Supplementary Material: Zero-shot Composed Text-Image Retrieval

Yikun Liu[1,2]
yikunliu@sjtu.edu.cn
Jiangchao Yao[1,3]
Sunarker@sjtu.edu.cn
Ya Zhang[1,3]
ya_zhang@sjtu.edu.cn
Yanfeng Wang[1,3,†]
wangyanfeng622@sjtu.edu.cn
Weidi Xie[1,3,†]
weidi@sjtu.edu.cn

[1] Coop. Medianet Innovation Center, Shanghai Jiao Tong University, China
[2] Beijing University of Posts and Telecommunications, China
[3] Shanghai AI Laboratory

A Appendix

In this supplementary material, we start by detailing the procedure for dataset construction, namely, Laion-CIR-Template dataset, then we present more detailed experiment comparison. Additionally, we also show the results for model training on a combined dataset of Laion-CIR-Template and Laion-CIR-LLM. Lastly, we present some failure cases from our dataset construction pipeline and several interpretable heatmaps to analyze the reasoning patterns of the TransAgg model.

A.1 Details on constructing Laion-CIR-Template

While constructing the Laion-CIR-Template dataset, we consider editing the captions from eight semantic aspects, as detailed in the following sections.

Cardinality. We identify the reference image captions that contain digits, then we construct the relative caption based on the templates shown in Table 1. Next, we replace "num1" in the reference image caption with "num2" or "a group of" to get the edited caption.

Predefined Template
change to {num2} {noun}.
change to a group of {noun}.
change {num1} {noun} to {num2} {noun}.
change {num1} {noun} to a group of {noun}.
change the num of {noun} from {num1} to {num2}.

Table 1: Predefined templates for cardinality type.

Addition. We randomly select a noun from the reference image caption, and then select another noun that has a similarity score between 0.5 to 0.7 to it. Next, we construct the corresponding relative caption based on the templates listed in Table 2, and obtain the edited caption by adding “with {noun}” to the reference image caption.

Predefined Template
add {noun}.
{noun} has been added.
{noun} has been newly placed.

Table 2: Predefined templates for addition type.

Negation. We randomly select a noun phrase from the reference image caption, then use the template defined in Table 3 to construct a relative caption. The edited caption is created by removing the corresponding noun phrase from the reference image caption.

Direct Addressing. We randomly select images with a similarity score of 0.5 to 0.7 as target images by comparing their description with the reference images. The caption of the selected target image is referred to as the relative caption.

Predefined Template
no {noun_phrase}.
remove {noun_phrase}.
{noun_phrase} is gone.
{noun_phrase} is missing.
{noun_phrase} is no longer there.

Table 3: Predefined templates for negation type.

Compare & Change. First, a noun phrase (noun_phrase1) is randomly selected from the reference image caption. Then, another noun phrase (noun_phrase2) with a similarity score in the range of 0.5 to 0.7 is chosen as the replacement for noun_phrase1. The resulting relative caption is generated using the templates defined in Table 4. The edited caption is obtained by substituting noun_phrase1 in the reference image caption with noun_phrase2.

Predefined Template
not {noun_phrase1}, but {noun_phrase2}.
replace {noun_phrase1} with {noun_phrase2}.
instead of {noun_phrase1}, show {noun_phrase2}.

Table 4: Predefined templates for compare & change type.

Comparative Statement. In this section, we focus on some common adjectives. We start by selecting the adjectives from the reference image caption, and replacing them with their antonyms to create the edited caption. The relative caption is then formed by using the comparative form of the antonym with the noun it modifies.

Viewpoint. We randomly select a noun from the reference image caption, and use the templates from Table 5 to construct a relative caption. We then append either “small” or “big” to the noun depending on the meaning of the relative caption to create an edited caption.

Statement with Conjunction. This section randomly selects two out of the seven scenarios mentioned earlier and combines them randomly. The final relative caption combines each of their respective relative captions using "and". The edited caption is then modified according to their respective rules.

A.2 Detailed Experimental Results

In this section, we present more detailed experimental results.

Predefined Template
focus on the {noun}.
zoom in the {noun}.
zoom out the {noun}.

Table 5: Predefined templates for viewpoint type.

A.2.1 Pretrained backbone and finetuning

The complete experimental results for different backbone and fine-tuning types on the CIRR and FashionIQ datasets are presented in Table 6 and Table 7, respectively.

		Recall@K				$\text{Recall}_{\text{Subset}}$@K		
Backbone	Fine-tuning	K=1	K=5	K=10	K=50	K=1	K=2	K=3
CLIP-B/32	✘	24.46	53.61	67.54	89.81	57.81	78.17	89.54
	only text enc.	27.08	57.21	70.31	90.39	62.70	82.41	92.15
	both	29.30	60.48	73.25	92.31	63.57	82.31	91.95
CLIP-L/14	✘	25.04	53.98	67.59	88.94	55.33	76.82	88.94
	only text enc.	27.90	58.27	71.01	91.30	60.48	80.31	90.75
	both	33.04	64.39	76.27	93.45	63.37	82.27	92.22
BLIP	✘	34.89	64.75	76.24	92.22	66.34	83.76	92.92
	only text enc.	38.10	68.42	79.08	93.51	70.34	86.42	94.28
	both	37.18	67.21	77.92	93.43	69.34	85.68	93.62

Table 6: Generalization for different backbones and fine-tuning types on CIRR.

		Shirt		Dress		TopTee		Average	
Backbone	Fine-tuning	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
CLIP-B/32	✘	25.37	42.69	19.44	42.04	26.93	49.31	23.91	44.68
	only text enc.	27.38	45.58	21.47	43.88	28.15	49.82	25.67	46.43
	both	27.48	46.52	20.58	43.28	27.38	48.50	25.15	46.10
CLIP-L/14	✘	29.54	47.79	23.85	44.57	32.33	52.52	28.57	48.29
	only text enc.	30.91	49.31	27.32	47.79	33.61	54.05	30.61	50.38
	both	34.79	53.39	27.71	49.68	35.39	57.88	32.63	53.65
BLIP	✘	28.07	45.63	21.67	41.89	31.11	50.79	26.95	46.10
	only text enc.	32.83	52.31	27.67	49.38	35.70	58.08	32.07	53.26
	both	34.84	53.93	31.28	52.75	37.79	60.48	34.64	55.72

Table 7: Generalization for different backbones and fine-tuning types on FashionIQ.

A.2.2 Traininig on combination of Laion-CIR-Template and Laion-CIR-LLM

In this section, we combine the Laion-CIR-Template dataset with Laion-CIR-LLM dataset to create a new dataset called Laion-CIR-Combined that consists of approximately 32k samples. Subsequently, we train our proposed TransAgg model on the combined dataset and the results are shown in Table 8 and Table 9. It can be observed that using more data tends to lead to better results.

	Recall@K				$\text{Recall}_{\text{Subset}}$@K		
Fine-tuning	K=1	K=5	K=10	K=50	K=1	K=2	K=3
✘	35.28	64.46	76.53	92.46	65.37	83.37	92.12
only text enc.	37.87	68.88	79.60	93.86	69.79	86.09	93.93
both	36.71	67.06	77.82	93.65	66.25	84.09	93.10

Table 8: Results on the CIRR test set.

Fine-tuning	Shirt		Dress		TopTee		Average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
✘	30.86	49.02	24.49	45.51	32.94	54.05	29.43	49.53
only text enc.	34.45	53.97	30.24	51.91	38.40	59.51	34.36	55.13
both	35.03	52.94	32.52	54.34	37.89	59.15	35.15	55.48

Table 9: Results on the FashionIQ validation set.

A.2.3 Comparison with state-of-the-art

Here, we compare our proposed approach with several existing zero-shot composed image retrieval methods on CIRR and FashionIQ datasets, as shown in Table 10 and Table 11.

Method	Zero-shot eval	# Training triplets	Recall@K				Recall$_{\text{Subset}}$@K		
			K=1	K=5	K=10	K=50	K=1	K=2	K=3
Pic2Word [9]$_{\text{CVPR'2023}}$	✔	-	23.90	51.70	65.30	87.80	-	-	-
PALAVRA [3]$_{\text{ECCV'2022}}$	✔	-	16.62	43.49	58.51	83.95	41.61	65.30	80.94
SEARLE-XL-OTI [2]$_{\text{arXiv'2023}}$	✔	-	24.87	52.31	66.29	88.58	53.80	74.31	86.94
CompoDiff w/T5-XL [5]$_{\text{arXiv'2023}}$	✔	18m	19.37	53.81	72.02	90.85	28.96	49.21	67.03
CASE Pre-LaSCo.Ca. [6]$_{\text{arXiv'2023}}$	✔	360k	35.40	65.78	78.53	**94.63**	64.29	82.66	91.61
TransAgg(Laion-CIR-Template)	✔	16k	**38.10**	68.42	79.08	93.51	**70.34**	**86.42**	**94.28**
TransAgg(Laion-CIR-LLM)	✔	16k	36.71	67.83	79.03	93.86	66.03	83.66	92.50
TransAgg(Laion-CIR-Combined)	✔	32k	37.87	**68.88**	**79.60**	93.86	69.79	86.09	93.93
CLRPLANT w/OSCAR [7]$_{\text{ICCV'2021}}$	✘	-	19.55	52.55	68.39	92.38	39.20	63.03	79.49
ARTEMIS [4]$_{\text{ICLR'2022}}$	✘	-	16.96	46.10	61.31	87.73	39.99	62.20	75.67
CLIP4CIR [1]$_{\text{CVPRW'2022}}$	✘	-	38.53	69.98	81.86	95.93	68.19	85.64	94.17
BLIP4CIR+Bi [8]$_{\text{arXiv'2023}}$	✘	-	40.15	73.08	83.88	96.27	72.10	88.27	95.93
CASE [6]$_{\text{arXiv'2023}}$	✘	-	48.00	79.11	87.25	97.57	75.88	90.58	96.00

Table 10: Comparasion on CIRR test set. The best and second-best numbers are shown in red and blue respectively.

Method	Zero-shot eval	# Training triplets	Shirt		Dress		TopTee		Average	
			R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Pic2Word [9]$_{\text{CVPR'2023}}$	✔	-	26.20	43.60	20.00	40.20	27.90	47.40	24.70	43.70
PALAVRA [3]$_{\text{ECCV'2022}}$	✔	-	21.49	37.05	17.25	35.94	20.55	38.76	19.76	37.25
SEARLE-XL-OTI [2]$_{\text{arXiv'2023}}$	✔	-	30.37	47.49	21.57	44.47	30.90	51.76	27.61	47.90
CompoDiff w/T5-XL [5]$_{\text{arXiv'2023}}$	✔	18m	**38.10**	52.48	**33.91**	47.85	**40.07**	52.22	**37.36**	50.85
TransAgg(Laion-CIR-Template)	✔	16k	32.83	52.31	27.67	49.38	35.70	58.08	32.07	53.26
TransAgg(Laion-CIR-LLM)	✔	16k	32.92	52.16	28.56	49.58	36.82	58.59	32.77	53.44
TransAgg(Laion-CIR-Combined)	✔	32k	34.45	**53.97**	30.24	**51.91**	38.40	**59.51**	34.36	**55.13**
CLRPLANT w/OSCAR [7]$_{\text{ICCV'2021}}$	✘	-	17.53	38.81	17.45	40.41	21.64	45.38	18.87	41.53
ARTEMIS [4]$_{\text{ICLR'2022}}$	✘	-	21.78	43.64	27.16	52.40	29.20	54.83	26.05	50.29
CLIP4CIR [1]$_{\text{CVPRW'2022}}$	✘	-	39.99	60.45	33.81	59.40	41.41	65.37	38.32	61.74
BLIP4CIR+Bi [8]$_{\text{arXiv'2023}}$	✘	-	41.76	64.28	42.09	67.33	46.61	70.32	43.49	67.31
CASE [6]$_{\text{arXiv'2023}}$	✘	-	48.48	70.23	47.44	69.36	50.18	72.24	48.79	70.68

Table 11: Comparasion on FashionIQ validation set. The best and second-best numbers are shown in red and blue respectively.

A.3 Explainability

In this section, we present some interpretable examples. As shown in the first row of Figure 1, the relative caption demands a focus on the head of the dog. Correspondingly, the model concentrates most of its attention on the dog. In the second row of Figure 1, the relative caption requires bent knees and knee pads to be worn. Consequently, the model prioritizes the knee and knee pads as the main focal points.

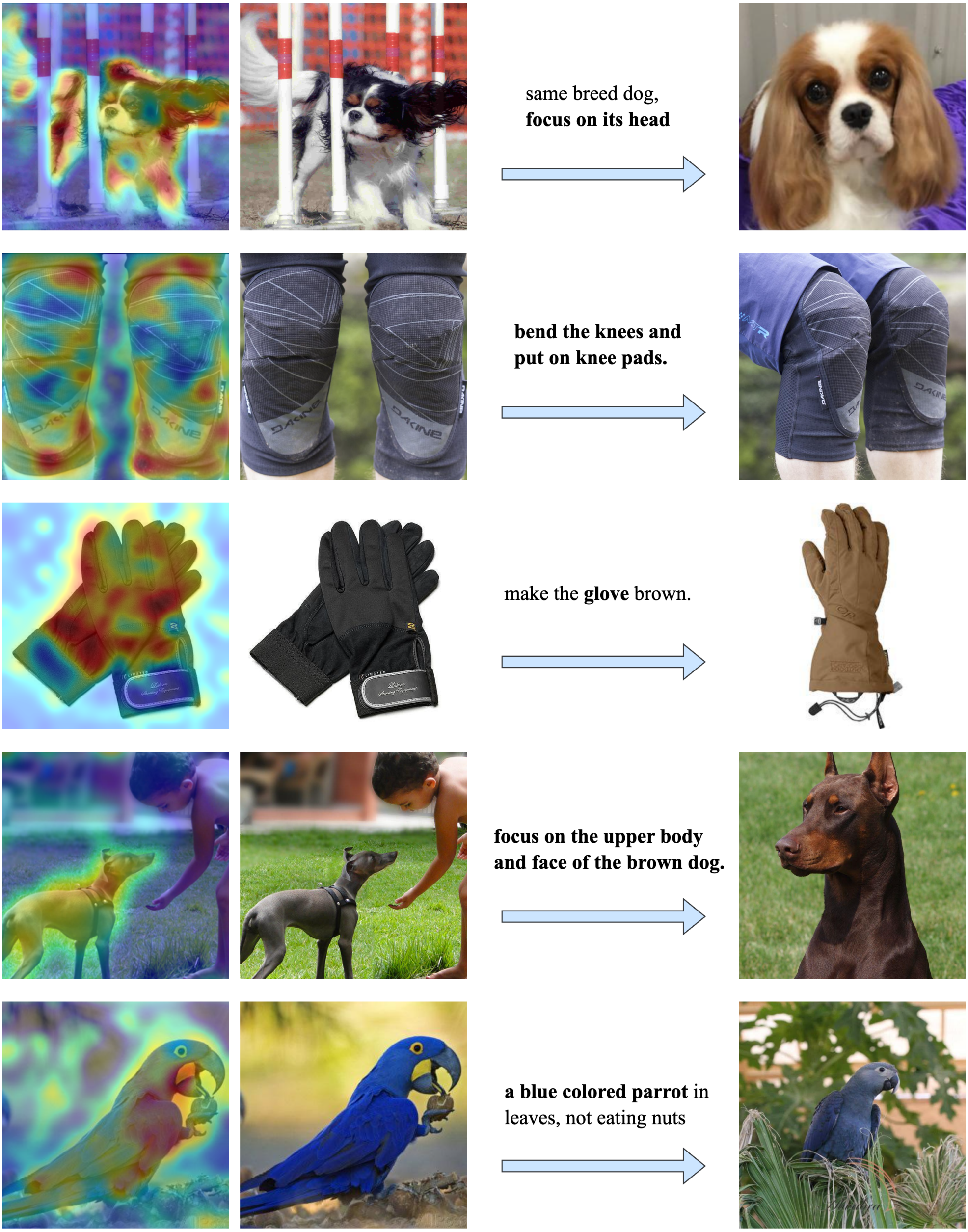
same breed dog,
focus on its head
bend the knees and
put on knee pads.
make the **glove** brown.
focus on the upper body
and face of the brown dog.
a blue colored parrot in
leaves, not eating nuts

Figure 1: Explainability heatmaps for CIR task. From left to right are the heatmap, reference image, relative caption and the target image. The heatmap is calculated through the attention between the bolded token in the relative caption and other image patches.

References

[1] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *CVPR Workshops*, 2022.

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

[3] Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, and Yuval Atzmon. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *ECCV*, 2022.

[4] Ginger Delmas, Rafael S Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *ICLR*, 2022.

[5] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023.

[6] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and early fusion for composed image retrieval. *arXiv preprint arXiv:2303.09429*, 2023.

[7] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, 2021.

[8] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bidirectional training for composed image retrieval via text prompt learning. *arXiv preprint arXiv:2303.16604*, 2023.

[9] 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 *CVPR*, 2023.