	**67.2**

Table 3. Results on Fashion-IQ for attribute manipulation. †indicates results from the original paper.

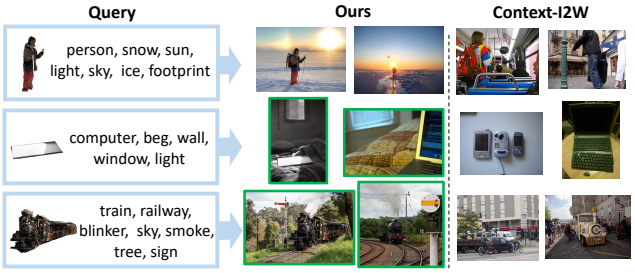

Figure 2. Retrieved results on the object composition task.

to capture positional uncertainty and generate high-level visual elements (*e.g.,* objects, senses, attributes, and different details) with accurate poses. These results highlight the model's capacity for fine-grained visual content prediction, which is crucial for accurate ZS-CIR.

D. More Qualitative Experiment on COCO

In the object composition experiments, PrediCIR significantly outperforms the current SoTA model by an average of 3.60%. These results underscore the remarkable effectiveness of our TCP module in predict missing objects relevant to manipulation text, which facilitates the combination of multiple objects, as shown in Figure 2.

E. Algorithm of Prediction-based Word Mapping Process.

Algorithm 1 presents the pseudo-code for our prediction-based image-to-word mapping process. We initiate the process by creating mask tokens for a target block. The mask tokens are parameterized by a shared learnable vector with an added positional embedding. These mask tokens are designed to predict the visual content missing in the reference image. These mask tokens are subsequently fed into a narrow Transformer architecture, which incorporates source local features and the action with manipulation intent to perform self-attention. To achieve a dynamic ratio during the fusion of source and predict embeddings, we utilize a tanh-gating mechanism [15].

F. Review of Image World Model

F.1. JEPA Framework Overview

The Image World Model (IWM) [10]. builds upon the Joint Embedding Predictive Architecture (JEPA) framework [18], as utilized in approaches like I-JEPA [1]. In JEPA-based methods, representations are learned by predicting the effect of transformations applied to an image in a latent space. This is achieved by conditioning the predictor on transformation parameters, allowing it to infer the relationship between source and target representations effectively.

Unlike contrastive methods that aim for invariance to

Backbones	Methods	Conferences	Cartoon		Origami		Toy		Sculpture		Average	
			R10	R50	R10	R50	R10	R50	R10	R50	R10	R50
ViT-L/14	Pic2Word†	CVPR 2023	8.0	21.9	13.5	25.6	8.7	21.6	10.0	23.8	10.1	23.2
	SEARLE-XL	ICCV 2023	9.6	24.9	16.1	27.3	7.6	25.4	11.3	26.4	11.2	26.0
	LinCIR	CVPR 2024	9.4	24.2	15.7	26.9	10.8	27.0	11.7	27.9	11.9	26.5
	Context-I2W†	AAAI 2024	10.2	26.1	17.5	28.7	11.6	27.4	12.1	28.2	12.9	27.6
	PrediCIR(50%)	–	11.4	27.5	18.2	31.4	13.0	28.4	13.1	30.6	13.9	29.5
	PrediCIR(100%)	–	**14.2**	**31.9**	**20.4**	**34.3**	**14.7**	**30.8**	**16.3**	**34.9**	**16.4**	**33.0**
ViT-G/14	LinCIR	CVPR 2024	13.7	30.2	19.5	32.9	13.8	30.2	15.2	34.0	15.5	31.8
	PrediCIR	–	**15.6**	**34.6**	**23.7**	**37.2**	**17.2**	**37.5**	**19.3**	**37.8**	**19.0**	**36.8**

Table 4. Results on ImageNet for domain conversion. †indicates results from the original paper.

Backbones	Methods	R1	R5	R10
ViT-L/14	Pic2Word†	23.9	51.7	65.3
	SEARLE-XL†	24.2	52.4	66.3
	LinCIR†	25.0	53.3	66.7
	Context-I2W†	25.6	55.1	68.5
	PrediCIR(50%)	26.3	55.7	69.2
	PrediCIR(100%)	**27.2**	**57.0**	**70.2**
ViT-G/14	CompoDiff†	26.7	55.1	74.5
	LinCIR†	35.3	64.7	76.1
	PrediCIR	**36.6**	**65.7**	**77.6**

Table 5. Results on CIRR for object manipulation. †indicates results from the original paper.

Backbones	Methods	R1	R5	R10
ViT-L/14	Pic2Word†	11.5	24.8	33.4
	SEARLE-XL	13.3	28.3	37.6
	LinCIR	11.7	24.9	34.2
	Context-I2W†	13.5	28.5	38.1
	PrediCIR(50%)	14.6	30.1	39.8
	PrediCIR(100%)	**15.1**	**33.0**	**42.8**
ViT-G/14	LinCIR	14.8	30.6	40.5
	PrediCIR	**17.2**	**34.8**	**45.9**

Table 6. Results on COCO for object composition. †indicates results from the original paper.

Backbones	Methods	mAP@5	mAP@10	mAP@25	mAP@50
ViT-L/14	Pic2Word	8.7	9.5	10.6	11.3
	SEARLE-XL†	11.7	12.7	14.3	15.1
	LinCIR†	12.6	13.6	15.0	15.9
	Context-I2W	13.0	14.6	16.1	17.2
	PrediCIR(50%)	14.3	15.7	17.2	18.1
	PrediCIR(100%)	**15.7**	**17.1**	**18.6**	**19.3**
ViT-G/14	CompoDiff†	15.3	17.7	19.5	21.0
	LinCIR†	19.7	21.0	23.1	24.2
	PrediCIR	**23.7**	**24.6**	**25.4**	**26.0**

Table 7. Results on CIRCO for object manipulation.

data augmentations, JEPA frameworks preserve semantic information through latent inpainting, enabling the predictor to model transformations explicitly. By working in the latent space, JEPA removes redundant or hard-to-predict details, improving representation quality without focusing on pixel-level reconstruction [6]. These features make JEPA a powerful tool for learning representations that are both semantically meaningful and capable of generalization.

F.2. Image World Model (IWM)

IWM extends the JEPA framework to learn robust and reusable world models. The predictor in IWM serves as the instantiation of the world model, capable of applying transformations in latent space. Unlike invariant predictors, which disregard transformation details, IWM learns equivariant representations by conditioning on transformation parameters [10].

The training process begins with the generation of source (x) and target (y) views from a given image I. Target views are created by applying random augmentations such as horizontal flips, cropping, and color jitter, ensuring the target retains as much semantic information as possible. In contrast, source views are derived from the target by introducing additional transformations, including grayscale, blur, solarization, and masking inspired by I-JEPA. These transformations enforce the predictor to learn transformation-aware latent representations.

Transformation Encoding. The transformation parameters $a_{x\to y}$ encode the differences between source and target views, including augmentation details such as color jitter and destructive transformations. These parameters serve as input to the predictor, allowing it to model the transformations explicitly.

Latent Prediction. The source and target views are processed by an encoder f_θ and its exponential moving average (EMA) f_θ^{EMA} to obtain latent representations z_x and z_y. The predictor p_ϕ is conditioned on the source embedding, transformation parameters, and masked token positions to predict the target representation $\hat{z_y}$. The learning objective minimizes the $L2$ distance between the predicted $\hat{z_y}$ and the actual target z_y over masked regions:

$$L(x,y) = \sum_{i \in M_x^C} \left\| p_\phi \left(f_\theta(x), a_{x\to y}, m_a \right)_i - f_\theta^{\text{EMA}}(y)_i \right\|_2^2.$$

Algorithm 1 Prediction-based Word Mapping process.

Input: batch of source image features $\boldsymbol{V}_x = \{\boldsymbol{v}_{x_i}\}_{i=1}^{m}$, where $\boldsymbol{v}_{x_1}$ is the global source feature $\boldsymbol{v}_{x_g}$, batch of action $a_{\boldsymbol{x}\to\boldsymbol{y}}$ with manipulation intent, N_{layer}.
Parameter: mask tokens m_a, parameterized by a shared learnable vector $\boldsymbol{x} \in \mathbb{R}^{d\times 1}$ with an added positional embedding, 8-heads attention layer $Attn$, 3-layers FC layers f_M, $gate_\alpha$.
Output: pseudo token S_*

Initialize $m_a \in \mathbb{R}^{d\times n}$, $Attn$, f_M randomly.
Let $\boldsymbol{X}_{att}^{i} = [a_{\boldsymbol{x}\to\boldsymbol{y}}, \{\boldsymbol{v}_{x_i}\}_{i=2}^{m}, m_a]$,$t = 1$
while $t \le N_{layer}$ **do**
  $\boldsymbol{X}_{att}^{i+1} = \boldsymbol{X}_{att}^{i} + Attn_t(\text{q}=\boldsymbol{X}_{att}^{i+1}, \text{k}=\boldsymbol{X}_{att}^{i+1}, \text{v}=\boldsymbol{X}_{att}^{i+1})$
  $\boldsymbol{X}_{att}^{i+1} = \boldsymbol{X}_{att}^{i+1} + f_{M_t}(\boldsymbol{X}_{att}^{i+1})$
  $t = t + 1$
end while
$S_* = f_{M_s}(\boldsymbol{v}_{x_g}) + tanh(gate_\alpha) \cdot avg(f_{M_p}(\boldsymbol{X}_{out}))$
return S_*

Architecture. The encoder of IWM adopts the ViT architecture [9], while the predictor uses a similar structure with modified depth and embedding dimensions. IWM instances are denoted as $\text{IWM}_{X,Y}^{Z}$, where X is the predictor depth, Y its embedding dimension, and Z specifies its capability, such as "Equi" for equivariant models.

F.3. The Reusability of IWM

IWM not only enhances representation learning but also enables effective downstream task adaptation. Finetuning the learned world model alongside the frozen encoder significantly improves task performance with minimal additional cost. Furthermore, inspired by instruction tuning [29], IWM can be adapted for multi-task learning, demonstrating its efficiency and versatility compared to traditional methods. This highlights the importance of incorporating the world model into inference processes, rather than discarding it after pretraining.

G. More Implementation Details

For training PrediCIR, We adopt ViT-B/32 and ViT-L/14 CLIP [23] pre-trained on 400M image-text paired data. The crop sizes and aspect ratios of random cropped images and blocked target images are the same, in the range of $(0.2, 0.25)$ and $(0.75, 1.5)$, respectively (ablation in the supplementary). For training PrediCIR, we utilize the Conceptual Caption dataset [25], which comprises 3M images. Our predictor is designed as a lightweight (narrow) ViT architecture. Specifically, the number of self-attention blocks is 12 with 384 dimensional embeddings. To improve training stability, we initialize the learnable scalar of tanh-gating to 0 [2]. We employ AdamW [21] with a learning rate of 1×10^{-5}, weight decay of 0.1, and a linear warmup of 10000 steps. The batch size is 1024. For training Pic2Word, SEARLE, Context-I2W, and LinCIR, we utilized their official code for training, and hyper-parameters were kept consistent with those reported in their respective papers. All models are trained on 4 NVIDIA A100 (80G) GPUs. Moreover, we conduct ablation studies on CIRR test sets and FashionIQ validation sets. For Fashion-IQ, we consider the average recall. To ensure reliable results, we report the performance averaged over three trials.

G.1. RCDM Visualizations Details.

In Figure 7 of our main paper and Figure 1, to visualize the representations of a pre-trained neural network in pixel space, we follow I-JEPA [1], freeze our PrediCIR, and train a decoder following the RCDM framework [5]. The RCDM framework trains a decoder network h_ω, comprising a generative diffusion model, to reconstruct an image $\boldsymbol{x}$ from the representation vector of that image $\boldsymbol{s}_x$ and a noisy version of that image $\hat{\boldsymbol{x}} := \boldsymbol{x} + \epsilon$, where ϵ is an additive noise vector. Concretely, the decoder objective is to minimize the loss function $\|h_\omega(\hat{\boldsymbol{x}}, \boldsymbol{s}_x) - \epsilon\|$. We train each RCDM network for 350,000 iterations using the default hyperparameters. After training the decoder, one can subsequently feed the representation vector of an unseen test image $\boldsymbol{s}_y$ into the decoder along with various random noise vectors to generate several pixel-level visualizations of the representation, thus providing insight into the features captured in the representations of the pre-trained network. Qualities that are common across samples represent information that is contained in the representation. On the other hand, qualities that vary across samples represent information that is not contained in the representations.

G.2. More Evaluation Datasets Details

We evaluate our model on six ZS-CIR datasets, *i.e.,* COCO [19] and GeneCIS [28] for object/attribute composition, ImageNet [8, 14] for domain conversion, CIRR [20] and CIRCO [4] for object/scene manipulation, and FashionIQ [30] for attribute manipulation. Following the original benchmarks, we use Recall@k (R@k) as the evaluation metric for CIRR, GeneCIS, and FashionIQ, COCO, ImageNet and mean average precision (mAP@k) for CIRCO to account for multiple positives. We also evaluate CIRR in a subset setting, where $\text{Recall}_{\text{Subset}}$@k measures retrieval performance within a limited selection of images relevant to the query in the database. The evaluation datasets are preprocessed, as explained in the main paper, we describe the details of the dataset, *i.e.,* number of query images and candidate images used for evaluation.

FashionIQ [30] is a dataset of fashion-related images across three categories: Shirt, Dress, and Toptee, comprising 30,134 triplets from 77,684 images. The dataset was curated by collecting image attributes and then tasking human annotators to write captions describing highly related images based on those attributes. FashionIQ simulates realistic user interactions, as captions were generated via a chat-based visual interface to mimic online shopping queries. The dataset is divided into training (60%), validation (20%), and test (20%) splits. For zero-shot CIR, we use only the validation split, as the test set annotations are not publicly available.

CIRR [20] contains 21,552 real-world images sourced from NLVR^2 [26]. The dataset includes training, validation, and test splits, with the latter evaluated via a remote server. Our analysis focuses on the validation split for model selection. Unlike FashionIQ, which targets fashion-specific queries, CIRR encompasses diverse domains with complex descriptions. The dataset was built by identifying visually similar images using ResNet-152 [12] pretrained on ImageNet [8] and employing human annotators to describe differences between paired images. However, CIRR suffers from two key issues: (1) image pairs identified by ResNet often lack true visual similarity, as they were not verified by human annotators; and (2) captions are often unrealistic or ambiguous, including unnecessary details. These limitations reduce CIRR's practical relevance compared to FashionIQ. Additionally, CIRR uses a small subset retrieval task (e.g., five items) to mitigate noise, but this approach is problematic, as the target image often relates only to the text condition rather than the reference image. Previous studies [4, 11, 24], have noted the prevalence of false negatives (FNs) in CIRR, complicating evaluation accuracy, as seen in other cross-modal retrieval tasks [7, 31].

Notably, both FashionIQ and CIRR face challenges from FN instances. While each query has a single labeled positive, multiple valid matches may exist in the dataset. FashionIQ mitigates this by reporting Recall@K with larger K values (*e.g.,* 10 or 50), whereas CIRR employs subset retrieval. However, these approaches fail to fundamentally resolve the FN issue, particularly for CIRR's noisy annotations.

CIRCO [4] builds on the COCO dataset [19], addressing the FN problem by including an average of 4.53 ground truths per query. This design enables more reliable evaluation using metrics like mAP. CIRCO contains no training split and provides validation (220 queries) and test (800 queries) splits, with the latter evaluated remotely.

GeneCIS [28] defines conditional retrieval tasks focusing on attributes (*e.g.,* "focus on an attribute," "change an attribute") and objects (*e.g.,* "focus on an object," "change an object"). Attribute tasks use VisualGenome [17] and VAW [22], while object tasks are based on COCO [19]. Each task comprises around 2,000 queries with a small gallery size (*e.g.,* 15 images, 10 for "focus on an attribute") to limit FNs. Text queries correspond to attributes or objects (*e.g.,* "color," "backpack").

COCO [19] contains images with corresponding lists of object classes and instance masks of query images. Following Pic2Word, we randomly crop one object and mask its background using its instance mask to create a query for each image. The list of object classes is used as text specification.

ImageNet [19] consists of 200 classes across diverse domains with domain annotations. Unlike previous benchmarks, the task involves retrieving an image in the specified domain for the same semantic object category (*e.g.,* retrieving a cartoon goldfish given a natural goldfish reference image and the modifier `"cartoon"`). This task requires no reasoning over image semantics, as the modifier independently specifies a domain change. Significant improvements over Pic2Word and Context-I2W can be achieved by leveraging the final description format, `"a domain of a caption"`.

Table 8. The number of images used for evaluation in each dataset.

Dataset	Query images	Candidate images
CIRR (Test)	4,148	2,315
CIRCO (Test)	800	123,403
Fashion (Dress)	2,017	3,817
Fashion (Shirt)	2,038	6,346
Fashion (TopTee)	1,961	5,373
GeneCIS (Focus Attribute)	2000	10
GeneCIS (Change Attribute)	2112	15
GeneCIS (Focus Object)	1960	15
GeneCIS (Change Object)	1960	15
COCO	4,766	4,766
ImageNet	10,000	16,983

G.3. More Inference Details

(1) Domain conversion. This setup evaluates the ability to compose real images and domain information to retrieve corresponding domain-specific images. We utilize ImageNet [8] and ImageNet-R [14], which comprises 200 classes with diverse domains and has domain annotations. Following Pic2Word, we pick cartoon, origami, toy, and sculpture as the evaluation target to avoid noise in the annotations. With this selection, we have 16,983 images as candidates. In the evaluation, given the real image from ImageNet and target domain names, we compose the query

following the procedure in (a) in the Inference section. *e.g.,* `a cartoon of [*].`

(2) Object/Attribute composition. We evaluate the GeneCIS [28] test split and the validation split (5000 images) of COCO [19],, which dataset contains images with corresponding lists of object classes and instance mask of query images. Following Pic2Word, we randomly crop one object and mask its background using its instance mask to create a query for each image. The list of object classes is used as text specification. Given the reference image and class list, we compose a query by following (b) in the Inference section. *e.g.,* `a photo of [*], [cat] and [dog].`

(3) Object/scene manipulation by text description. In this setup, a reference image is provided alongside a text description containing instructions for manipulating either an object or the background scene depicted in the reference image. This composition of the reference image and text description enables the retrieval of manipulated images. We evaluate the test split of CIRR [20] and CIRCO [4] using the standard evaluation protocol following previous works [4, 24, 27], and query texts are composed following the procedure `a photo of [*], [sentence].`

(4) Attribute manipulation. We employ Fashion-IQ [30], which includes various modification texts related to image attributes. These attribute manipulations are given as a sentence. As with CIRR, we adopt the standard evaluation protocol and create query texts following the procedure `a photo of [*], [sentence]`. In evaluation, we employ the validation set, following previous works [3, 4, 24, 27].

G.4. More Effectiveness and Efficiency Analysis

Our approach achieves significant improvements across six ZS-CIR tasks, with performance gains ranging from 1.73% to 4.45% Over SoTA models. Due to our predictor design for prediction-based mapping, our model size (99.8M) is larger than the simple 3-layer MLP mapping (0.9M) of Pic2Word. As a result, in the same setting, our training time (28 hours) is 6 hours longer than Pic2Word and 18 hours longer than SEARLE. Despite requiring 25 additional training hours compared to LinCIR, an efficient training model, PrediCIR completes training 203 hours faster than the diffusion-based semi-supervised CompoDiff, achieving significant performance gains. Our inference time(0.03s) is only 0.01s slower than LinCIR and four times faster than CompoDiff (0.12s).

References

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 1, 3, 5

[2] Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. In *Uncertainty in Artificial Intelligence*, pages 1352–1361, 2021. 5

[3] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21466–21474, 2022. 7

[4] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. *arXiv:2303.15247*, 2023. 5, 6, 7

[5] Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about. *arXiv preprint arXiv:2112.09164*, 2021. 5

[6] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*, 2024. 4

[7] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008. 6

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5, 6

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 5

[10] Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning, 2024. 3, 4

[11] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, , Yoohoon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval. In *Conference on Computer Vision and Pattern Recognition*, 2024. 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1

[14] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 5, 6

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[16] 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. 3

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, page 32–73, 2017. 6

[18] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022. 1, 3

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 5, 6, 7

[20] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021. 5, 6, 7

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 5

[22] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *CVPR*, 2021. 6

[23] 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 International Conference on Machine Learning*, pages 8748–8763, 2021. 5

[24] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314, 2023. 6, 7

[25] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565, 2018. 5

[26] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. 6

[27] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5180–5188, 2024. 7

[28] Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *CVPR*, 2023. 5, 6, 7

[29] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 5

[30] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11307–11317, 2021. 5, 6, 7

[31] Jason Zhu, Yanling Cui, Yuming Liu, Hao Sun, Xue Li, Markus Pelger, Tianqi Yang, Liangjie Zhang, Ruofei Zhang, and Huasha Zhao. Textgnn: Improving text encoder via graph neural network in sponsored search. In *Proceedings of the Web Conference*, pages 2848–2857, 2021. 6