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

Supplementary Material

A. Attention Visualization

To demonstrate that optimizing ConText-CIR for the Text Concept-Consistency and contrastive objectives improves the ability of the framework to ground text features to their relevant image features, we visualize the attention of concepts to the image query at various points in the training process. Figure 1 gives the concept attentions, averaged across noun phrases, for 5 concept-image pairs at 5 points during training (downwards is later in training). We observe that attention becomes significantly more focused to the image patches that text concepts refer to as training progresses, with very little spurious attention. In all cases, fully trained ConText-CIR shows precise attentions to the image patches that a noun phrase refers to, while earlier models either ground attention to irrelevant objects or spurious image features.

B. Synthetic Dataset Pipeline

good4cir employs a three-stage pipeline to generate precise, high-quality annotations for CIR model training. In Stage 1, the model curates a list of key objects and descriptors from the query image. During Stage 2, the model derives a similar list from the target image by comparing it against the list of objects from the query image, ensuring consistency and making modifications when necessary. Stage 3 compares both lists to generate a list of fine-grained difference captions that describe the addition, removal, and modification of objects from the query to the target image. Figure 3 and Figure 2 give examples of synthesized examples from the good4cir framework. More details about the prompts used for generation and qualitative results validating the usefulness of good4cir's generated data for CIR models may be found in the accompanying paper [3].

C. Additional Benchmarks

We also present zero-shot composed image retrieval metrics on FashionIQ [6] and the domain conversion on ImageNet-R [1, 2] introduced by Pic2Word [5]. The FashionIQ evaluation is stratified by classes of clothing items as defined by the dataset, and we report R@10 and R@50 on the test set. As described by Pic2Word, we use the 200 classes and domains outlined by ImageNet and ImageNet-R and perform domain level evaluation on retrieval results. The retrieval prompt is generated by picking a class from the set {`cartoon, origami, toy, sculpture`}. We report average R@10 and R@50 across target domains.

Table 1 demonstrates that ConText-CIR is consistently state-of-the-art, demonstrating convincing performance gains across both FashionIQ an ImageNet-R's domain translation task.

Method	FashionIQ								ImageNet-R	
	Shirt		Dress		Toptee		Average		Average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
LinCIR	20.92	42.44	29.10	46.81	28.81	50.18	26.28	46.49	-	-
iSEARLE-XL	28.75	47.84	22.51	46.36	31.31	52.68	27.52	48.96	15.52	33.39
CIReVL	29.49	47.40	24.79	44.76	31.36	53.65	28.55	48.57	23.75	43.05
CoVR-BLIP-2	34.26	56.22	41.22	59.32	38.96	59.77	38.15	58.44	-	-
ours	**38.52**	**61.09**	**46.81**	**65.43**	**50.28**	**71.05**	**45.20**	**65.86**	**25.62**	**44.84**

Table 1. Zero-shot composed image retrieval metrics on FashionIQ and ImageNet-R.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[2] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 1

[3] Pranavi Kolouju, Eric Xing, Robert Pless, Nathan Jacobs, and Abby Stylianou. good4cir: Generating detailed synthetic captions for composed image retrieval, 2025. 1

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

[5] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. *CVPR*, 2023. 1

[6] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *CVPR*, 2021. 1

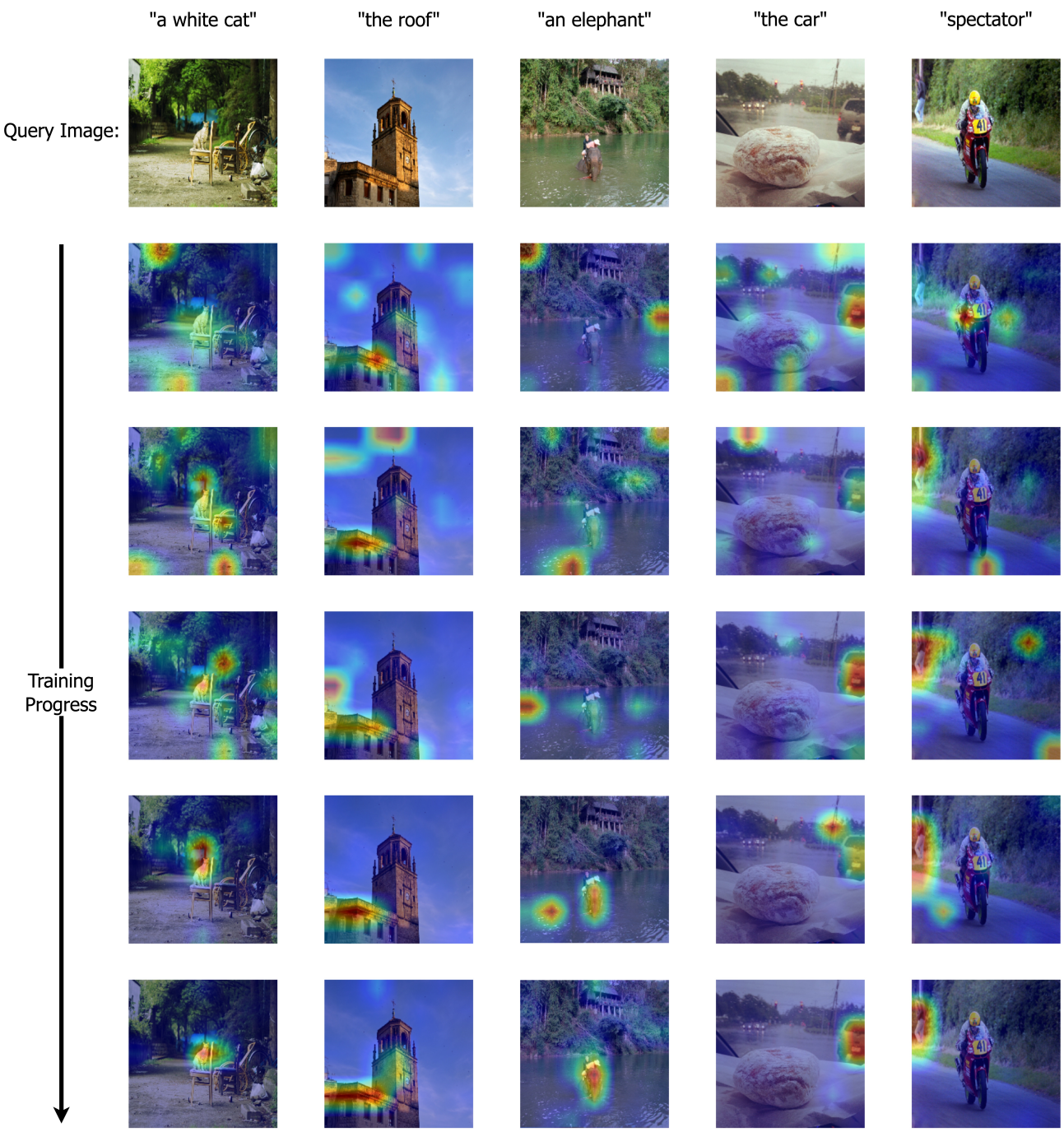


Figure 1. Visualization of attentions averaged over noun phrases to image queries. Concept attentions become more grounded to relevant image features with training.

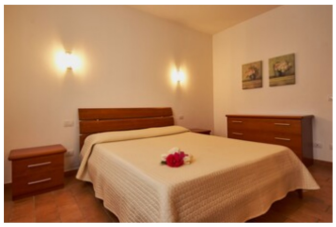
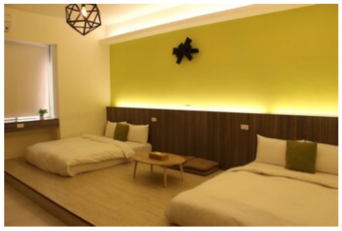
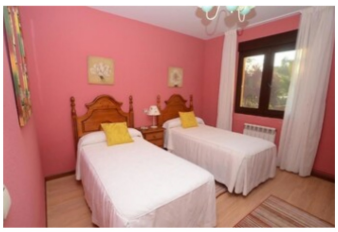
Query Image	Target Image	Text Differences
		**Difference Captions:** ~ Remove the bathtub, Add a toilet with a white porcelain material, standard size, compact shape, attached tank, and chrome handle ~ Place a white rectangular bath mat with a soft texture and non-slip backing on the floor ~ Add decorative plants with green leaves in small white pots and place them on the countertop ~ Install a wooden door with light brown color, modern handle, and smooth surface ~ Ensure the door has a hinged design ~ Ensure the decorative plants are artificial and neatly arranged ~ Ensure the added door has a clean appearance and light finish ~ Remove the shower curtain
		**Difference Captions:** ~ Replace the tan-colored blanket with a beige-tan blanket on the bed, Remove the two pillows from the bed ~ Change the nightstand to have a medium brown color instead of tan ~ Modify the nightstand to have two drawers instead of a single drawer, ~ Replace the simple headboard of the bed with a headboard that has medium-height horizontal slats and a smooth texture ~ Add a wooden table with three drawers, matching the nightstand and bed frame, positioned against the wall ~ Include two pieces of floral artwork with green and beige coloring, framed and mounted on the wall side by side above the wooden table ~ Install two mounted lights on the walls ~ Ensure no bedskirt is visible around the bed ~ Make sure that the bedspread on the bed is centered
		**Difference Captions:** ~ Lower the bed closer to the ground ~ Remove the visible bed frame ~ Reduce the visible pillows to two per bed ~ Use green decorative pillows instead of dark ones ~ Place one decorative pillow centrally on each bed ~ Attach the headboard to the wall instead of the bed ~ Extend the headboard horizontally across the wall ~ Integrate lighting fixtures into the headboard ~ Introduce light yellow walls for a brighter appearance ~ Ensure the window has a Roman blind with a light-colored cover ~ Add a small, light-colored wooden table between the beds ~ Include a piece of small, abstract artwork with a black, geometric design on the wall ~ Incorporate a modern, dark-colored, geometric-shaped lighting fixture hanging from the ceiling above the foot of the first bed
		**Difference Captions:** ~ Change the white bedding to include a small pattern ~ Place the beds side by side to create a matching set ~ Add an additional pink pillow to each bed ~ Modify the headboard to have intricate cuts at the top ~ Replace the nightstand with one that has three drawers and squared legs ~ Swap out the floral bedside lamp for a modern white-colored one ~ Center the large rectangular window on the wall and ensure it has a white frame ~ Adjust the curtains to have a sheer white outer layer and a solid pink inner layer ~ Transform one wall into a pink accent wall while keeping other walls white ~ Darken the wooden flooring to a medium brown color and use hardwood material instead of laminate ~ Add a light brown wooden bench against the wall for additional sitting or storage space

Figure 2. Example generated text differences for the Hotel-CIR dataset using our synthetic data generation pipeline.

Query Image	Target Image	Text Differences
		**CIRR:** have an avocado in the background **Ours:** ~ Remove the glass of chocolate smoothie with banana slice garnish ~ Replace the banana with an avocado sliced in half, showing light green flesh and a large seed ~ Change the chocolate smoothie to a yellow smoothie with a creamy texture, pale yellow color, and tiny bubbles on top ~ Swap the transparent glass for a mason jar with an embossed logo, open top, and add a green and white spiral-patterned straw ~ Add a circular sticker label with multicolored text reading 'DE LIGHT FUL' affixed near the top ~ Introduce a flat, light brown, smooth wooden surface with natural grain lines as the background
		**CIRR:** more yellow and the light come from the left side **Ours:** ~ Remove the ancient cathedral and add a church with sandstone color, a dome with a cross, and detailed carvings ~ Replace the round tower with a conical dome featuring arched openings and decorative columns, topped with a cross ~ Modify the surrounding landscape to include a rocky cliff with a reddish-brown hue and patches of vegetation ~ Add a stone wall with rectangular stones, partially obscured by plants, acting as a boundary marker ~ Introduce trees and bushes with green foliage and varied species to add contrast to the stone structures
		**CIRR:** instead of rabbits dog is sitting in grass **Ours:** ~ Remove both guinea pigs from the scene ~ Introduce a tan and white short-haired dog with alert ears and a black nose in a reclining position ~ Add a bright green fern with feathery leaves behind the dog to enrich the backdrop ~ Include a cylindrical tree trunk with brown bark beside the dog to enhance the natural setting ~ Alter the vivid green grass to appear in tandem with the new elements, supporting a cohesive natural environment
		**CIRR:** The same type of cabinet but blue color **Ours:** ~ Change the cabinet to a light blue color with ornate trim and floral motifs, and replace the existing handles with round brass handles in a vintage style ~ Remove the ceramic pitcher, yellow bowl, stacked plates, wooden shelves, scale, ceramic mug, jug, yellow pot, and light source from the image ~ Add a crown molding to the top of the cabinet with intricate carvings and a center floral detail to enhance its aesthetic appeal ~ Incorporate wooden legs with a curved shape and matching blue color to support the cabinet and complement its design ~ Include lower cabinet doors with ornate carvings and floral embellishments in a light blue color to add elegance to the bottom section
		**CIRR:** Shot from a different angle **Ours:** ~ Replace the off-white whipped cream topping with white fluffy peaks that are generously spread ~ Remove the mini eggs and add bright red cherries with glossy finishes and stems positioned on the peaks ~ Introduce chocolate shavings with a dark brown color and fine texture scattered on top ~ Replace the jelly layer with a drizzled dark brown chocolate sauce with a glossy finish ~ Incorporate pecans with a rough texture and brown color scattered on top for added crunch ~ Substitute fruit pieces with banana slices having a pale yellow color and soft texture layered within ~ Integrate thin, vivid red strawberry slices layered visibly for a juicy appearance

Figure 3. Example original reference texts from CIRR [4] and generated text differences from the CIRR$_R$ dataset generated using our synthetic data generation pipeline.