

Pick-and-Draw: Training-free Semantic Guidance for Text-to-Image Personalization

Supplementary Material

A DETAILS OF BASELINES

We reproduce all the baselines according to official implementations from the diffusers library.

Textual Inversion [2] is a fine-tuning-based method that optimizes a placeholder embedding of the diffusion text encoder, so as to invert the subject into the diffusion text space. Textual Inversion requires 2000 ~ 3000 training steps for learning a new subject and we report results using 2500 steps across all instances in the experiments on DreamBench.

DreamBooth [5] is a fine-tuning-based method that optimizes the parameters of the whole diffusion UNet for image personalization. It learns to bind the specified subject with a rare text token via a reconstruction loss. It further utilizes a class prior preservation loss to avoid language drift and reduced generative diversity but at the cost of worsening identity consistency. DreamBooth requires around 400 ~ 800 steps in general and we report the results using 600 steps in the experiments on DreamBench. We do not use class prior preservation loss since we find it significantly reduces the overall identity consistency, complicates convergence, and requires heavy tuning for each instance.

BLIP-Diffusion [4] is an encoder-based method which pre-trains a multimodal encoder following BLIP-2 to produce visual representation aligned with the text. It exhibits zero-shot capability but needs further fine-tuning to achieve better performance. BLIP-Diffusion requires 40 ~ 120 steps for different subjects and we report the results using 80 steps in the experiments on DreamBench.

B MORE ABLATION AND ANALYSIS

Visualization of activation selection. In Fig. 2 we visualize different activation selection strategies for the appearance picking guidance. We choose Textual Inversion as an example baseline and generate samples with Pick-and-Draw guidance using activations of resolution 16×16 , 32×32 , and 64×64 . Activations of resolution 16×16 contain mainly layout information, which is coarse-grained and may introduce unwanted layout leakage (*i.e.* the window frame) to the guided samples (3rd column). Activations of resolution 64×64 focus on high-frequency details such as edges, which are not sufficient for local appearance transfer. The corresponding guided samples (4th column) fail to maintain identity consistency with the reference image. In comparison, activations of resolution 32×32 encode rich semantic information and focus on different salient regions of the subject, facilitating the local appearance transfer to achieve the best visual result in identity preservation. Therefore, the guided samples (last column) not only exhibit consistent appearances but also eliminate the interference of background from reference images.

Impact of guidance step selection. We conduct ablation study on guidance steps for appearance-aware loss ℓ_{app} and layout-aware loss ℓ_{lay} on DreamBooth. Results are presented in Fig. 1. Since early denoising steps exert a significant impact on the generated object layout [1, 3], we perform layout guidance solely (Fig. 1 *left*) from the very beginning and find that the optimal performance is achieved when stopping guidance at step 10. More layout guidance steps result in a significant performance drop on DINO score. We perform appearance guidance (Fig. 1 *middle*) in a similar manner and find that 10 steps are sufficient for appearance transfer. Additionally, applying only appearance guidance leads to a substantial decrease in the CLIP-T score, indicating a severe overfitting problem. We then set the guidance schedule of ℓ_{lay} as [0, 10], fix the length of appearance guidance as 10 steps, and perform the two types of guidance simultaneously (Fig. 1 *right*). The scatter plot shows that starting the appearance guidance at step 10 achieves the best trade-off performance.

Considering the above, we set the range of guidance steps for layout guidance and appearance guidance as [0, 10] and [10, 20], respectively. The optimal setting is in line with intuition, where we initially employ layout-aware loss to constrain subject shape and background to ensure generative diversity and image-text fidelity, followed by the utilization of appearance loss to enforce object identity consistency and subject fidelity. We adhere to this setting throughout all other experiments.

C FAILURE CASES

Pick-and-Draw fails to generate images aligned with text prompts if the template image by Stable Diffusion provides a false layout prior. In addition, it may suffer from incomplete appearance transfer when the subjects generated by the baseline model differ too much from the reference. We present two possible failure cases in Fig. 3.

D MORE QUALITATIVE RESULTS

Results on DreamBooth. We provide more qualitative results in Fig. 4 to show the improvement of DreamBooth when equipped with Pick-and-Draw. Our method consistently improves the performance of DreamBooth on both subject fidelity (row 5 ~ 7) and image-text fidelity (row 1 ~ 6).

Results on Vanilla Stable Diffusion. We apply Pick-and-Draw to Vanilla Stable Diffusion for zero-shot text-to-image personalization. Visual results can be shown in Fig. 5. We observe favorable outcomes of the tree and flower cases. As illustrated in Section 4.5 of the main body, our method is capable of personalizing simple subjects even without strong prior of the fine-tuning-based personalization baselines like Dreambooth [5].

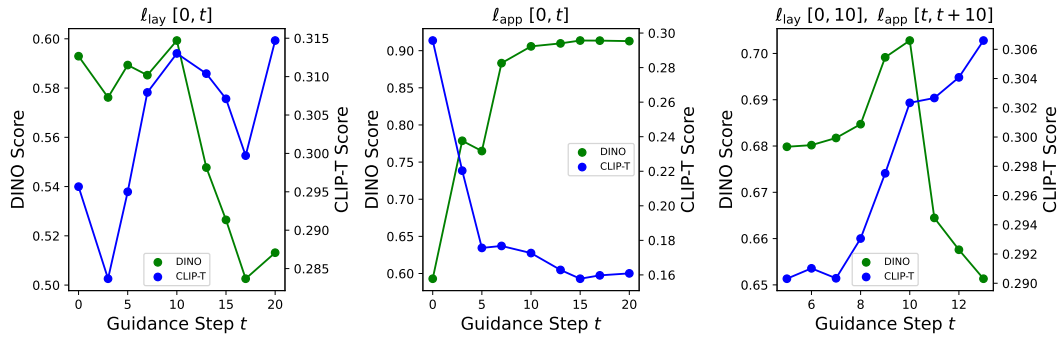


Figure 1: Ablation study for guidance step selection. We conduct ablation experiments under three conditions: employing only the layout loss (left), employing only the appearance loss (middle), and employing both losses simultaneously (right). The guidance steps of both losses are labeled.

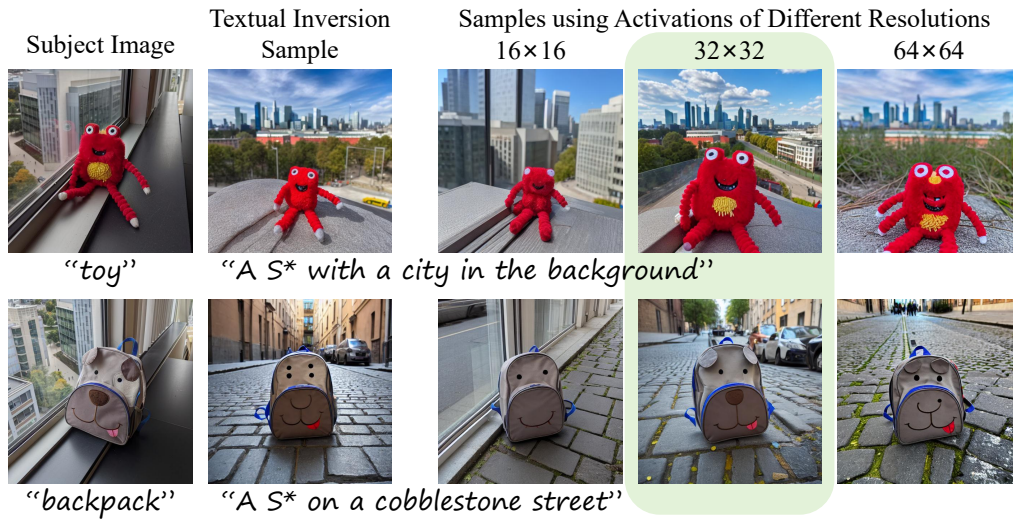


Figure 2: Ablation study for visualizing different activation selection strategies for the appearance picking guidance, conducted on Textual Inversion. The best selection strategy is marked in green.

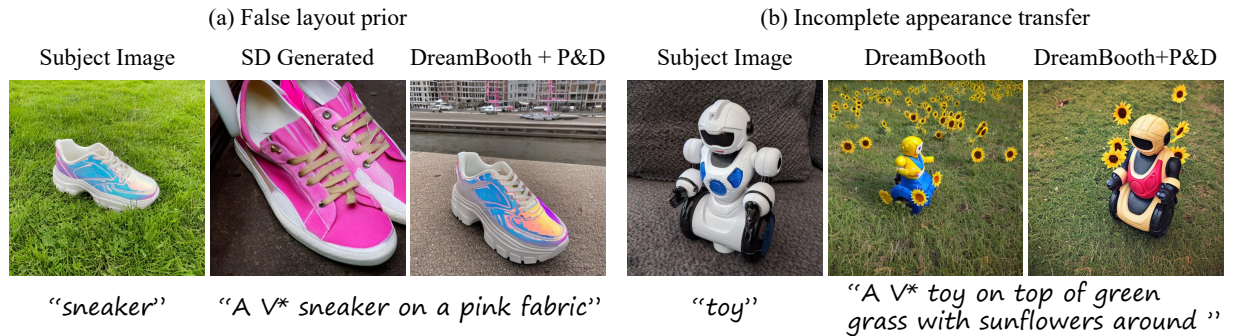


Figure 3: Example failure generations. SD stands for Stable Diffusion and P&D stands for our method Pick-and-Draw.

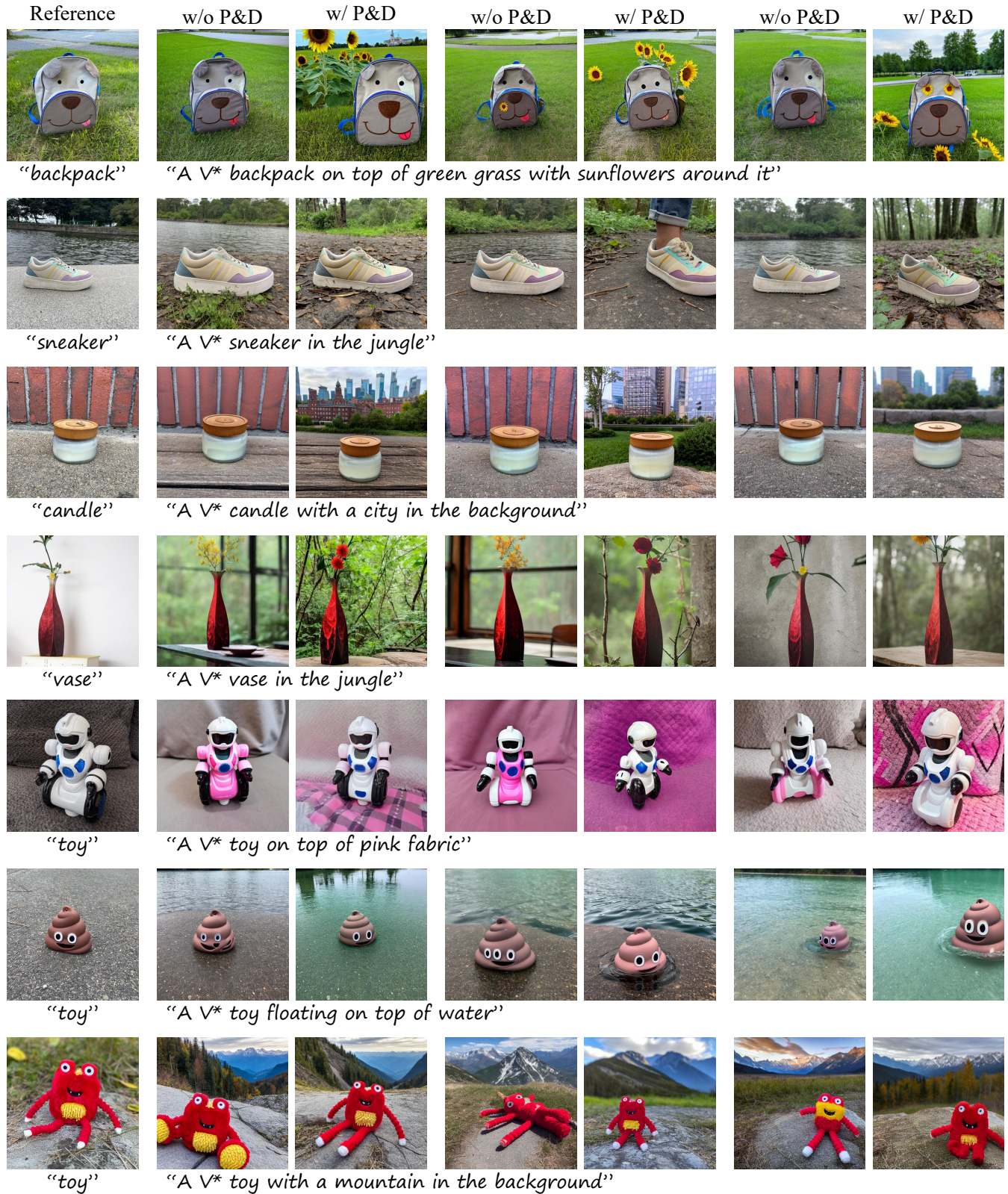


Figure 4: More qualitative results on DreamBooth before and after applying Pick-and-Draw.

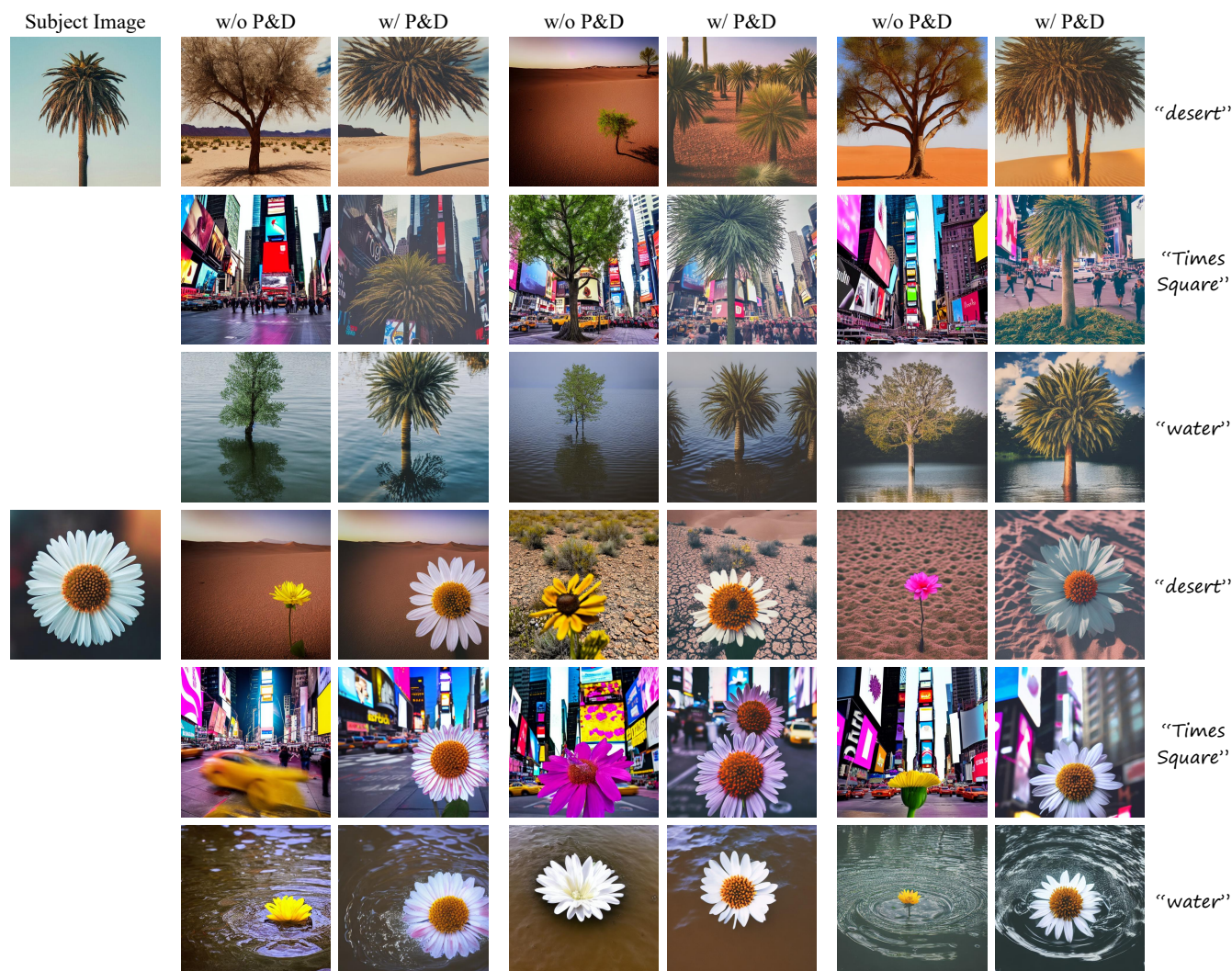


Figure 5: More qualitative results on Vanilla Stable Diffusion before and after applying Pick-and-Draw.

REFERENCES

- [1] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–10.
- [2] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022).
- [3] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).
- [4] Dongxu Li, Junnan Li, and Steven CH Hoi. 2023. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720* (2023).
- [5] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. 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*. 22500–22510.