# Adapting In-context Generation for Enhanced Composed Image Retrieval

## Supplementary Material

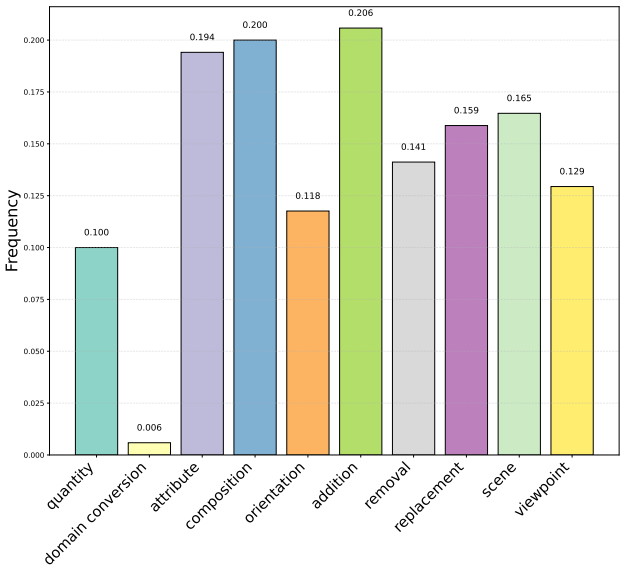



Figure 1. **Visualization of different semantic categories in the relative captions** (since some relative captions encompass multiple semantic categories, the sum of proportions does not equal 1).

## 1. More Details about Domain-Adaptive In-context Generation

This section provides a detailed elaboration of the DAIG implementation, with the complete pipeline illustrated in Figure 4. It begins with ***the construction of in-context descriptions***. After acquiring a small set of domain-specific triplets (e.g., 32-shot), a captioner such as Qwen2.5-VL [2] is employed to generate descriptions for all images, yielding $T_r$ and $T_t$, which, combined with the provided relative caption $T_c$, form the in-context descriptions. This step requires minimal time and computational resources, as captioning is performed only on a small number of samples.

Once these in-context samples are obtained, ***in-context generative tuning*** is conducted via CIR-LoRA, enabling a text-to-image (T2I) model, specifically FLUX 1.dev [4], to learn domain and task priors. This phase typically takes a few hours ($15,000$ training steps with a batch size of 1). Upon completion, a domain-adapted T2I model is obtained, ready for the data augmentation stage of DAIG.

In the augmentation stage, a carefully designed instruction template is used to guide an LLM, i.e., Qwen2.5-Instruct [9], to generate diverse textual triplets. In our template, we incorporate both $object$ and $edit$. The distribution of $edit$ is illustrated in Figure 1. $object$ is sampled from two predefined sets constructed by GPT-5 [1] and tailored for the fashion domain and real-life scenarios, respectively. An illustration of these objects is presented below, with the full lists omitted due to space constraints, but all project materials, including code, templates and the object lists, will be made publicly available to ensure reproducibility. After obtaining 20K triplets, T2I generation is performed by inputting them into the domain-adapted T2I model for in-context synthesis. This produces well-aligned, domain- and task-adapted CIR triplets, which are subsequently used for training robust CIR models.

> ***Real-world Object List***: pencil, clock, dog, llama, telephone, giraffa, marmot, touchpad, peppers, beaker, cake slice, floor lamp, bottle cap, museum, scarf, spoon, pineapple, tree, lime, face mask, vizsla dog, apron, sandals, cake, pelican, headphones, lobster, bird, lion, chopsticks...
> ***Fashion-domain Object List***: black, A-line dress, V-neck neckline, sleeveless, midi length, pleated skirt, striped, navy/white, fit-and-flare dress, scoop neckline, short sleeves, knee-length, side pockets, white sheath dress, square neckline, cap sleeves, ruched waist, floral print, Mario graphic, pink...

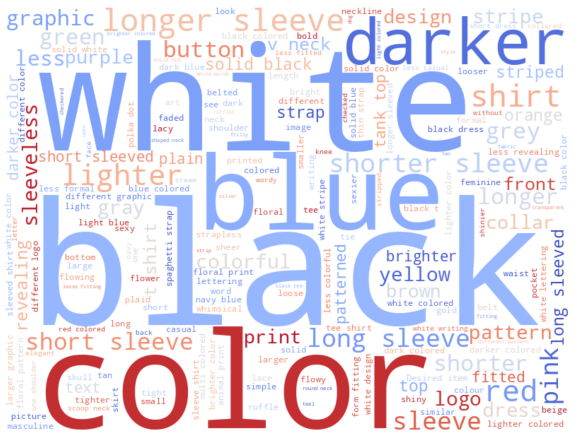
Finally, we produce 20K domain-adapted triplets, with a word cloud visualization of their relative captions shown in Figure 2. Additional samples produced by DAIG in real-life scenarios and the fashion domain are visualized in Figures 5 and 6, respectively.

## 2. Benchmarks for Composed Image Retrieval

FashionIQ [8] is established to advance research on conversational interfaces for online fashion shopping, aiming to move beyond traditional keyword-based retrieval systems that often fail to capture user intent. The dataset focuses on interactive fashion image retrieval, a canonical form of the CIR task. It comprises 30,134 triplets derived from 77,684 fashion images, organized into three main categories: Dress, Shirt and Toptee. Each triplet consists of a reference image, a relative caption describing the intended modification and a target image. In addition to the triplet structure, the dataset also provides product descriptions and attribute-level annotations, which support more fine-grained evaluation. The official split follows a 6:2:2 ratio for training, validation and testing. Representative examples are visualized in Figure 3b.

CIRR [5] is proposed to broaden CIR research to open-domain retrieval scenarios, addressing the limitation that


(a) Word cloud visualization of DAIG under real-life scenarios.




(b) Word cloud visualization of DAIG in the fashion domain.

Figure 2. **The relative captions word cloud visualization of triplets generated by DAIG.**

most existing benchmarks, such as FashionIQ, are confined to a single domain. CIRR is constructed by first sampling a large set of visually similar natural images from NLVR [7], using ResNet-152 [3] pre-trained on ImageNet [6] as the similarity backbone. Pairs of highly similar images are then manually annotated with relative captions to form triplets. In total, CIRR comprises 36,554 annotated triplets, which are randomly divided into training, validation, and testing splits with an 8:1:1 ratio. Evaluation is conducted through a remote server submission system, ensuring fair benchmarking across methods. Despite its strengths, CIRR is not without challenges: some relative captions contain vague or redundant descriptions, and the dataset includes a considerable number of false negatives (FNs). Visual examples are provided in Figure 3a.

## 3. Qualitative Results

We visualize several inference results on CIRR [5] and FashionIQ [8], as shown in Figure 7. The top four cases demonstrate that integrating DAIG effectively enhances the recognition and retrieval performance of SPRC. The improved capability to recognize modifications such as "two antelopes" and "white with a black logo" stems from DAIG's sampling of diverse semantic categories during sample generation (Figure 1), enabling fine-grained modifications including quantity, attribute and composition.

Naturally, our method also has certain limitations. For instance, in the last two cases shown in Figure 7, the left example demonstrates incorrect analysis of the dog's orientation, which may be attributed to inherent fine-grained errors in current T2I models, an issue that is difficult to avoid at this stage. The right example highlights an annotation problem, where the relative caption "more revealing" is overly ambiguous, leading to false negatives. This issue is relatively common in current CIR benchmarks, primarily because the annotation process relies solely on paired reference-target images without access to global context.

## References


[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report, 2024. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[3] 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. 2

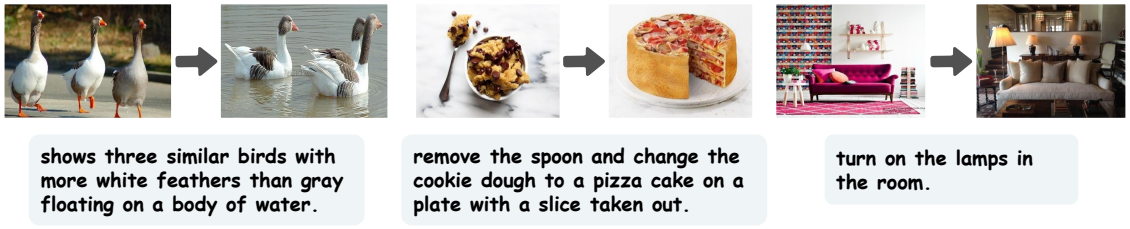
[4] Black Forest Labs. Flux: Official inference repository for flux.1 models. https://github.com/black-forest-labs/flux, 2024. Accessed: 2024-11-12. 1

[5] 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. 1, 2, 7

[6] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 2

[7] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs, 2019. 2

[8] 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




(a) CIRR exampels, covering a wide range of real-life scenarios.


(b) Fashion IQ exampels. Left: Shirt; Middle: Dress; Right: Toptee.

Figure 3. **Examples of CIR benchmarks.** In all instances, the left side represents the reference image, while the right side is the target image obtained through the given relative caption.

feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021. 1, 2, 7

[9] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, et al. Qwen2.5 technical report, 2025. 1

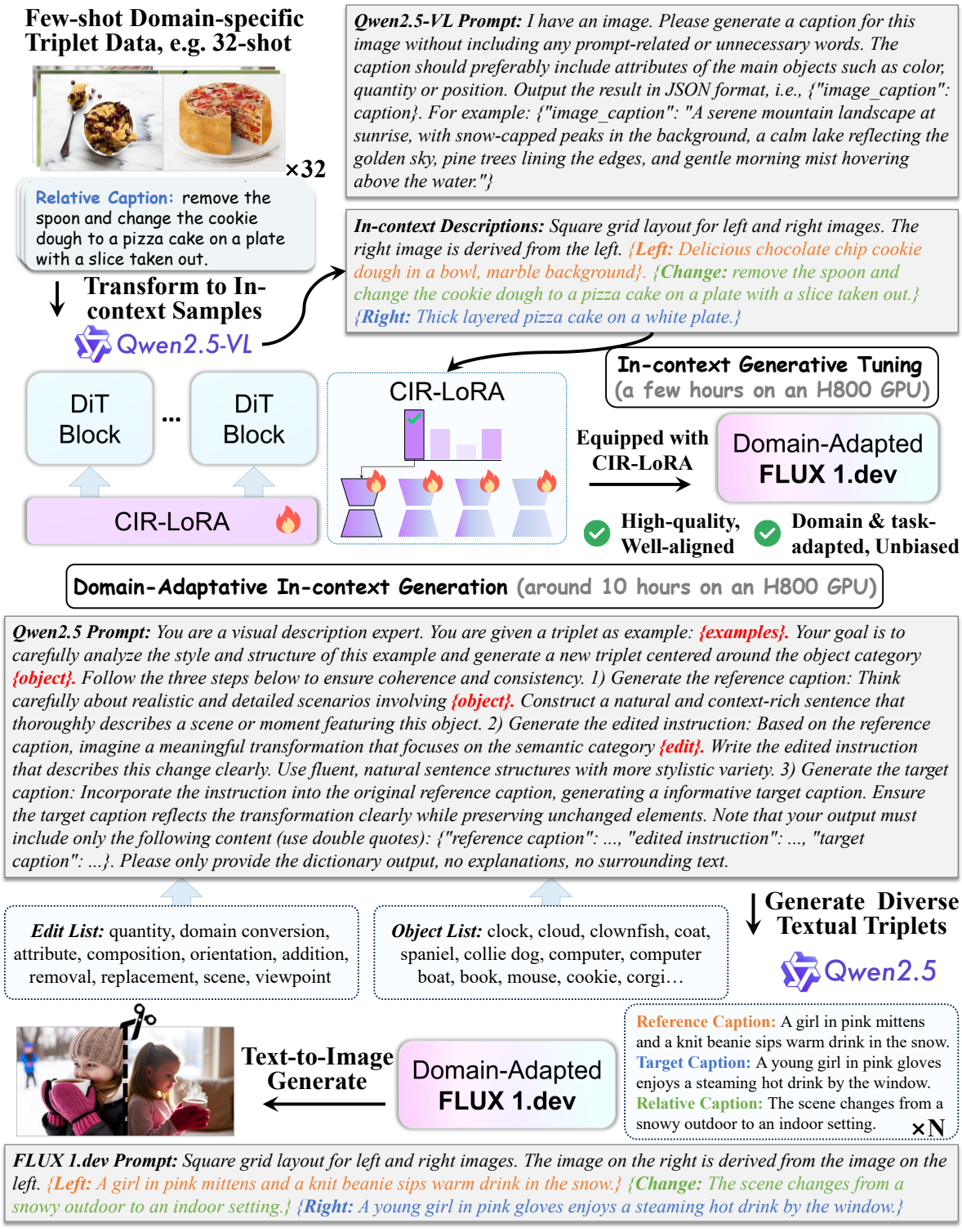
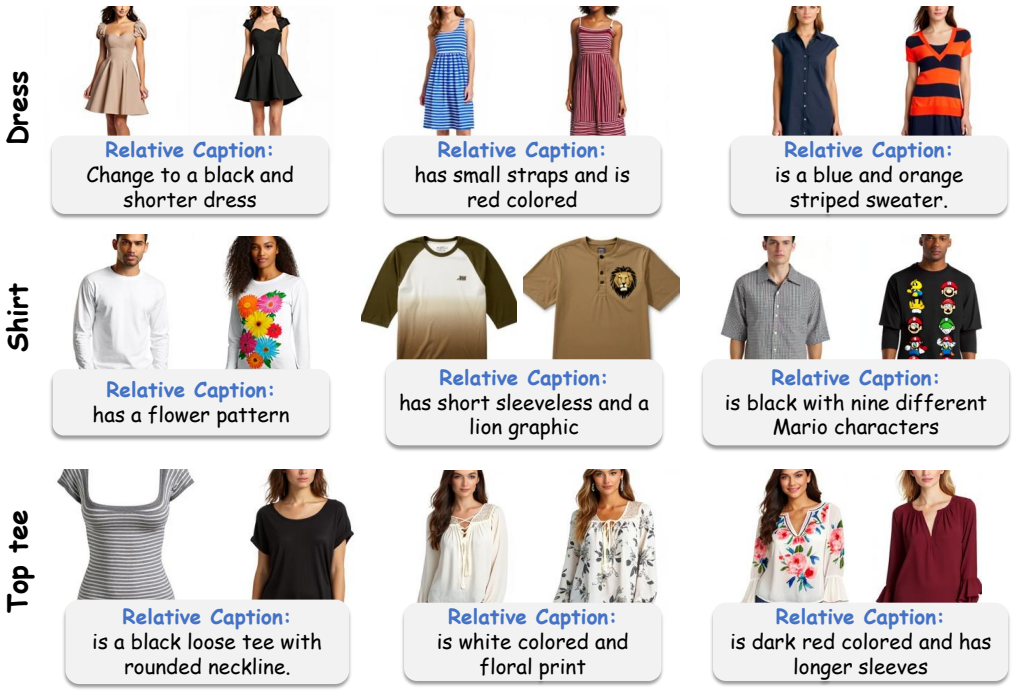



Figure 4. **An overview of the DAIG pipeline,** encompassing the construction of in-context samples, in-context generative tuning and domain-adaptive in-context generation.

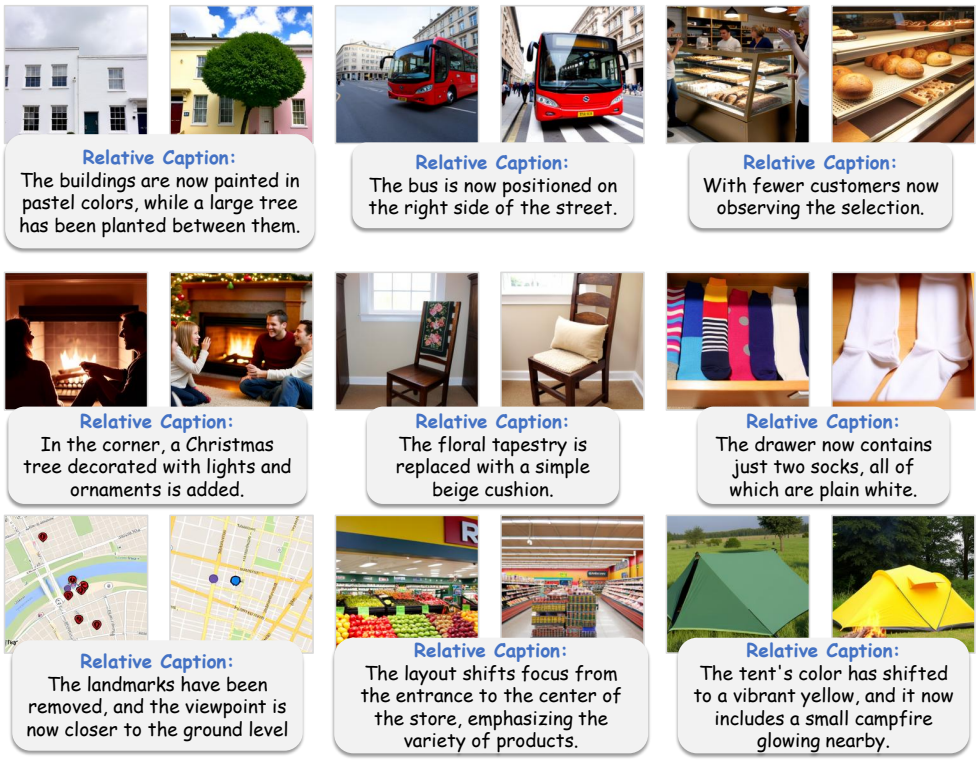

(a) Triplets generated by training DAIG on FashionIQ.


(b) Triplets generated by training DAIG on CIRR in real-life scenarios.

Figure 5. **More triplet examples generated by on DAIG in the fashion domain and real-life scenarios.**




(a) Triplets generated by training DAIG on FashionIQ.




(b) Triplets generated by training DAIG on CIRR in real-life scenarios.

Figure 6. **More triplet examples generated by on DAIG in the fashion domain and real-life scenarios.**


Figure 7. **Qualitative results on CIRR [5] and FashionIQ [8],** along with failure case analysis. We present the top-5 predictions, with correct items annotated by green bounding boxes.