Large-Scale Text-to-Image Model with Inpainting is a Zero-Shot Subject-Driven Image Generator

Supplementary Material

A. Baselines

We provide the details of the encoder-based image prompting baselines that we compared in human preference study, as well as in qualitative and quantitative evaluations. All of them utilize a specialized image encoder which extracts image feature from the reference image and injects it into the TTI model. While these models train the specialized image encoder to enable image prompting for zero-shot subject-driven text-to-image generation, they compromise subject alignment, especially in the granular details of the subject. For qualitative results and the human preference study, we compare our method only to the baselines with available open-source weights.

- **ELITE**[1] [16] encodes the visual concepts into textual embeddings, leveraging global and local mapping networks to represent primary and auxiliary features separately, ensuring high fidelity and editability in subject-driven text-to-image generation.
- **BLIP-Diffusion**[2] [3] pre-trains a multimodal encoder following BLIP-2 [4] which produces the text-aligned visual representation of the target subject, and learns the subject representation to enable the TTI model to perform efficient subject-driven text-to-image generation.
- **Kosmos-G** [7] aligns the output space of Multimodal Large Language Models (MLLMs) with the CLIP [10] space by anchoring the text modality, and bridges the MLLM with a frozen TTI model using AlignerNet and instruction tuning. As there are no available weights for this baseline, we cannot conduct the human preference study and can only compare using automatic quantitative metrics based on the values reported in their paper.
- **Subject-Diffusion** [6] utilizes an image encoder trained on their own large-scale subject-driven dataset to incorporate both coarse and fine-grained reference information into the pre-trained TTI model, enabling high-fidelity subject-driven text-to-image generation without test-time fine-tuning. Subject-Diffusion also has no available open-source weights, so we only conduct the quantitative comparisons with their reported values in the paper.
- **λ-Eclipse**[3] [8] employs a CLIP-based latent space and image-text interleaved pre-training and contrastive loss to project text and image embeddings into a unified space, preserving subject-specific visual features and reflecting the target text prompt.
- **MS-Diffusion**[4] [15] introduces a layout-guided framework for multi-subject zero-shot subject-driven text-to-image generation by employing a grounding resampler for detailed feature integration and a multi-subject cross-attention mechanism to ensure spatial control and mitigate subject conflicts.
- **IP-Adapter**[5] [6] [18] trains an effective lightweight adapter to enable image prompting for pre-trained TTI models, using a decoupled cross-attention mechanism with separate cross-attention layers for text and image prompts. At the time the IP-Adapter paper was released, SD-v1.5 [11] was used; however, more recent versions, including SD-XL [9], SD-3 [1], and FLUX [2], have since been made available. For quantitative comparisons, we referenced the results for the SD-XL version from another study [8], while we conducted our own evaluations for the FLUX version to ensure a fair comparison. In all experiments using IP-Adapter, regardless of the base model version, the conditioning scale is set to 0.6.

B. Subject-Driven Text-to-Image Generation

B.1. Evaluation Setting

We conduct the main comparisons with baselines on 30 subjects in DreamBench [13]. These consist of 21 objects and 9 live subjects, with 25 evaluation prompts for the objects or live subjects. Diptych Prompting uses the subject name to refer to the target subject and utilizes evaluation prompts that include the subject name for the target description in diptych text. In all zero-shot baselines and our method, we enhance the subject names by adding descriptive modifiers to more accurately refer to the target subjects in the text prompt. The subject names for each subject are summarized as follows in the form of (directory name, subject name):

- backpack, backpack
- backpack_dog, backpack
- bear_plushie, bear plushie
- berry_bowl, 'Bon appetit' bowl
- can, 'Transatlantic IPA' can
- candle, jar candle
- cat, tabby cat
- cat2, grey cat
- clock, number '3' clock

[1] ELITE: https://github.com/csyxwei/ELITE
[2] BLIP-Diff: https://github.com/salesforce/LAVIS/tree/main/projects/blip-diffusion
[3] λ-Eclipse: https://github.com/eclipse-t2i/lambda-eclipse-inference
[4] MS-Diff: https://github.com/MS-Diffusion/MS-Diffusion
[5] IP-Adapter (SD-XL): https://huggingface.co/h94/IP-Adapter
[6] IP-Adapter (FLUX): https://huggingface.co/XLabs-AI/flux-ip-adapter

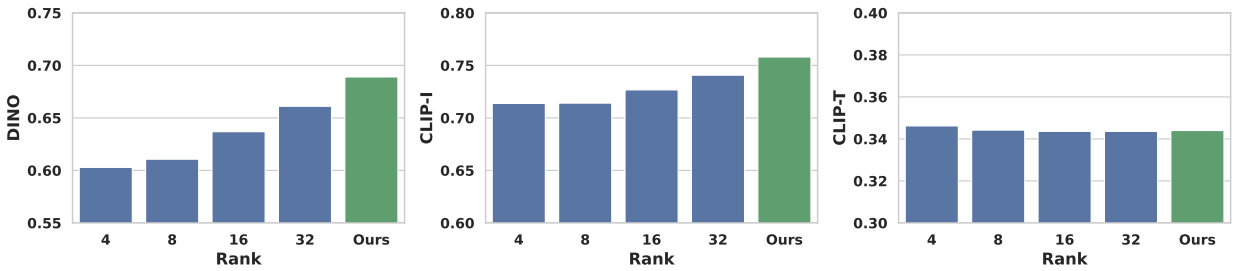

Figure S1. **DreamBooth Comparisons.** Quantitative comparisons to DreamBooth-LoRA with various rank values.

- colorful_sneaker, colorful sneaker
- dog1, fluffy dog
- dog2, fluffy dog
- dog3, curly-haired dog
- dog5, long-haired dog
- dog6, puppy
- dog7, dog
- dog8, dog
- duck_toy, duck toy
- fancy_boot, fringed cream boot
- grey_sloth_plushie, grey sloth plushie
- monster_toy, monster toy
- pink_sunglasses, sunglasses
- poop_emoji, toy
- rc_car, toy
- red_cartoon, cartoon character
- robot_toy, robot toy
- shiny_sneaker, sneaker
- teapot, clay teapot
- vase, tall vase
- wolf_plushie, wolf plushie

B.2. Comparison with Fine-Tuning-Based Method

To provide a more comprehensive comparison, we also compare with DreamBooth [13], a representative fine-tuning-based method. For efficient training, we attach a LoRA adapter to the pre-trained FLUX and perform fine-tuning by training only the LoRA adapter while freezing the FLUX. We train for 300 steps using the Adam optimizer with a learning rate of 1×10^{-4}. Additionally, to compare different fine-tuning model capacities, we adjusted the rank of the LoRA adapter and conducted comparative experiments using the same metrics (DINO, CLIP-I, CLIP-T). The results are presented in Fig. S1, where our Diptych Prompting demonstrates superior performance across various model capacities.

B.3. Additional Results

We include additional samples of Diptych Prompting in Fig. S2 and Fig. S3 for diverse objects and contexts. As demonstrated in the results, our methodology achieves high-quality image generation and satisfies both subject alignment and text alignment in a zero-shot manner by leveraging FLUX's capabilities. Notably, this is accomplished without any specialized training for subject-driven text-to-image generation. We also note that the fine details in the target subject are well reflected in the generated results, even for challenging subjects that previous zero-shot methods struggled with (e.g., robot toy, 'Bon appetit' bowl).

C. Human Preference Study

Following the previous work [13], we perform the human preference study by pairwise comparison in two separate questionnaires for each aspect: subject alignment and text alignment. In both questionnaires, users are presented with a reference image, a target text, and two images generated by each method. They are then asked to select which image better satisfies the desired objective according to the following instructions.

For subject alignment:

- Inspect the reference subject and then inspect the generated subjects.
- Select which of the two generated items reproduces the identity (item type and details) of the reference item
- The subject might be wearing accessories (e.g., hats, outfits). These should not affect your answer. Do not take them into account.
- If you're not sure, select Cannot Determine / Both Equally.
- Which Machine-Generated Image best matches the subject of the reference image?

For text alignment:

- Inspect the target text and then inspect the generated items.
- Select which of the two generated items is best described by the target text.
- If you're again not sure, select Cannot Determine / Both Equally.
- Which Machine-Generated Image is best described by the reference text?

Model	Arch	Param	DINO	CLIP-I	CLIP-T
SD-v2	U-Net	1.2B	0.504	0.744	0.260
SD-XL	U-Net	3.5B	0.941	0.954	0.288
SD-3	MM-DiT	7.7B	0.705	0.821	0.340
FLUX	MM-DiT	16.9B	0.720	0.828	0.352

Table S1. **Diptych Generation Comparisons.** Quantitative comparisons of the diptych generation capabilities of various TTI models based on the total number of parameters, including the autoencoder, main network, and text encoder.

D. Diptych Generation

Our framework relies on the emerging property of the large-scale TTI model, FLUX, particularly its strong understanding of diptych property and the ability to represent diptych accurately. We verify this by synthesizing a total of 2100 diptychs, using 20 objects, each with a pair of two random prompts for each panel among 15 prompts, and comparing the diptych generation performance with those of other previous TTI models. The prompt for diptych generation follows the setup mentioned in Sec. 3.1 of the main paper. We assessed the quality of each diptych by evaluating the interrelation and text alignment of each panel. This is measured through splitting the generated image in half and measuring DINO and CLIP-I scores between each panel, as well as the CLIP-T score between each panel and its description. The results are shown in Tab. S1, in which the diptych generation performance and total number of parameters including the autoencoder, main network, and text encoders are reported. These results exhibit the superior diptych generation capability of FLUX, where smaller models are insufficient. This allows us to extend to inpainting and propose a zero-shot subject-driven text-to-image generation method via diptych inpainting-based interpretation.

E. Background Removal Ablation

We provide additional samples for the ablation study conducted with and without the background removal process G_{seg} in Fig. S4. Consistent with the findings in the main paper, including the background leads to content leakage, where irrelevant elements such as background, pose, and location are mirrored in the generated results. This hinders the accurate reflection of the desired context described by the text and reduces diversity in pose and location. In contrast, removing the background and retaining only the subject information in the reference image on the left panel allows the generated outputs to better align with the desired context while exhibiting greater diversity in pose and location."

Method	DINO	CLIP-I	CLIP-T
RB-Mod [12]	0.295	0.598	**0.372**
IP-Adapter [18]	0.337	0.602	0.371
Diptych Prompting	**0.357**	**0.623**	0.349

Table S2. **Stylized Image Generation Comparisons.** Quantitative comparisons of stylized image generation with previous zero-shot methods.

F. Reference Attention Enhancement Ablation

We further present the actual sample quality variations according to the reference attention rescaling factor λ values to support the quantitative ablations in the main paper. These variations are visualized in Fig. S5. As seen in the qualitative results, the absence of reference attention enhancement ($\lambda = 1.0$) can lead to a loss of fine details of the subject, resulting in subtle discrepancies such as the left eye of the backpack dog, the patch on its right eye, the fur color on the dog's face, or the texture of the bear plushie's fur. As the λ value increases, these missed details are better preserved, leading to improved subject alignment performance. However, excessive enhancement can negatively impact the quality of the generated images, causing the subject to appear slightly blurred or exhibit minor color shifts.

G. Stylized Image Generation

For stylized image generation, Diptych Prompting places the style image in the left panel and inpaints the right panel using the text prompt *"A diptych with two side-by-side images of same style. On the left,* {*original image description*}*. On the right, replicate this style exactly but as* {*target image description*}*"* without attention enhancement ($\lambda = 1.0$) for referencing only the stylistic elements except the content. Additional samples are provided in Fig. S6. Beyond the qualitative results, we also include quantitative comparisons using the same metrics (DINO, CLIP-T, CLIP-I) applied to a total of 2000 generated images in Tab. S2. These images include 4 samples per prompt and per style image, across 25 prompts and 20 style images collected from previous work [14]. As shown in the result, our method demonstrates comparable results to existing zero-shot style transfer methods specialized in stylized image generation, further proving the versatility of our approach.

H. Subject-Driven Image Editing

Diptych Prompting is extended to the subject-driven image editing by placing the reference subject image on the left panel and the editing target image on the right panel in the incomplete diptych. By masking only the desired area in the right panel and applying diptych inpainting, the reference subject from the left panel is generated in the masked

region on the right panel, resulting in the subject-driven image editing. Following the previous work [17], we conduct the subject-driven image editing with selected images from a subset of the MSCOCO [5] validation dataset, in which each image contains a bounding box and the bounding box is smaller than half of image size. We applied masking to the inside of the bounding box, enabling the generation of the reference subject within the specified region. More samples of various subjects and editing target images are available in Fig. S7.

I. Limitations

Currently, FLUX is the only model with sufficient capability to effectively generate diptychs. However, as more advanced text-to-image (TTI) models become available, we anticipate that our method will be applicable to a wider range of models in the future. In line with advancements in other encoder-based zero-shot approaches, there is a need to explore multi-subject-driven text-to-image generation. We leave this exploration for future work. Furthermore, diptych generation requires the generated image to have an aspect ratio of $2:1$. Due to the limitation in the generatable resolution of FLUX, we were unable to produce the diptych image at a size of 2048×1024 pixels and confirmed results up to 1536×768 pixels, resulting in subject-driven image (right panel) being 768×768 pixels in size. We expect that this issue can be easily addressed by utilizing super-resolution models such as ControlNet [19] or advanced TTI models for high-resolution image generation in the future.

Figure S2. **Subject-Driven Text-to-Image Generation.** More samples of subject-driven text-to-image generation using Diptych Prompting.

Figure S3. **Subject-Driven Text-to-Image Generation.** More samples of subject-driven text-to-image generation using Diptych Prompting..

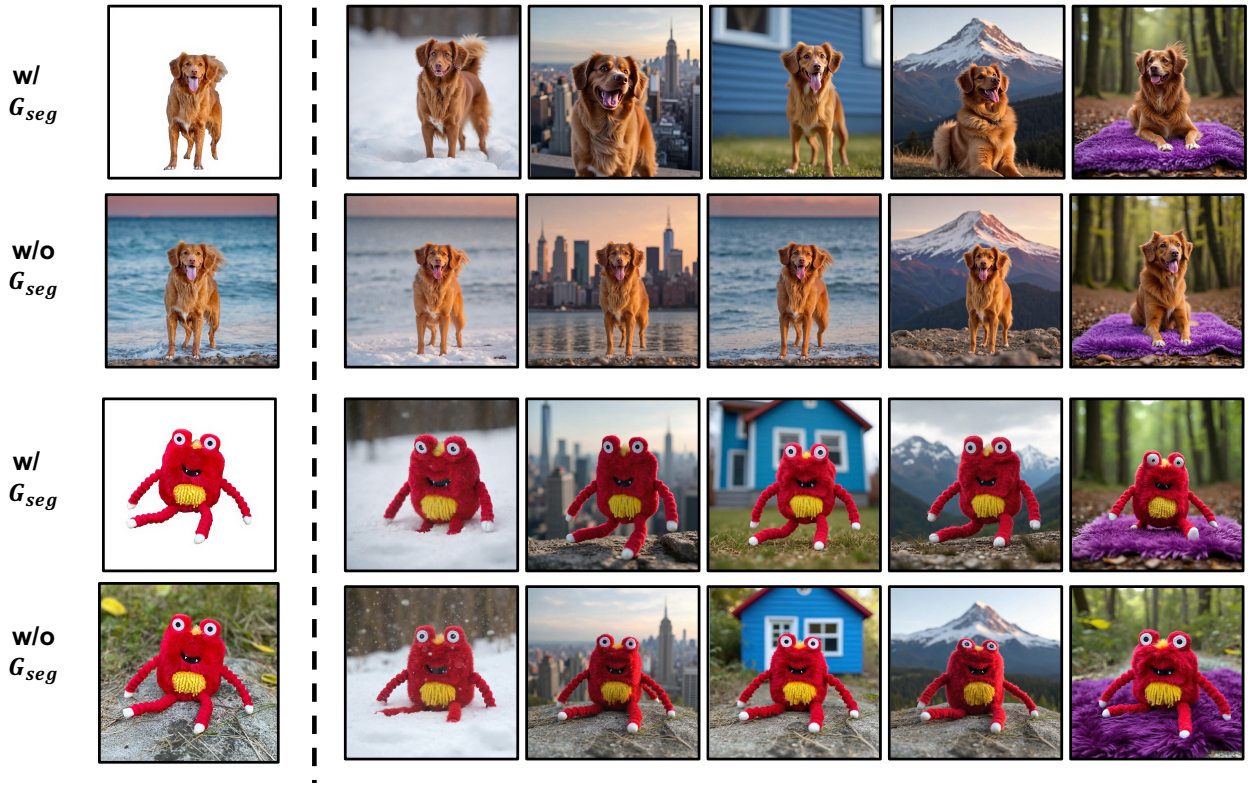

Figure S4. G_{seg} **Ablation.** Qualitative comparisons with and without the background removal process.

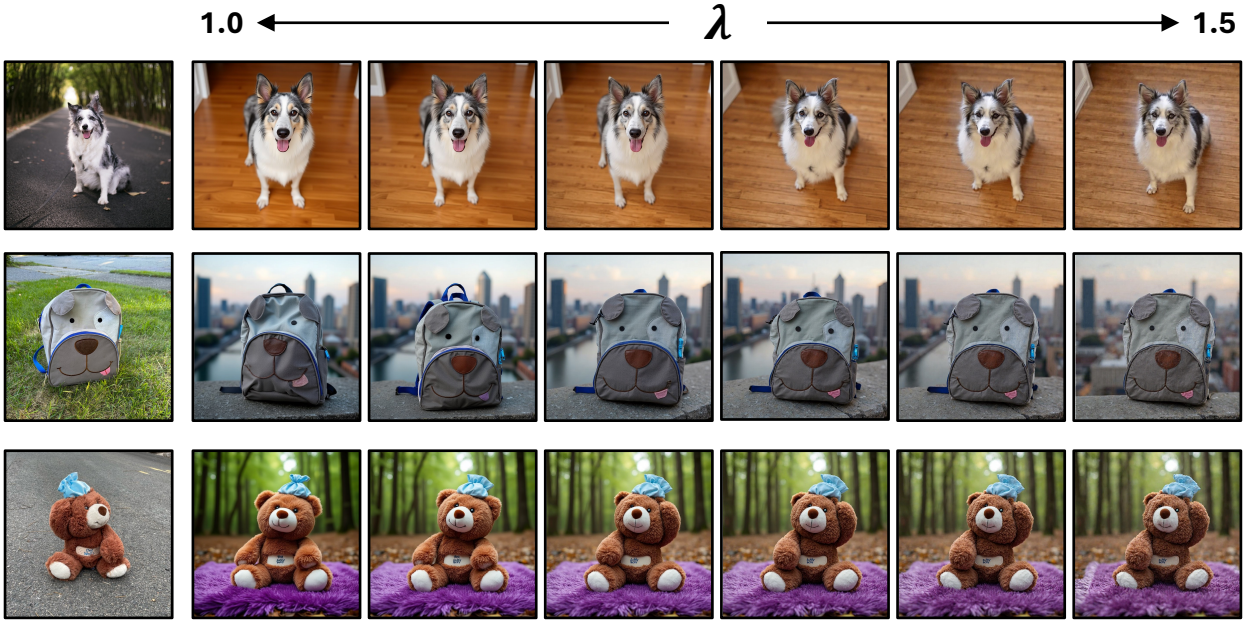

Figure S5. λ **Ablation.** Qualitative transitions according to the varying λ values. we control the λ from 1.0 (without reference attention enhancement) to 1.5. For a detailed view, please zoom in.

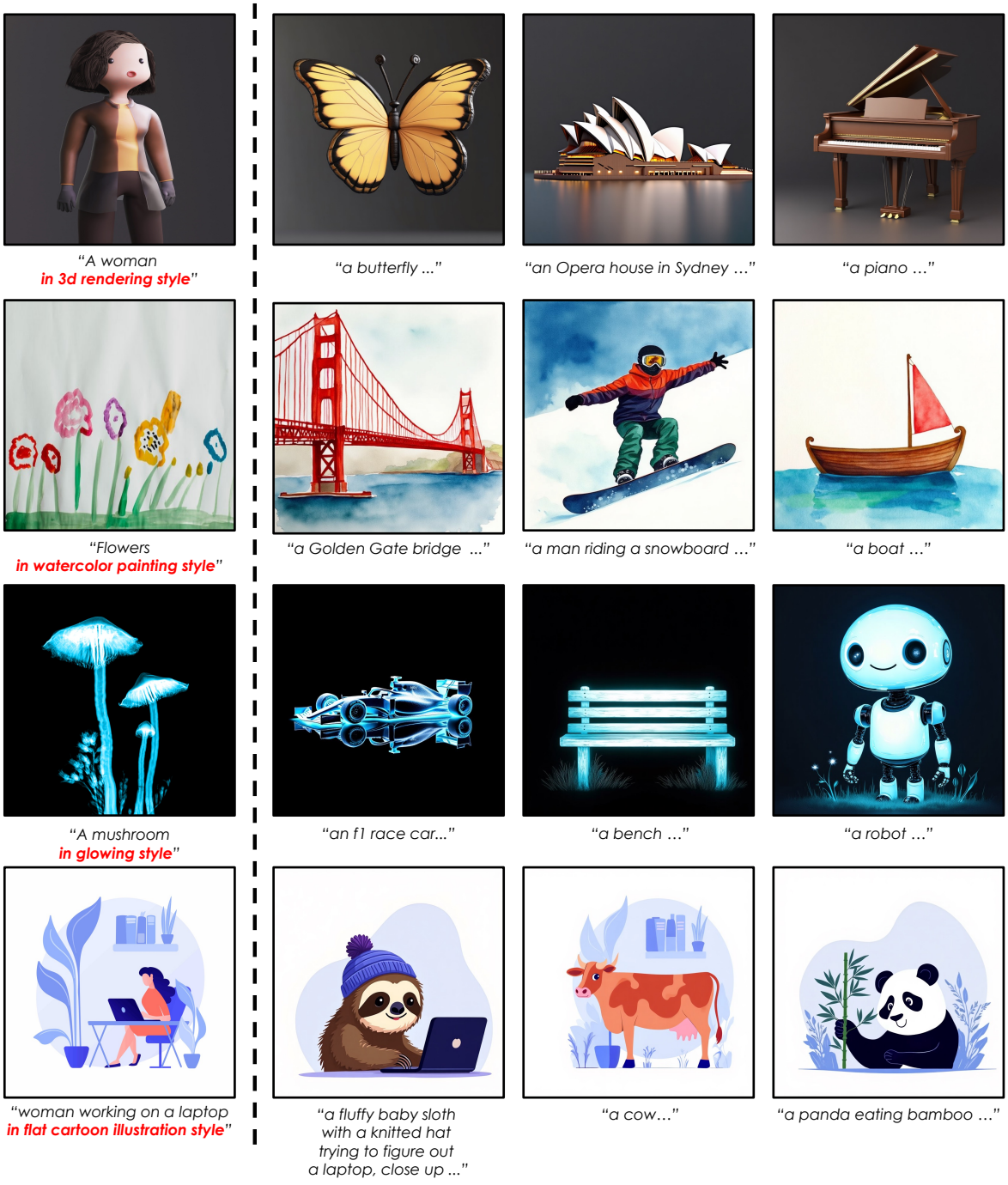

Figure S6. **Stylized Image Generation.** More samples of stylized image generation using Diptych Prompting.

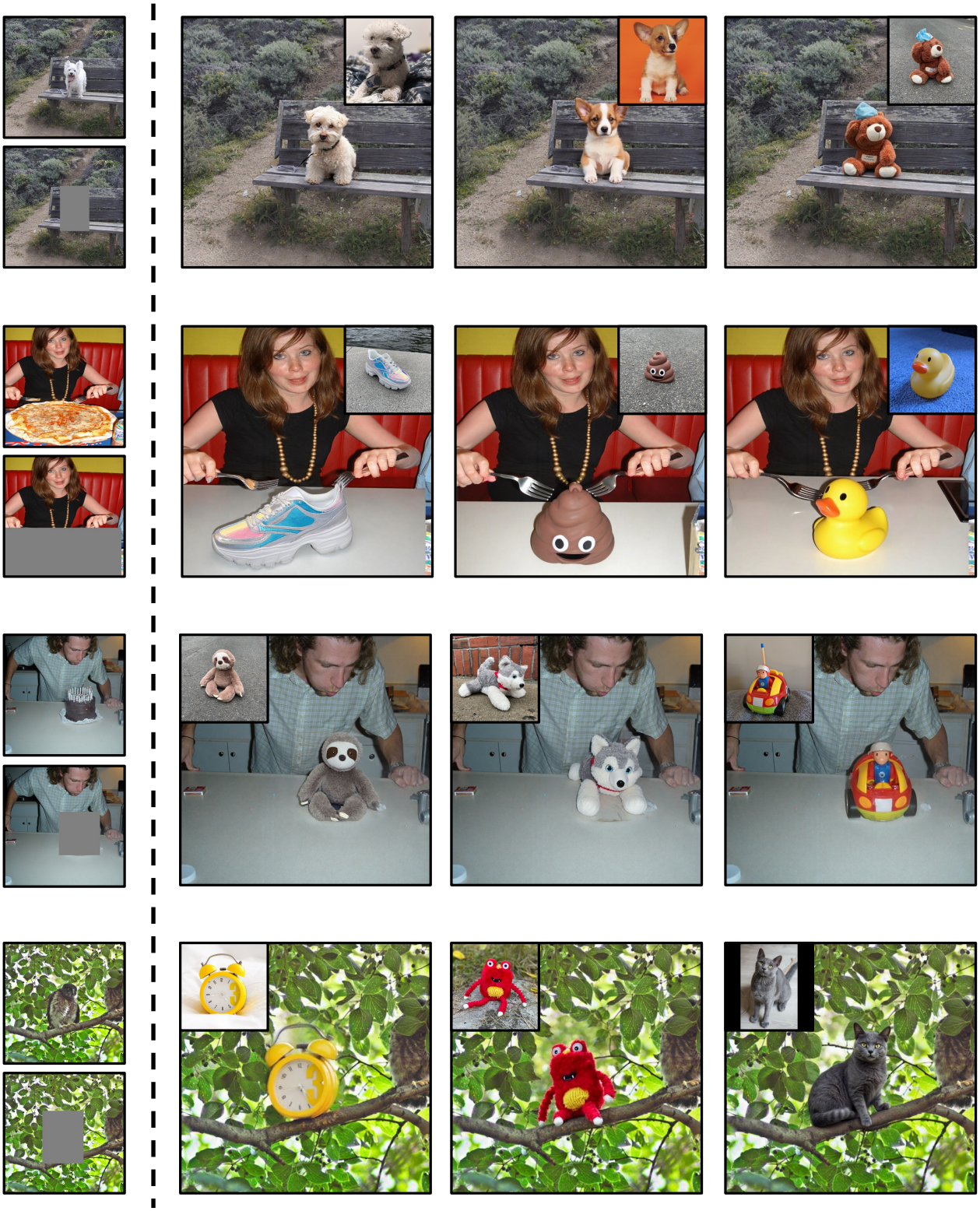

Figure S7. **Subject-Driven Image Editing.** More samples of subject-driven image editing using Diptych Prompting.

References

[1] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1

[2] Black Forest Labs. Flux.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024. 1

[3] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. 2024. 1

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4

[6] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 1

[7] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[8] Maitreya Patel, Sangmin Jung, Chitta Baral, and Yezhou Yang. λ-eclipse: Multi-concept personalized text-to-image diffusion models by leveraging CLIP latent space. *arXiv preprint arXiv:2402.05195*, 2024. 1

[9] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[12] Litu Rout, Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Rb-modulation: Training-free personalization of diffusion models using stochastic optimal control. *arXiv preprint arXiv:2405.17401*, 2024. 3

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

[14] Kihyuk Sohn, Lu Jiang, Jarred Barber, Kimin Lee, Nataniel Ruiz, Dilip Krishnan, Huiwen Chang, Yuanzhen Li, Irfan Essa, Michael Rubinstein, et al. Styledrop: Text-to-image synthesis of any style. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[15] X Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024. 1

[16] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023. 1

[17] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 4

[18] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 3

[19] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 4