A Sanity Check on Composed Image Retrieval

Supplementary Material

A. Details of Generating Different Semantic Subsets

Addition&Negation&Change&Background. The generation details for these four semantic subsets are already covered in Section 3.1 of the main paper. In this section, we focus on the specifics of how reference image captions are sampled for each semantic aspect. For the Addition aspect, captions are sampled from the COCO validation set. For Negation, we utilize both the validation and test sets of Flickr30K [15]. Sampling for the Change aspect is done from the validation set of NoCaps [1], while for Background, we draw from the validation set of the CC3M [13] dataset.

Given the image caption for the reference image, we prompt Mixtral-8x7B-Instruct-v0.1 [8] to simultaneously generate a relative caption and the caption of the target image with the following prompt:

> I need you to perform one reasonable editing step on an image. I will provide the caption of the original image, and I need you to specify the following: instruction and the caption of edited image. The semantic aspect of the instruction you generate is {1}. Reference image caption: {2}

where the first placeholder refers to the corresponding semantic aspect, namely addition, negation, change or background, while the second placeholder refers to the reference image caption.

Cardinality. We adopt *"a real-life image of* {*num*} {*noun*}*."* as a template caption, where num is chosen from ten numbers ranging from 1 to 10, and noun is selected from the object category in COCO and ImageNet [4]. Then, we employ it to drive the diffusion model to generate corresponding images. After generating the images, we manually adjust the number of objects in the images. Subsequently, we select three images with different numbers of objects belonging to the same category: one serves as a reference image, one as a target image, and the other as a hard negative image. The relative captions are randomly sampled from the templates in Figure 1. In these templates, the position of the placeholder corresponds to the respective quantity.

Adjust the object count to match {}.	Adjust the object count from {} to {}.
Modify the total objects to {}.	Modify the total objects from {} to {}.
Transition the object quantity to {}.	Transition the object quantity from {} to {}.
Transform the number of objects to {}	Transform the number of objects from {} to {}.
Update the object tally to {}.	Shift the object count from {} to {}.
Alter the quantity of objects to become {}.	Alter the quantity of objects from {} to {}.
Convert the object number to {}.	Convert the object number from {} to {}.
Set the number of objects to {}.	Set the number of objects from {} to {}.
Change the number of objects to {}.	Change the number of objects from {} to {}.

Figure 1. The template for cardinality relative captions.

Complex. We utilize HQ-Edit [7] as the data source for the complex subset. HQ-Edit leverages GPT-4V to provide high-quality reference image captions, relative captions, and target image captions. We select samples with relative captions of more than 25 words as our caption source.

B. Prompts of Our Multi-round Evaluation Pipeline

Prompts for MLLM. We use MLLM and LLM to simulate user interaction. Specifically, we first prompt MLLM to generate the caption for both candidate and target images. For CIRR, CIRCO, and FISD, we employ the prompt: *Give me a short*

and precise English description of the image. For the FashionIQ dataset, which emphasizes clothing, we adopt the following prompt: *Give a short and precise English description of the clothes*.

Prompts for LLM. After getting the captions, we prompt LLM to generate relative captions to describe the difference between the candidate and target images. For CIRR and CIRCO, we use the prompt as follows:

> I will give you two sentences, one of which is reference image caption and one is target image caption. I need you to give me an instruction to transition from reference image to target image. The type of instruction you give can be cardinality, addition, negation, direct addressing, compare&change, comparative, conjunction, spatial relation&background, viewpoint. The instruction you give does not need to have additional information, be detailed and highlight the key points. The instruction you generate is about 30 words. Reference Image Caption: {1}. Target Image Caption: {2}. Instruction:

For FashionIQ, we employ the following prompt that focuses on the semantic aspects potentially contained within the FashionIQ dataset as mentioned in [10].

> I will give you two sentences, one of which is reference image caption and one is target image caption. I need you to give me an instruction to transition from reference image to target image. The type of instruction you give can be addition, negation, direct addressing, compare&change, comparative, conjunction. The instruction you give does not need to have additional information, be brief and highlight the key points. The instruction you generate is about 10 words. Reference Image Caption: {1}. Target Image Caption: {2}. Instruction:

For FISD, we use the following prompt for addition, negation, cardinality, background, and change semantic aspect, where the first placeholder refers to the specific semantic aspect, the second placeholder denotes the reference image caption, and the third one represents the target image caption.

> I will give you two sentences, one of which is reference image caption and one is target image caption. I need you to give me an instruction to transition from reference image to target image. The type of instruction you give can only be {1} and can not contain other semantics. The instruction you give does not need to have additional information, be detailed and highlight the key points. The instruction you generate is about 10 words. Reference Image Caption: {2}. Target Image Caption: {3}. Instruction:

Regarding the complex instruction semantic aspect of the FISD benchmark, which may encompass diverse semantic instructions, we do not limit the semantic type. The prompt is as follows:

> I will give you two sentences, one of which is reference image caption and one is target image caption. I need you to give me an instruction to transition from reference image to target image. The instruction you give does not need to have additional information, be detailed and highlight the key points. The instruction you generate is about 30 words. Reference Image Caption: {1}. Target Image Caption: {2}. Instruction:

C. Evaluation Details

When conducting multi-round evaluations on the FISD, which distinguishes between different semantic aspects, we ensure that the feedback provided by the user simulator contains only the relevant semantic content. For supervised models, when evaluating on the CIRR, CIRCO, and FISD benchmarks, we utilize their official checkpoints, which have been fine-tuned on the CIRR dataset. This is because CIRR, CIRCO, and FISD encompass real-life image scenarios. Conversely, for the evaluation on the FashionIQ dataset, we employ the official checkpoint specifically fine-tuned on the FashionIQ.

D. Detailed Experimental Results

D.1. Additional Validation Experiments for FISD

Table 1. Comparison of performance on negation and overall semantics in the CIRR validation set.

Method	Negation	All
	R@1	R@1
Pic2Word	17.81	23.25
Context-I2W	20.55	26.96
SPRC	31.51	55.39

As shown in Section 4.2 of the main paper, CIR models perform poorly in handling negation semantics on FISD. To determine whether this issue is caused by the gap between synthetic and natural images, we use Llama3-8B to select and manually verify queries with negation semantics from the CIRR validation set. The experimental results are presented in Table 1, revealing that even with natural images, CIR models struggle with negation compared with other semantic aspects. This is consistent with our findings on FISD.

D.2. Additional Multi-round Evaluation Results

In this section, we present the multi-round evaluation results of various CIR models using the Recall@K metric. The experimental results are shown in Table 2.

E. Detailed Analysis of User Study

Analysis of User Study Experimental Results. As shown in Section 4.3 of the main paper, in the evaluation of the CIRR and FashionIQ benchmarks, feedback from real users is sometimes less effective compared to that from user simulators. As illustrated in Figure 2, this disparity may stem from the fact that feedback from real users tends to be concise and lacks details, whereas simulators provide more comprehensive and detailed feedback.

Implementation Details of the User Study. Each time, we present the user with two images: one reference/candidate image and one target image. We ask the user to provide reasonable feedback that describes the differences between the two images. We set the maximum number of interaction rounds to 5, meaning the user provides feedback up to 4 times in one multi-round interaction session. We offer users a wage of $15 per hour. Note that, our user study is conducted internally within the laboratory, with no potential risks. Additionally, we do not store any feedback provided by the users.

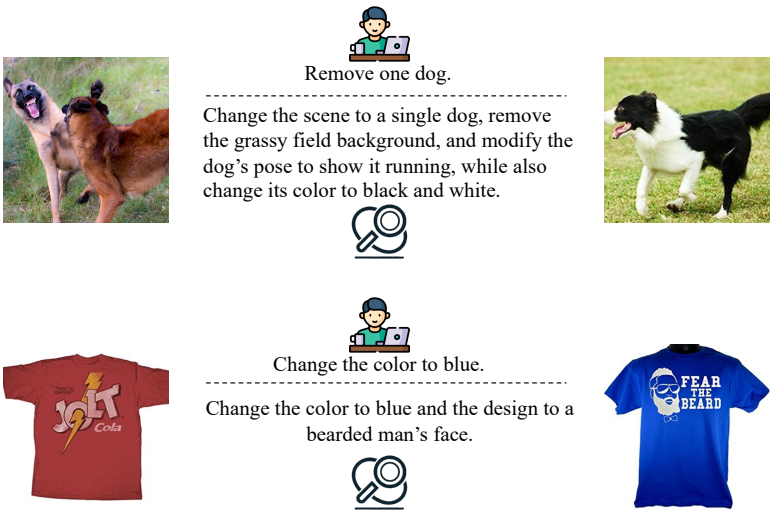


Figure 2. Some examples of feedback from both users and user simulators.

Table 2. Multi-round evaluation on various state-of-the-art CIR models across a range of benchmarks.

Method	FashionIQ-Dress		FashionIQ-Shirt		FashionIQ-Toptee		CIRR		FISD
	Recall@10	Recall@50	Recall@10	Recall@50	Recall@10	Recall@50	Recall@1	Recall@5	Recall@1
Round 1									
Pic2Word [12]	20.00	40.20	26.20	43.60	27.90	47.40	23.25	51.42	21.33
Context-I2W [14]	23.10	45.30	29.70	48.60	30.60	52.90	26.96	56.59	27.83
LinCIR [6]	20.92	42.44	29.10	46.81	28.81	50.18	25.09	54.41	32.42
TransAgg [9]	30.24	51.91	34.45	53.97	38.40	59.51	38.79	69.58	34.75
CLIP4CIR [3]	39.46	64.55	44.41	65.26	47.48	70.98	45.37	78.47	47.17
BLIP4CIR+Bi [11]	42.09	67.33	41.76	64.28	46.61	70.32	42.36	75.48	45.17
SPRC [2]	49.18	72.43	55.64	73.89	59.35	78.58	55.39	84.26	50.08
SPN4CIR [5]	50.57	74.12	57.70	75.27	60.84	79.96	56.47	85.29	55.83
Round 3									
Pic2Word [12]	21.81	44.42	33.12	50.93	31.72	51.50	40.28	66.01	64.58
Context-I2W [14]	30.84	54.69	43.62	62.66	44.97	65.83	43.00	69.31	50.08
LinCIR [6]	24.94	48.34	45.04	62.41	39.32	60.38	42.86	70.65	50.67
TransAgg [9]	40.65	64.95	55.84	74.48	60.12	78.79	61.83	85.77	77.67
CLIP4CIR [3]	46.16	69.71	59.52	79.39	61.35	80.62	67.26	93.95	74.08
BLIP4CIR+Bi [11]	49.78	73.23	49.36	71.54	55.69	76.03	52.31	84.62	69.67
SPRC [2]	57.91	78.78	71.05	85.92	71.49	87.35	80.63	96.22	77.08
SPN4CIR [5]	63.16	82.85	74.39	88.52	75.57	89.39	83.11	96.48	79.33
Round 5									
Pic2Word [12]	22.71	43.18	34.79	51.52	32.02	51.30	57.43	68.93	83.42
Context-I2W [14]	31.73	54.39	46.61	65.60	47.42	68.18	56.66	71.30	68.92
LinCIR [6]	25.43	46.75	46.61	65.01	40.03	60.28	58.19	72.95	68.42
TransAgg [9]	42.69	64.90	60.99	78.56	66.29	84.14	78.02	88.23	88.33
CLIP4CIR [3]	46.11	70.15	62.22	81.11	62.06	82.20	82.35	94.91	81.83
BLIP4CIR+Bi [11]	49.98	71.99	51.37	71.59	56.45	76.29	61.97	85.22	78.83
SPRC [2]	57.11	76.80	73.21	86.90	73.28	87.61	88.38	97.46	83.33
SPN4CIR [5]	64.20	83.74	77.53	91.41	79.30	91.59	89.26	97.54	84.92

F. Additional Failure cases in Current Public Benchmark

In this section, we show the failure cases of the current single-round CIR models on the CIRR dataset and the FashionIQ dataset. As shown in Figure 3 and 4, it can be seen that these failure cases are basically caused by insufficient correlation between the composed query and the target image.

G. Additional Examples of Our Proposed Benchmark

In this section, we present additional examples of our proposed FISD benchmark, illustrated in Figure 5 and 6. These figures display six data types: cardinality, addition, negation, change, background, and complex instruction, respectively.

H. Qualitative Analysis

In this section, we show qualitative examples in multi-round CIR. Figure 7, 8 and 9 illustrate successful results for two to four rounds. These examples demonstrate how our multi-round system incrementally approaches the target image by continuously incorporating user feedback. In Figure 10, we showcase some failure cases on the SPRC model under a multi-round setting. These failures primarily occur when receiving either excessively vague feedback or certain semantic feedback types (e.g., negation) that current CIR models can not effectively process. These observations highlight the critical need for both (i) improving the CIR model's overall performance and (ii) developing cleaner, more standardized benchmarks.

I. Discussion on Inference Cost

Although the CIR model achieves significant performance gains through multi-round interaction, it inevitably introduces additional overhead. However, in practical applications, there are effective strategies to maximize revenue while minimizing time costs during multi-round explorations. For instance, by dividing the queries into sub-domains, and applying the multi-round exploration only to specific sub-domains that yield better gains, we can reduce the cost of multi-round exploration by shifting from an instance-wise to a sub-domain-wise approach. Moreover, it is also possible to narrow down the candidate pool based on the first-round search, which can also exponentially reduce the time cost in the search space.

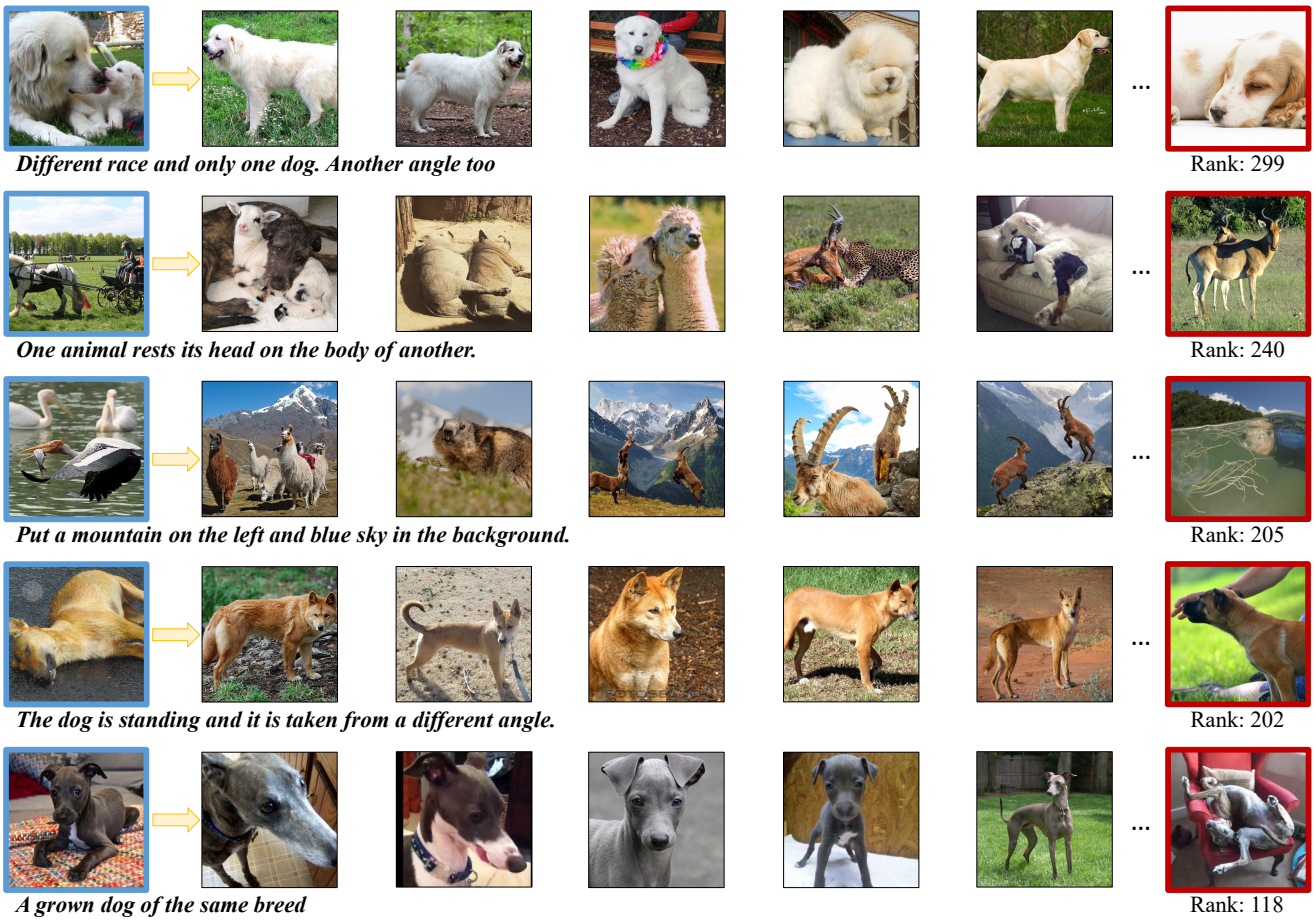
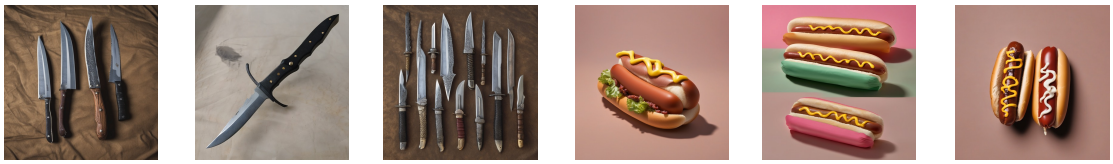


Figure 3. Failure cases of current CIR models on CIRR validation set. The reference image is marked with blue borders and the target image with red borders.

Figure 4. Failure cases of current CIR models on FashionIQ validation set. The reference image is marked with blue borders and the target image with red borders.

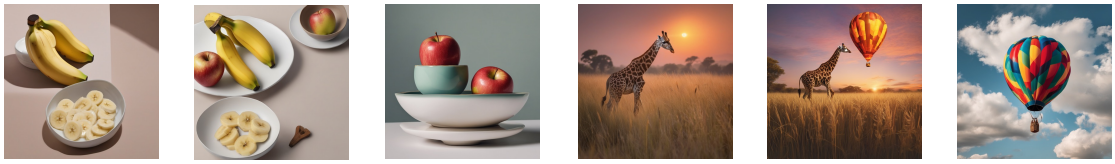


Shift the object count from four to one. Convert the object number to three.

Set the number of object from one to three. Update the object tally to eight.

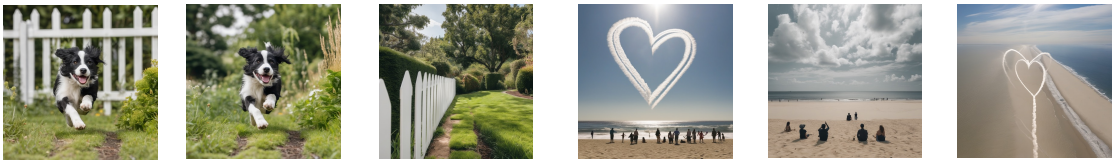
(a) Cardinality

Add a cat standing on top of a black car. Add two rubber ducks floating in the bathtub.

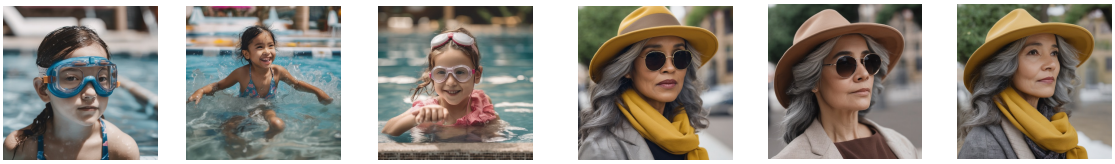


Add two apples. Add a vibrant hot air balloon floating in the sky.

(b) Addition



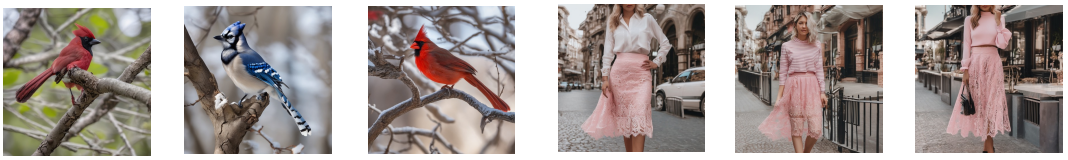
Exclude the white fence. Remove the large heart written in the air.



Exclude the goggles from the image. Exclude the woman's scarf.

(c) Negation

Figure 5. Some examples of our proposed benchmark, the order of the images is a reference image, target image, and hard negative image.



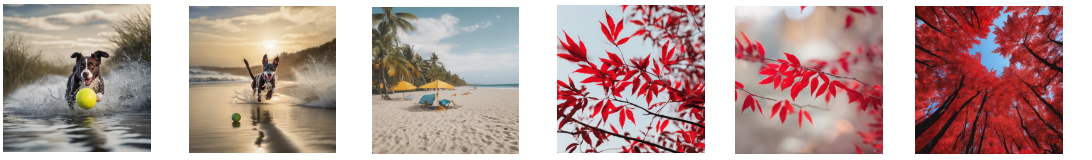
Change the color of the jay from red to blue.

Change the pattern of the woman's shirt from solid to striped.

Change the basketball to a soccer ball.

Replace the watermelon juice with pineapple juice

(e) Change



Change the background to a bench scene.

Make the background more blurred to emphasize the red leaves

Remove the busy background.

Change the background to a desolate and barren landscape.

(f) Background

Transition the scene to **nighttime**, add a starry sky with a **full moon**, and illuminate the scene with **warm artificial lighting** from the buildings and street lamps.

Change the setting to **a winter scene** with **a red wooden house**, replace the greenery with **snow-covered pine trees**, and dress the children **in winter clothing**.

Replace the sunflower with **a cherry blossom** tree with full pink blooms, adjust the **background to include more cherry blossoms**, and modify the sky to have a **soft pink hue**.

Transform the elderly gentleman into **a young boy**, change the clothing to a **casual t-shirt with a graphic print, shorts, and sneakers.** Replace the walking cane with **boy's right hand resting on the bench.**

(g) Complex

Figure 6. Some examples of our proposed benchmark, the order of the images is a reference image, target image, and hard negative image.

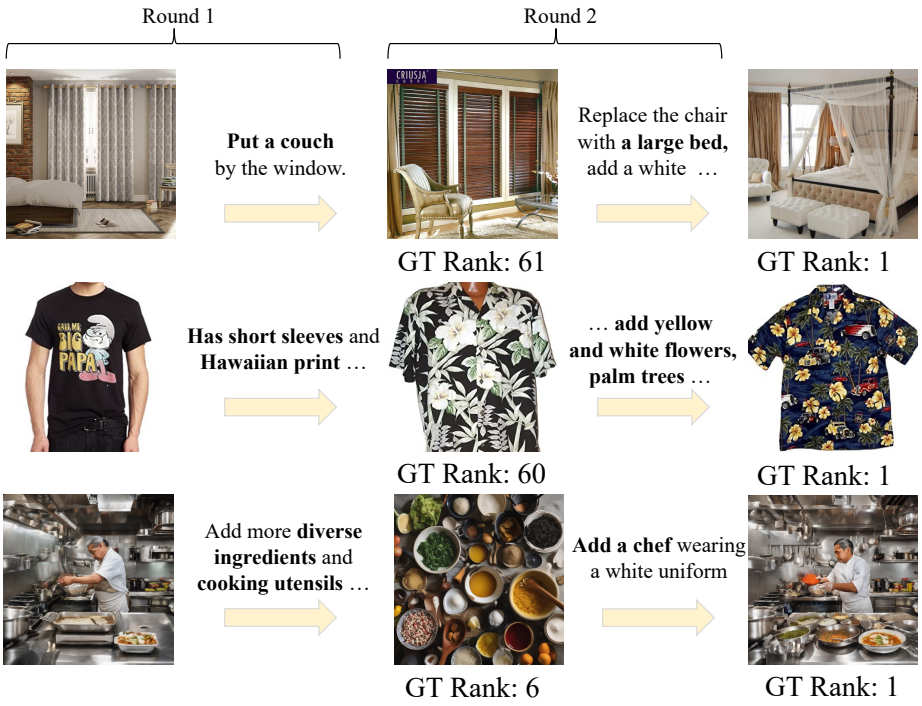


Figure 7. Qualitative results for two rounds.

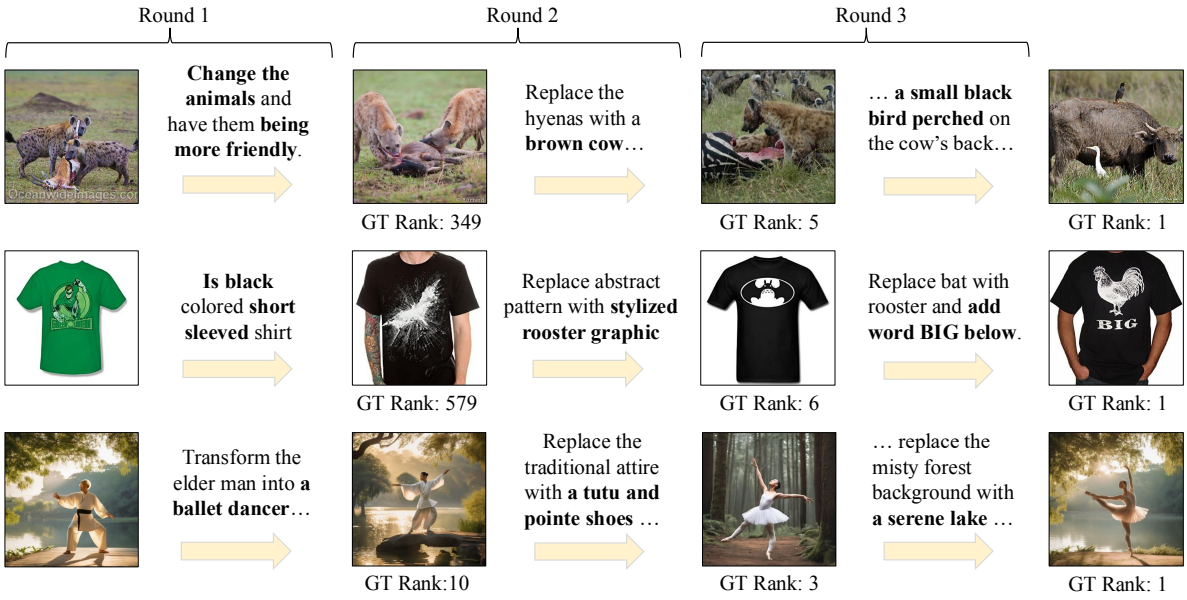


Figure 8. Qualitative results for three rounds.

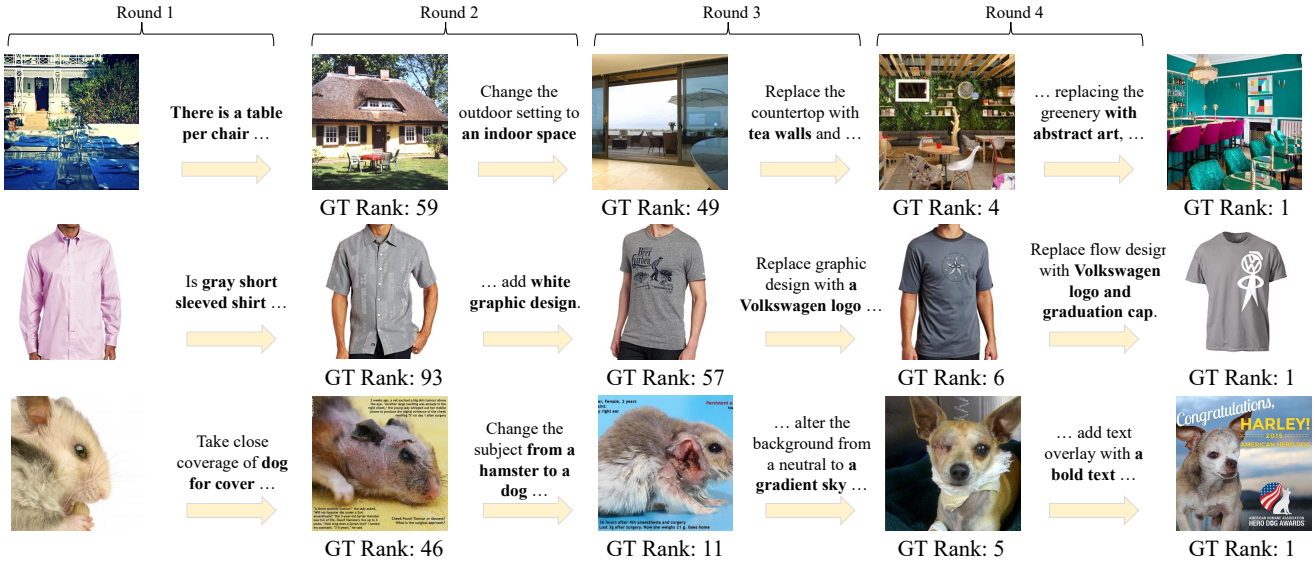


Figure 9. Qualitative results for four rounds.

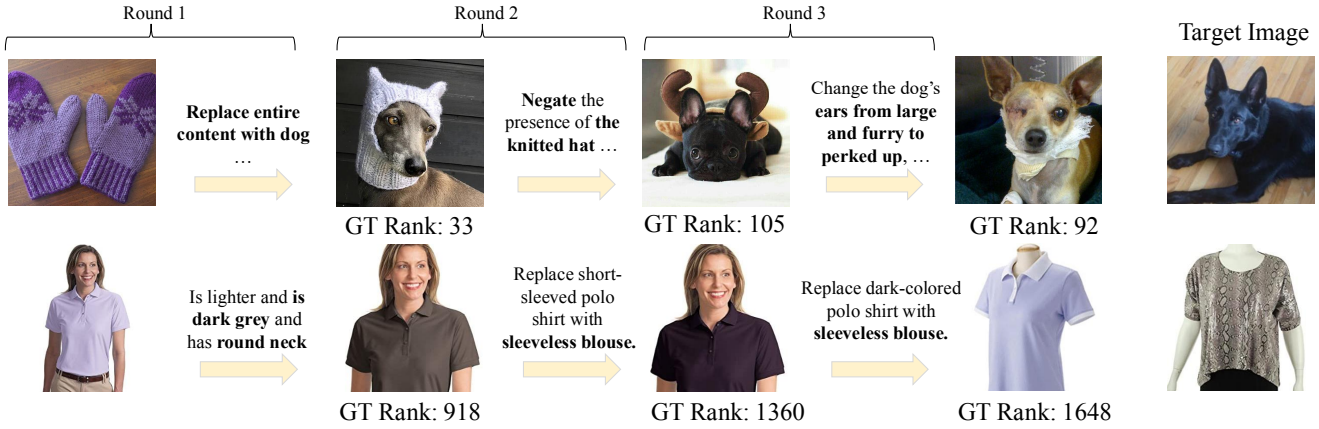


Figure 10. Failure cases in multi-round CIR.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *ICCV*, 2019. 1

[2] Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval. In *ICLR*, 2024. 4

[3] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Composed image retrieval using contrastive learning and task-oriented clip-based features. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023. 4

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[5] Zhangchi Feng, Richong Zhang, and Zhijie Nie. Improving composed image retrieval via contrastive learning with scaling positives and negatives. *arXiv preprint arXiv:2404.11317*, 2024. 4

[6] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval. In *CVPR*, 2024. 4

[7] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024. 1

[8] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 1

[9] Yikun Liu, Jiangchao Yao, Ya Zhang, Yanfeng Wang, and Weidi Xie. Zero-shot composed text-image retrieval. In *BMVC*, 2023. 4

[10] 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. 2

[11] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning. In *WACV*, 2024. 4

[12] 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. 4

[13] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 1

[14] 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 *AAAI*, 2024. 4

[15] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *TACL*, 2014. 1