

Generative Zero-Shot Composed Image Retrieval

Supplementary Material

A. Additional Ablation Study and Performance Analysis

A.1. Discussion on λ

We conducted an ablation study to assess the effect of varying the hyperparameter λ on model performance. Table 1 reports the results on the CIRR validation set using a ViT-L/14 backbone across different λ values. While performance remains relatively stable across the tested settings, $\lambda = 0.4$ yields the highest Recall@1 and is therefore selected for our final model configuration. For the CIRR and CIRCO datasets, we selected the optimal λ through the validation set and applied it to the testing set. For other datasets, we experimentally chose a relatively optimal lambda. For CIRR, CIRCO, Fashion-IQ, and GeneCIS, the λ parameters are set to 0.4, 0.5, 0.1, 0.1 respectively.

Table 1. Ablation study of λ . Performance is reported using Recall@K. $\lambda = 0.4$ achieves the highest Recall@1.

Lambda	Recall@K			
	K=1	K=5	K=10	K=50
0.2	25.02	54.99	68.52	89.79
0.3	25.42	55.37	68.62	89.72
0.4	26.17	54.75	68.14	89.24
0.5	25.97	55.32	68.69	89.79

A.2. Additional ZS-CIR Benchmark Comparisons

We further validate the effectiveness of CIG on both the CIRR and CIRCO validation sets, as presented in Table 2 and Table 3. CIG consistently delivers notable performance gains on these benchmarks. Specifically, when applied to the SEARLE baseline, CIG improves Recall@1 on the CIRR validation set by 2.06%. On the CIRCO validation set, CIG also enhances SEARLE’s mAP by nearly 1%, further demonstrating its ability to generalize and boost retrieval performance across different datasets.

Table 2. Quantitative results on the CIRR validation set.

Backbone	Method	Recall@K			
		K = 1	K = 5	K = 10	K = 50
ViT-L/14	Pic2Word [4]	22.6	52.6	66.6	87.3
	SEARLE [2]	24.11	54.68	68.02	89.09
	Pic2Word + CIG	23.42	52.67	65.63	87.37
	SEARLE + CIG	26.17	54.75	68.14	89.24

A.3. Discussion on different number of pseudo-target images.

We conduct experiments using multiple pseudo-target images. As shown in Table 4, leveraging two or three pseudo-

Table 3. Quantitative results on the CIRCO validation set.

Backbone	Method	mAP@K			
		K=5	K=10	K=25	K=50
ViT-B/32	Image-only	1.61	2.16	2.73	3.10
	Text-only	2.96	3.29	3.74	3.89
	Image + Text	2.63	3.58	4.52	4.94
	SEARLE [2]	6.82	7.83	9.15	9.77
	SEARLE + CIG	7.62	8.22	9.36	10.01
ViT-L/14	Pic2Word [4]	7.92	9.02	10.18	10.83
	SEARLE [2]	10.09	11.15	12.83	13.60
	SEARLE + CIG	10.98	12.12	13.65	14.41

target images consistently improves performance on the CIR task by providing richer contextual clues.

A.4. Additional domain exploration: Domain Conversion Setting.

We report domain conversion results on the ImageNet-R [3] benchmark in Table 5. CIG consistently outperforms Pic2Word across most domain shifts, while also surpassing supervised baselines. These results highlight CIG’s stronger generalization and transferability to novel visual domains beyond the standard training distribution.

B. Additional Visualization

Qualitative Evaluation on Composed Image Generation.

Figure 1 visualizes the results on the CIRCO validation dataset, including the reference image, generated image, and the top 3 ground truth target images. In this dataset, the delta captions are well-designed and almost never include keywords from the reference image. Despite this, our generated images still effectively retain the characteristics of the reference images: the architecture style in the first example, the object in the second example, and the color information of the last example are accurately preserved.

Figure 2 visualizes the results on the GeneCIS dataset. In this dataset, the text query is given by the name of the attribute or the object, like "color" or "table". Therefore, the text contains very limited information. Without context information, it poses a challenge to generation tasks. However, CIG still manages to produce good results. In the focus attribute, we generated signs with the same white letter color. In the change attribute example, we successfully changed the train to olive green and better preserved the information from the reference image compared to the target image. In the focus object example, we generated the same kitchen scene and retained the refrigerator. In the change object example, we retained the same bathroom scene and presented a table, better preserving the scene information

Table 4. **Discussion on different number of pseudo-target images.** Incorporating multiple pseudo-target images leads to consistent improvements in CIR by supplying more informative context.

Numbers of Images	CIRR				CIRCO			
	Recall@K				mAP@k			
	@1	@5	@10	@50	k=5	k=10	k=25	k=50
1	26.17	54.75	68.14	89.24	10.98	12.12	13.65	14.41
2	26.17	56.42	68.67	89.69	11.24	12.24	13.75	14.40
3	26.43	56.25	68.72	89.67	11.83	12.51	14.13	14.76

Table 5. **Domain conversion using ImageNet-R.** CIG consistently surpasses Pic2Word in most domain conversions, highlighting its effectiveness in adapting to various image domains.

Supervision	Methods	Cartoon		Origami		Toy		Sculpture		Average	
		R10	R50	R10	R50	R10	R50	R10	R50	R10	R50
ZERO-SHOT	Image-only	0.3	4.5	0.2	1.8	0.6	5.7	0.3	4.0	0.4	4.0
	Text-only	0.2	1.1	0.8	3.7	0.8	2.4	0.4	0.4	0.5	2.3
	Image+Text	2.2	13.3	2.0	10.3	1.2	9.7	1.6	11.6	1.7	11.2
	Pic2Word [4]	8.0	21.9	13.5	25.6	8.7	21.6	10.0	23.8	10.1	23.2
	Pic2Word + CIG	10.1	22.7	16.5	27.3	10.1	23.1	10.7	22.5	11.9	23.9
CIRR	Combiner [1]	6.1	14.8	10.5	21.3	7.0	17.7	8.5	20.4	8.0	18.5
Fashion-IQ	Combiner [1]	6.0	16.9	7.6	20.2	2.7	10.9	8.0	21.6	6.0	17.4

compared to the target image.

Qualitative Evaluation on Composed Image Retrieval.

Figure 3 shows some examples of retrieval results on CIRR validation dataset with SEARLE and SEALE + CIG settings. SEARLE is able to retrieve close images but still misses some details. In the first example, it retrieves the dog on chair instead of coach. Also in the third example, SEARLE is able to find a wood stairs but no glass. However, our generated image is able to provide more accurate information, like gazebo outside, stairs with glass, which is useful for retrieval task.

Failure Cases. Figure 4 visualized failure cases over different datasets. In the example from CIRR, although we can generate dogs of the same breed, the dog’s actions are unreasonable. The caption requires the dog to hold a toy, and typically dogs would use their mouths to hold objects. However, in the generated image, the dog attempts to hold the toy with its paws. In the Fashion-IQ examples, there are changes in the style of the dress, especially in the neckline. In the CIRCO examples, the generated images fail to correctly judge the quantity of objects. Lastly, in the GeneCIS example, although the generated images contain windows, it fails to retain the scene from the reference image.

C. Limitation and Future Work

The proposed pseudo target image generation algorithm not only helps composed image retrieval but also provides a byproduct, the generated image. In cases like e-commerce and recommendation, real images from the database are

necessary. I Although we improve the performance of composed image retrieval and provide the function of composed image generation, the proposed algorithm introduces extra computational cost and process to CIR. An alternative future direction is to further accelerate the composed image generation to real-time.

References

- [1] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022. 2
- [2] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1
- [3] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1
- [4] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2



Figure 1. **Qualitative Evaluation on CIRCO validation datasets.** In this dataset, the delta captions are carefully crafted and rarely include keywords from the reference images. Nevertheless, our generated images successfully preserve the key characteristics of the reference images: the architectural style in the first example, the object in the second example, and the color details in the last example are accurately maintained.

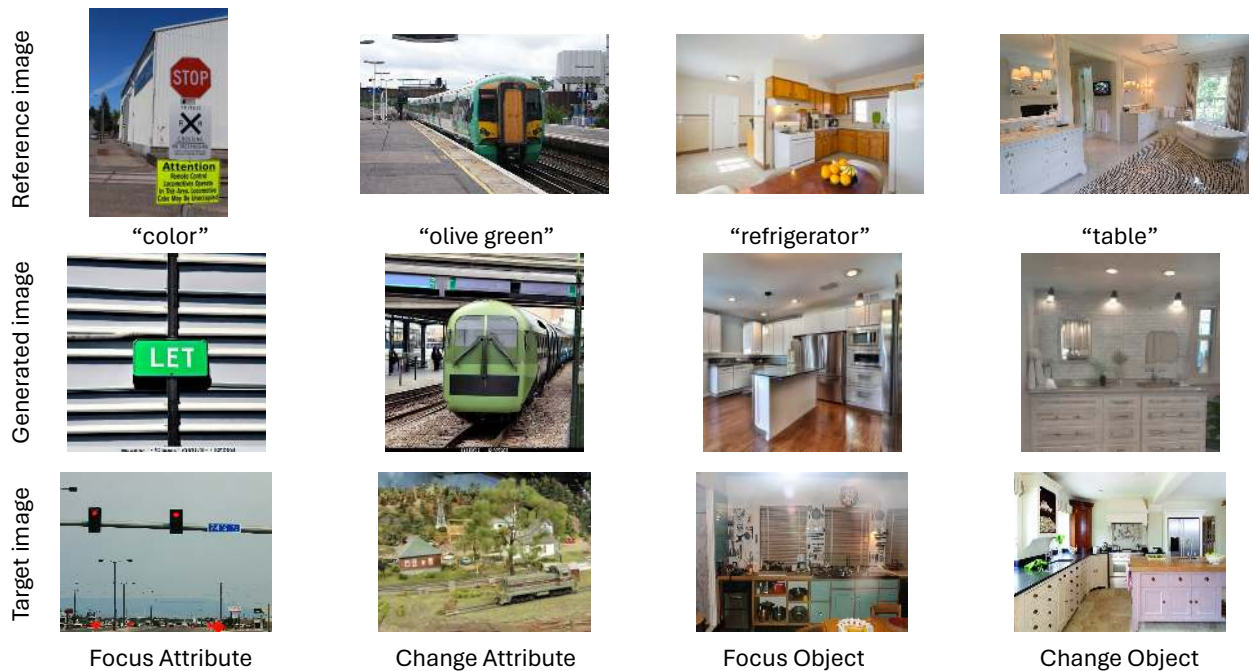


Figure 2. **Qualitative Evaluation on GeneCIS datasets.** In this dataset, the text query consists of simple attribute or object names, such as "color" or "table," providing very limited information. For the focus attribute, we generated signs with consistent white lettering. In the change attribute example, we successfully altered the train's color to olive green while preserving more reference image details than the target image. For the focus object example, we recreated the kitchen scene, retaining the refrigerator. Similarly, in the change object example, we preserved the bathroom scene while adding a table, better maintaining the scene's overall integrity compared to the target image.

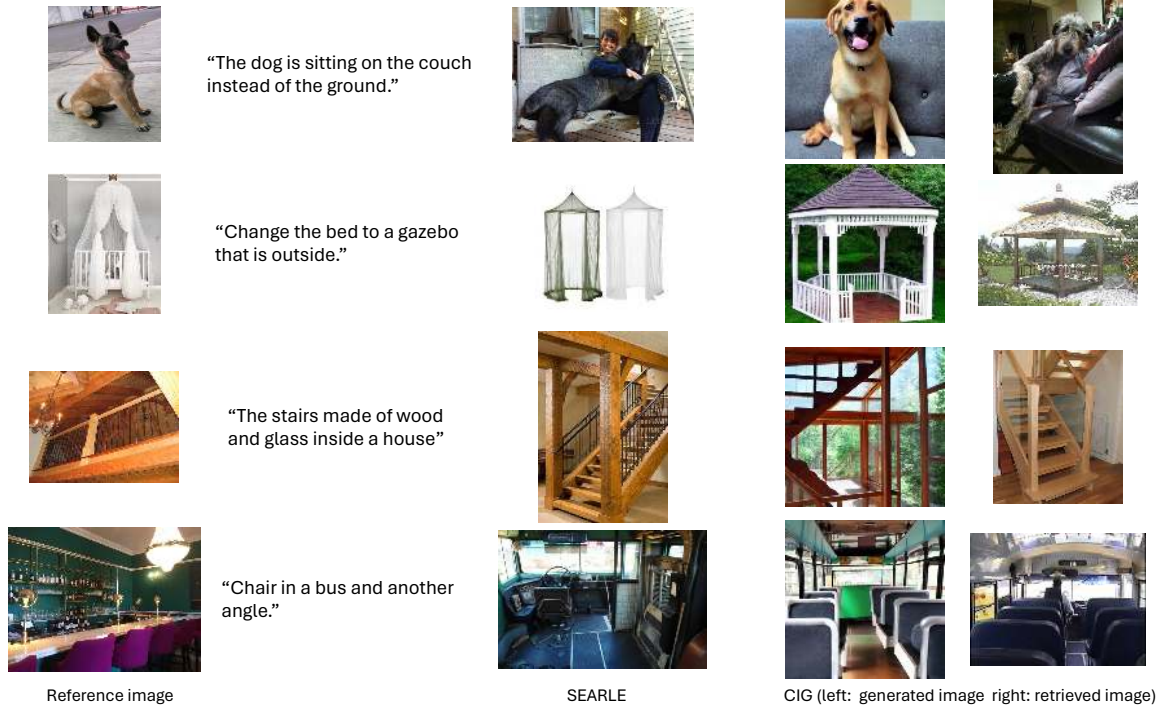


Figure 3. **Visualization on Composed Image Retrieval.** SEARLE demonstrates the ability to retrieve visually similar images but occasionally misses key details. For example, in the first case, it retrieves an image of a dog on a chair instead of a couch. In the third case, while it identifies wooden stairs, it fails to include the glass details. In contrast, our generated image captures more accurate information, such as the gazebo outside and stairs with glass, which enhances its utility for retrieval tasks.



Figure 4. **Failure cases of CIG on different benchmarks.** In the CIRR example, while we successfully generate dogs of the same breed, the dog’s action appears unnatural. In the Fashion-IQ examples, the generated images reflect changes in the dress style, particularly around the neckline. For the CIRCO examples, the generated images struggle to accurately determine the quantity of objects. In the GeneCIS example, although the generated images include windows, they fail to preserve the scene context from the reference image.