Tuning-Free Image Customization with Image and Text Guidance Supplemental Materials

Pengzhi Li[1*], Qiang Nie[2,3*], Ying Chen[3], Xi Jiang[4], Kai Wu[3], Yuhuan Lin[3], Yong Liu[3], Jinlong Peng[3], Chengjie Wang[3], Feng Zheng[4†]

[1]Tsinghua University [2] HKUST (GZ) [3]Tencent Youtu Lab [4] Southern University of Science and Technology [5] Research Institute of Multiple Agents and Embodied Intelligence, Peng Cheng Laboratory

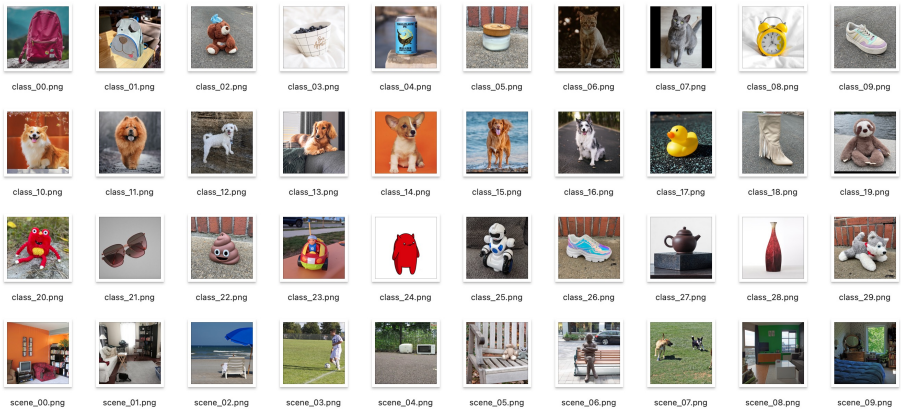

Fig. 1: We present our collected benchmark. The top three columns show the 30 different categories of objects from DreamBooth [6]. The bottom row displays representative scene images.

1 Benchmarks

Our approach achieves image customization with text and image guidance simultaneously, so existing image customization evaluation datasets, such as the one proposed by [4,6], cannot adequately assess our method. Therefore, we base our work on DreamBooth [6], maintaining consistency with its settings, and use 30 categories of objects, including both objects and living subjects. In order to generate scene images, we consider various indoor and outdoor lighting scenarios and selected 10 representative scenes from COCO [5] dataset, along with generating location information masks. Subsequently, we generate 10 sets of texts

†Corresponding author, * Equal contribution.

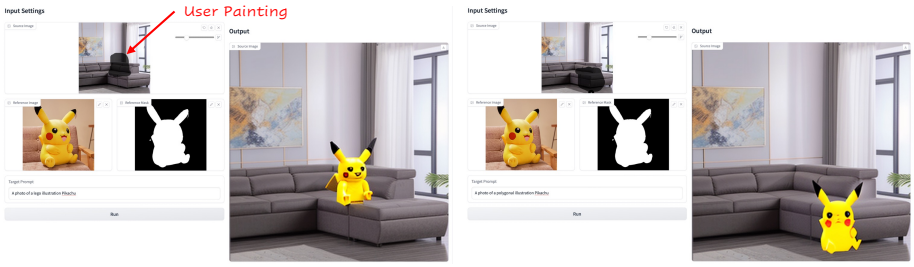

Fig. 2: We demonstrate some features of our preliminary demo, including generating the final image based on the user's painted area, text description, and reference images, among other information.

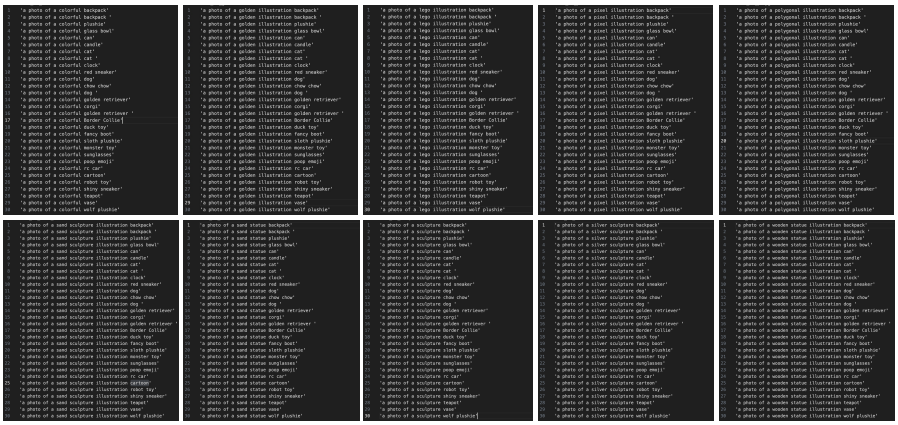

Fig. 3: We display the text descriptions, which are used to modify text attributes.

to modify attributes and applied them to all categories of scenes. As shown in Fig. 1, we present the 30 selected categories of objects in the first three rows, and the last row shows the scene images, the detail of each group is in Fig 4. In Fig. 3, we further demonstrate the text prompts used for content generation.

2 User Interaction

We show a preliminary demo in Fig. 2, users can use a brush to paint on the scene image to indicate the specific area that needs editing. The system generates image content based on different text descriptions input by the user.

3 User Study

We conduct a user study to evaluate the generated image quality of our method with six related approaches. We set evaluation metrics from both text and image

Fig. 4: We display our dataset details of each group, which consist of scene images, position masks, subjects, and subject masks.

perspectives. Each set consists of a scene image, a subject image, a text description, and six generated images. We ask users to rate the fidelity, quality, and text matching of the images based on the following criteria:

Fidelity: Consider the consistency of features between the background and the generated contents within the red area with the original subjects and scene images.

Quality: Whether the image is harmonious.

Text alignment: Whether the content within the red area is consistent with the text description.

User ratings range from 1-5, representing the worst and the best, respectively. We show more details in Fig. 6.

4 Comparisons with Attention-Based Editing Method

We select the state-of-the-art attention-based image editing algorithm PnP for comparison. As shown in Fig. 5, our method solves the problem of previous approaches where using text descriptions would change global information. We achieve accurate object attribute modification, while PnP [7] fails to accurately capture the subject and may alter surrounding environment information. Specifically, we collage the subject target with the scene image and then use PnP [7] for editing.

5 More Results

As shown in Fig. 7 and Fig. 8, we show additional high-quality comparisons of our method.

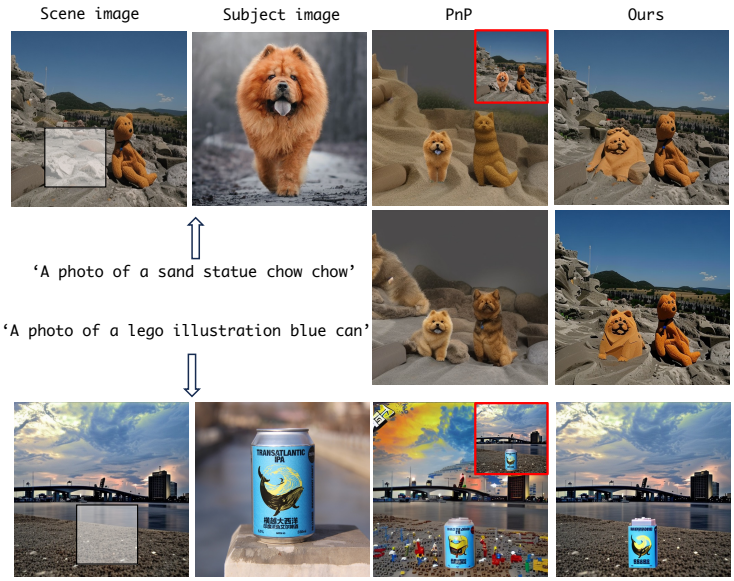

Fig. 5: Compared to the image editing method PnP [7] that also uses attention mechanisms, our approach achieves accurate local image editing without changing global information.

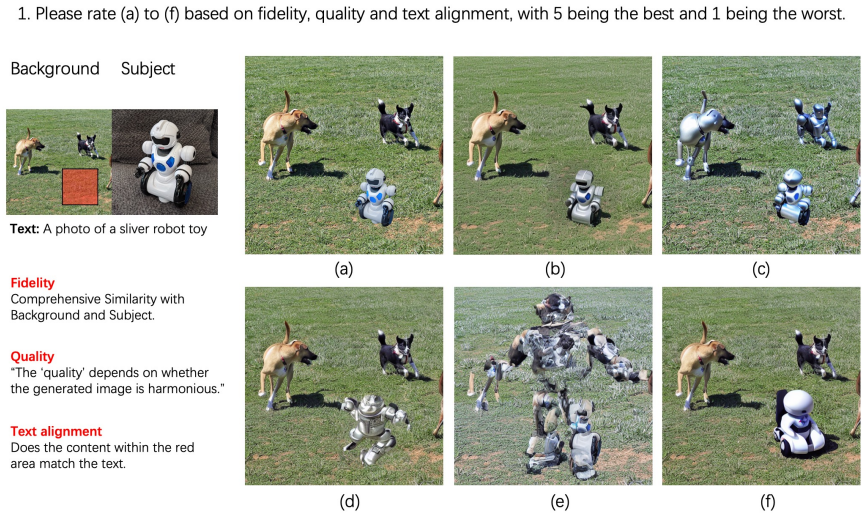

Fig. 6: We present the details of the questionnaire. We rank the six methods based on scores and comprehensively evaluate our approach according to the three criteria: fidelity, quality, and text matching.

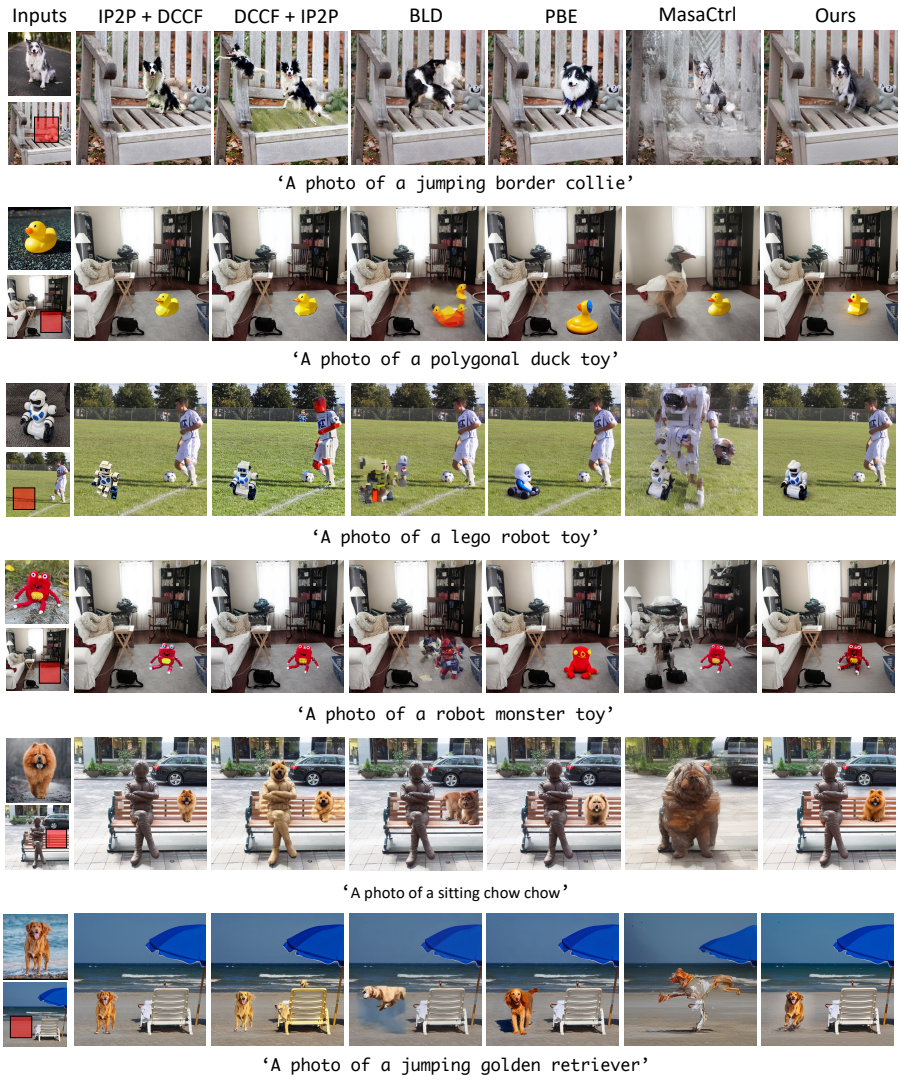

Fig. 7: We show more visual comparison results with IP2P [2] + DCCF [8], DCCF [8] + IP2P [2], BLD [1], PBE [9] and MasaCtrl [3]. Our method outperforms all these methods and overcomes their limitations, achieving outstanding generative performance.

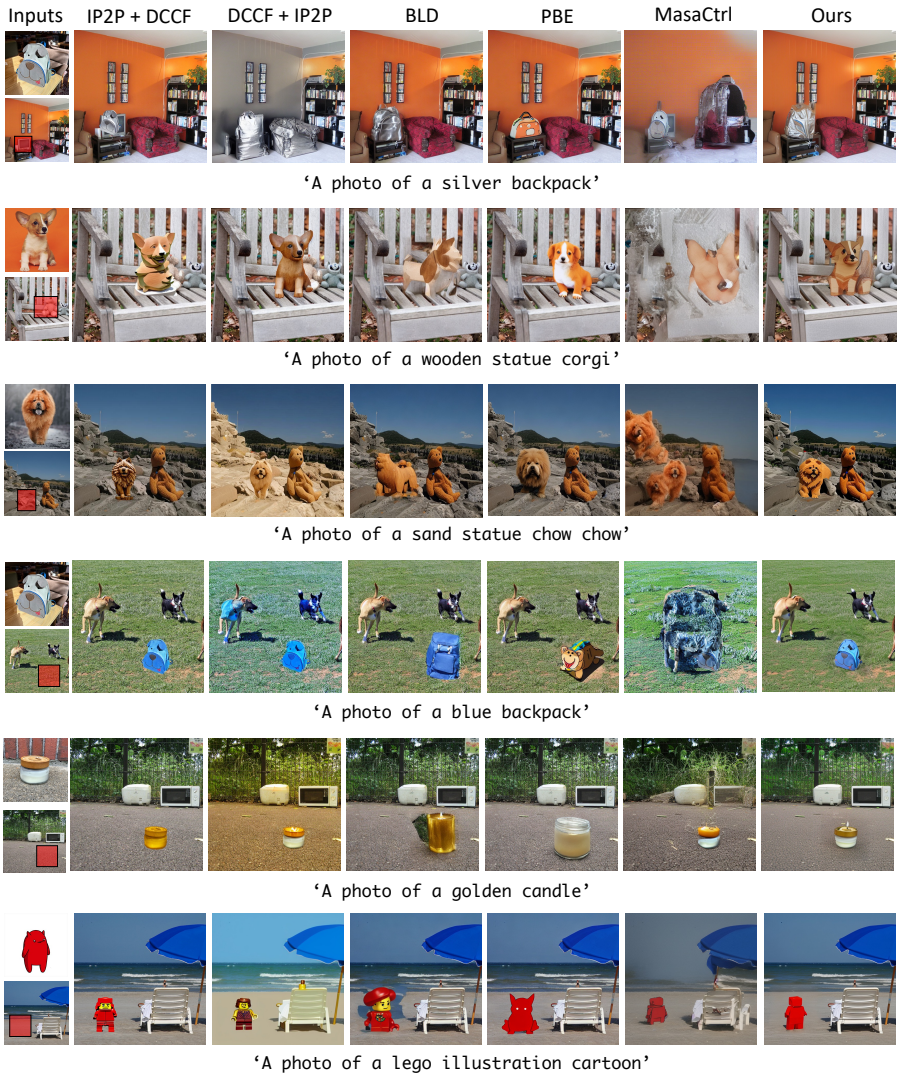

Fig. 8: We show more visual comparison results with IP2P [2] + DCCF [8], DCCF [8] + IP2P [2], BLD [1], PBE [9] and MasaCtrl [3]. Our method outperforms all these methods and overcomes their limitations, achieving outstanding generative performance.

References

1. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Transactions on Graphics (TOG) **42**(4), 1–11 (2023) 5, 6
2. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023) 5, 6
3. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465 (2023) 5, 6
4. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. arXiv preprint arXiv:2210.09276 (2022) 1
5. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 1
6. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242 (2022) 1
7. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. arXiv preprint arXiv:2211.12572 (2022) 3, 4
8. Xue, B., Ran, S., Chen, Q., Jia, R., Zhao, B., Tang, X.: Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In: European Conference on Computer Vision. pp. 300–316. Springer (2022) 5, 6
9. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18381–18391 (2023) 5, 6