RealGeneral: Unifying Visual Generation via Temporal In-Context Learning with Video Models

Supplementary Material

Figure 1. Examples from our constructed dataset for subject-driven text-to-image generation. Each image pair is carefully evaluated based on key subject components and multiple similarity metrics, ensuring high subject consistency.

1. Dataset Details

Similar to the approach of OminiControl [3], we construct a dataset of paired 1024×1024 images for subject-driven text-to-image generation. Our dataset comprises 2000 common object categories (*e.g.*, belt, flower, etc.), forming 220,000 image pairs. Firstly, we use ChatGPT-4o to generate a diverse set of object names. For each object, we further generate two distinct textual descriptions that depict the same subject in different scenarios. These descriptions are then used as input prompts to FLUX1. [1], which produces image pairs featuring the same object.

To ensure high subject similarity between the image pairs, we implement a three-stage evaluation process using ChatGPT-4o. In the first stage, ChatGPT-4o identifies the key components of the subject. In the second stage, it conducts a detailed description of these components. In the final stage, similarity scores are assigned across multiple dimensions (*e.g.*, overall shape and structure, materials, and colors). We compute the average score for each pair and retain only those with an average score exceeding 4. A representative subset of our constructed dataset is shown in Fig. 1.

Furthermore, we evaluate the quality of the Subject200k [3] using the same assessment process. The filtered-out data, which do not meet the similarity threshold, are presented in Fig. 2.

2. More Results

In this section, we present more generation results using RealGeneral.

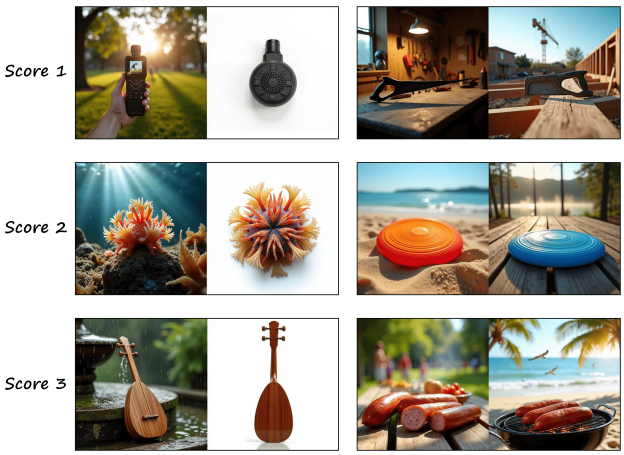

Figure 2. Examples of image pairs filtered out due to low subject similarity scores. Each row presents two pairs of images with similarity scores of 1, 2, and 3, respectively.

Figure 3. Ablation study for our proposed module. The full version presents more subject consistency while demonstrating comparable textual controllability.

Fig. 3 presents the qualitative results of our proposed module. Furthermore, Fig. 4 presents the qualitative results of various attention masks, as depicted in the paper. Fig. 5 showcases more results on the DreamBench [2] for customization task.

Fig. 6 shows more results of canny-to-image task. Fig. 7 shows more results of depth-to-image task. Fig. 8 shows more results of inpainting and coloring tasks. Fig. 9 shows more results of image-to-depth and deblurring tasks.

References

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

[2] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine

Figure 4. Qualitative results of various attention masks.

tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1

[3] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024. 1

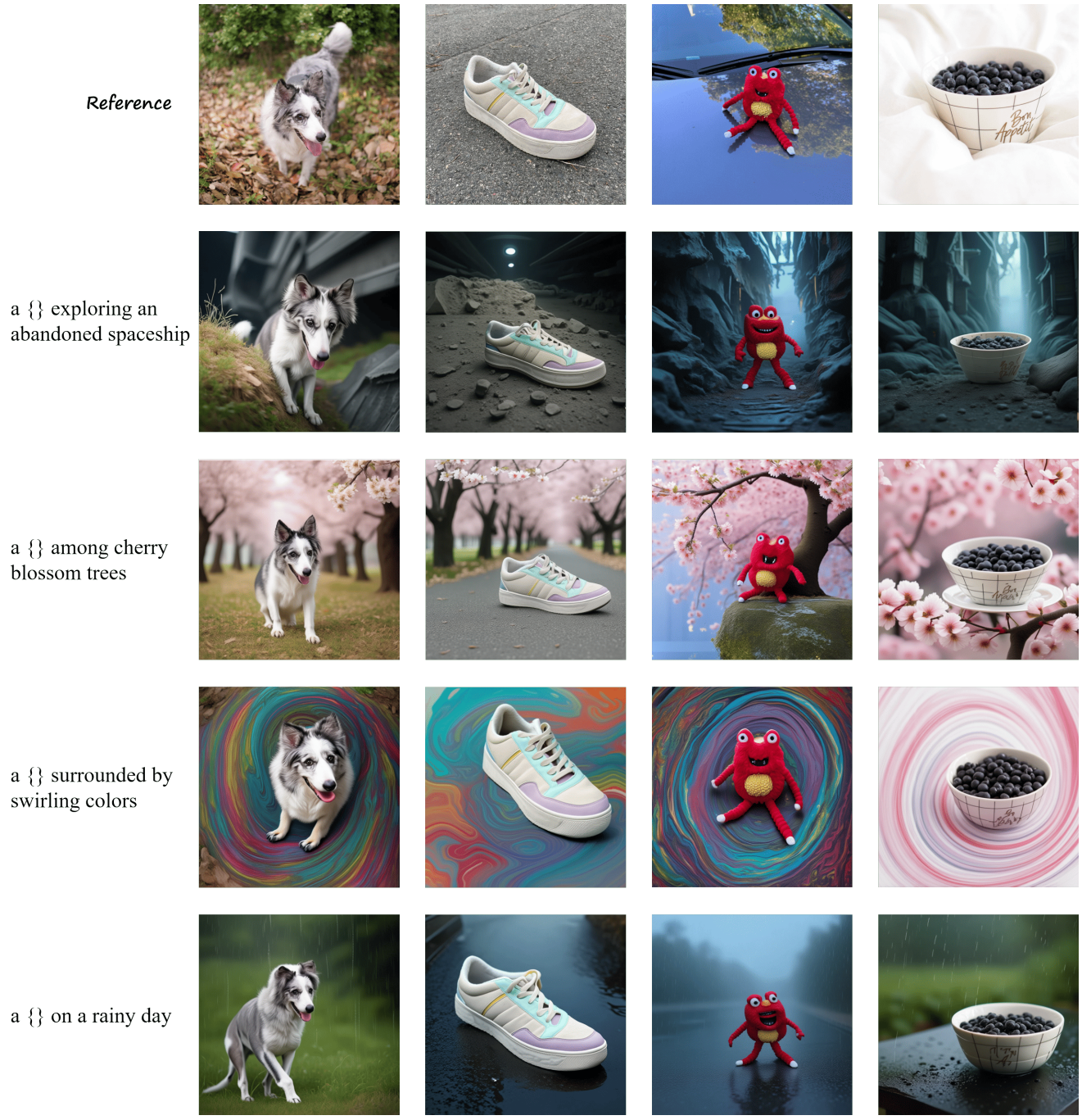

Figure 5. The results of customization task.

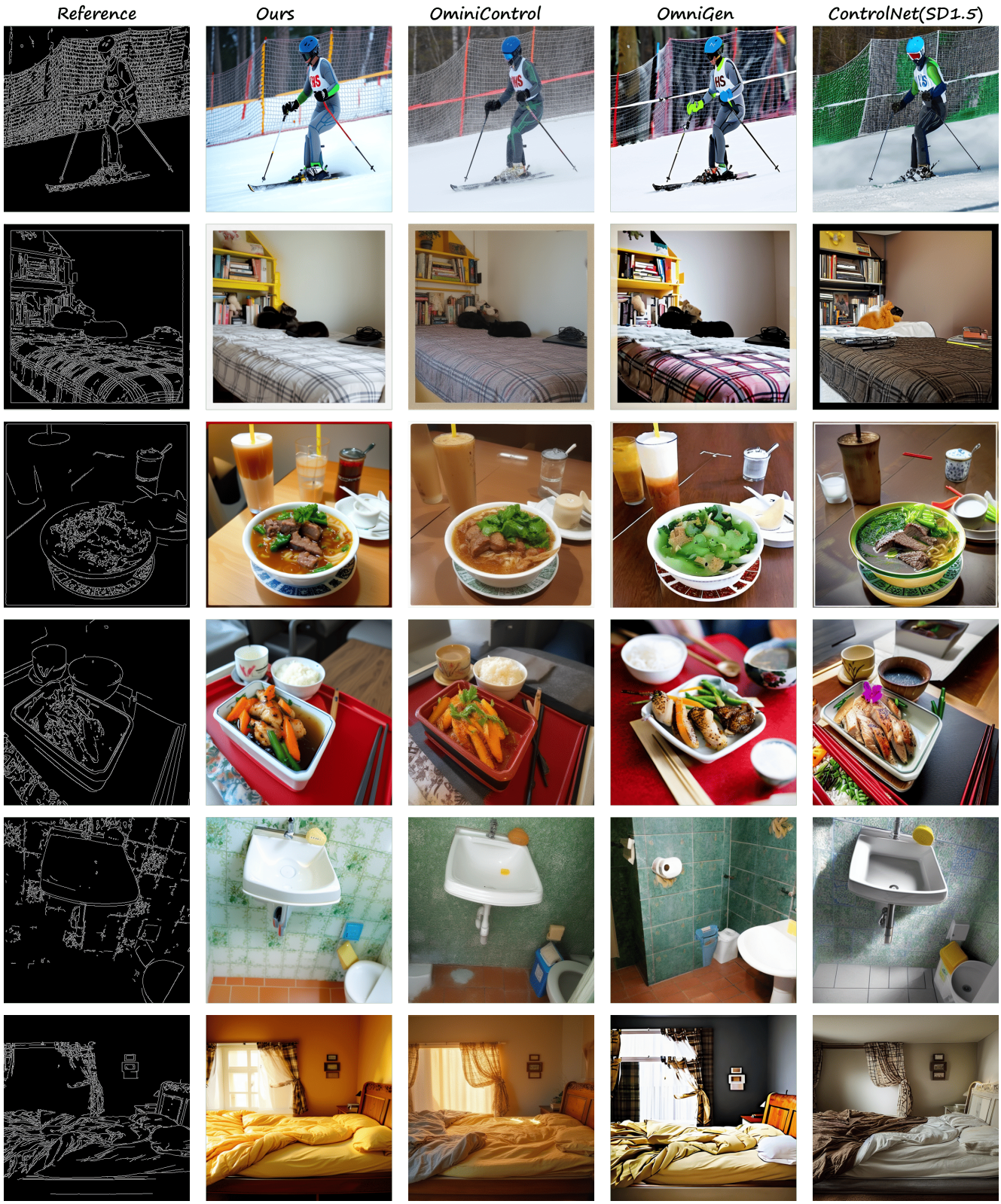

Figure 6. The results of canny-to-image task.

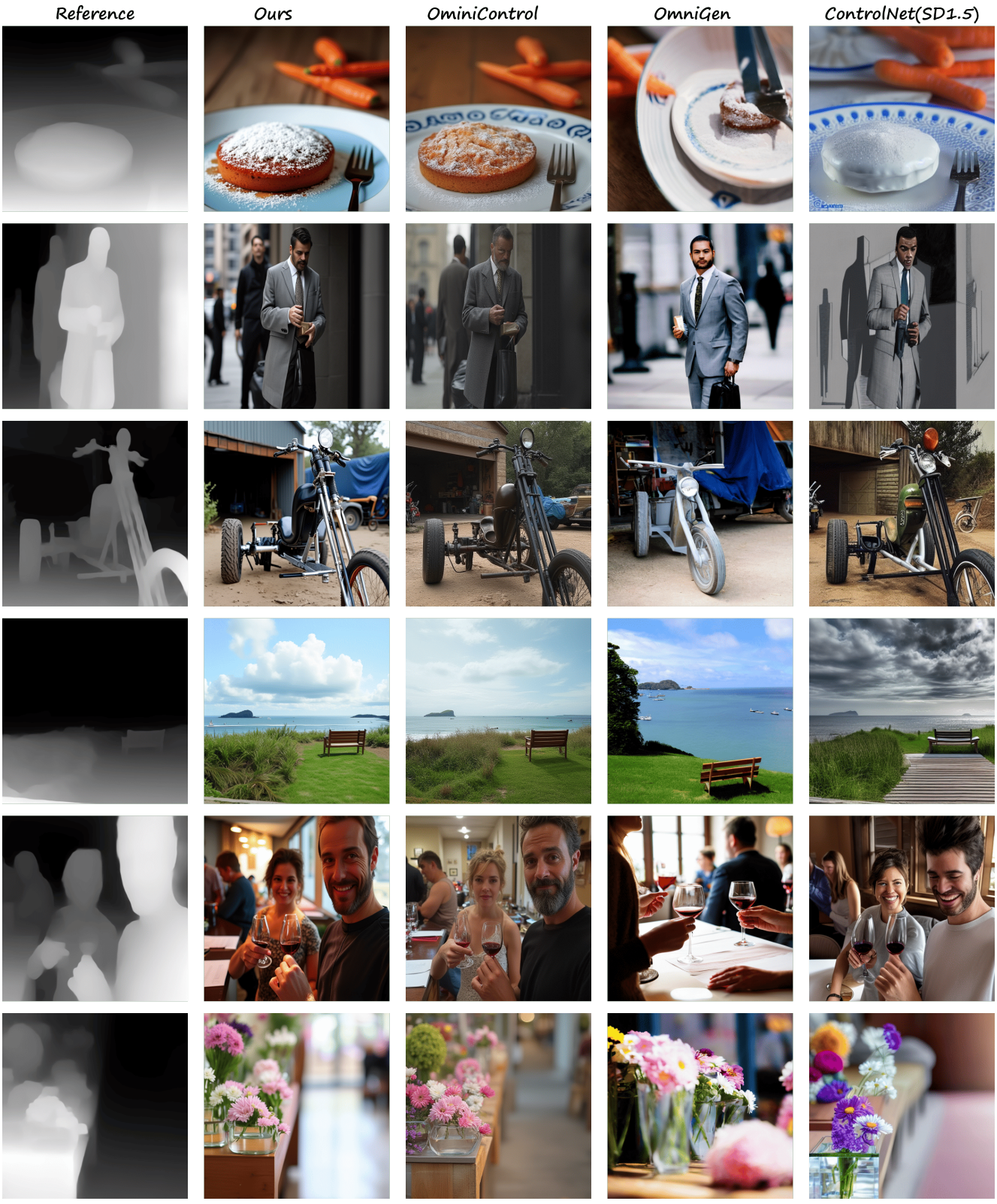

Figure 7. The results of depth-to-image task.

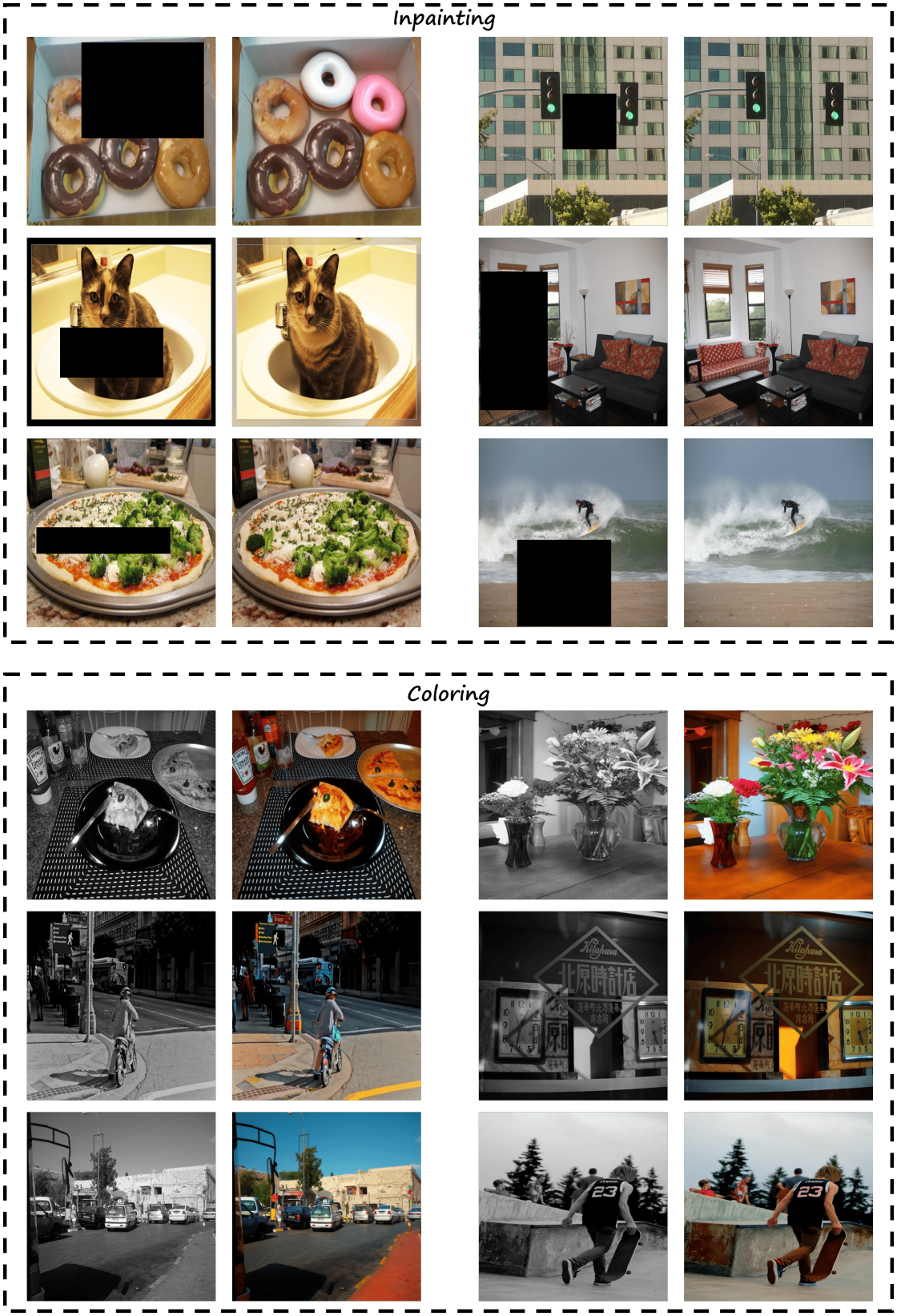

Figure 8. The results of inpainting and coloring tasks.

Figure 9. The results of image-to-depth and deblurring tasks.