

Fine-Tuning Visual Autoregressive Models for Subject-Driven Generation

Supplementary Material

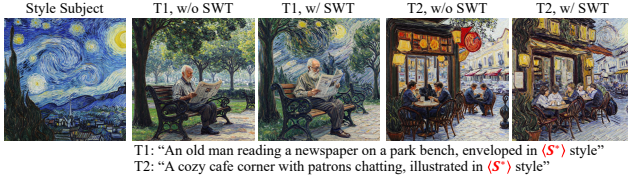


Figure 1. Qualitative comparison demonstrating the effectiveness of Scale-wise Weighted Tuning (SWT) for style personalization.

A. Additional Ablations

A.1. Scale-wise Weighted Tuning for Style Personalization

Unlike diffusion models, where later timesteps capture fine, high-frequency details, our analysis indicates that later scales of large-scale pretrained VAR models contribute minimally to output variation. Consequently, our proposed Scale-wise Weighted Tuning (SWT) strategy de-emphasizes these scales, improving robustness in style personalization tasks, particularly for capturing high-frequency details.

To validate SWT effectiveness, we fine-tune on eight distinct style concepts from DreamBench++ [5], each with nine text prompts, for 100 iterations per style. We generate eight outputs per pair and present representative qualitative comparisons in Fig. 1. As shown in Fig. 1, SWT achieves clearer stylistic expressions, such as distinct swirls and brushstrokes.

A.2. Prior Distillation vs. Prior Preservation Loss

We further compare our Prior Distillation (PD) method against the Prior Preservation Loss (PPL) approach [6]. For integrating PPL into VAR [7], we replace the original MSE objective with cross-entropy objective, using 100 class-specific images generated with the Infinity-2B checkpoint.

Quantitative results in Tab. 1 and qualitative results in Fig. 2 demonstrate comparable performance between PD and PPL across various metrics. Notably, PD eliminates the requirement for same-class dataset collection, substantially reducing the preparation time and computational overhead.



Figure 2. Qualitative comparison between our Prior Distillation (PD) and Prior Preservation Loss (PPL).

Variant	$I_{\text{dino}} \uparrow$	$I_{\text{clip}} \uparrow$	$T_{\text{clip}} \uparrow$	$\text{PRES} \downarrow$	$\text{DIV} \uparrow$
Ours w/ PD	0.786	0.853	0.267	0.709	0.272
Ours w/ PPL	0.653	0.822	0.265	0.608	0.378

Table 1. Quantitative comparison between our Prior Distillation (PD) and Prior Preservation Loss (PPL).

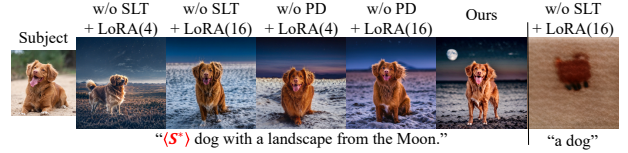


Figure 3. Qualitative comparison of LoRA adaptation versus our SLT + PD approach.

A.3. Selective Layer Tuning with LoRA

We evaluate Selective Layer Tuning (SLT) combined with Prior Distillation (PD) against the popular LoRA method [2]. Specifically, we apply LoRA across all VAR layers at multiple ranks (4 and 16) without employing SLT or PD, to isolate their contributions. As shown in Tab. 2, LoRA without SLT or PD results in unstable fine-tuning, causing visual artifacts or failing to accurately capture subject concepts.

The qualitative results in Fig. 3 further confirm that our combined approach (SLT + PD) stabilizes personalization and preserves subject identity, validating their critical role in achieving robust VAR personalization.

B. User Study Details


We conduct a user study with 20 participants to evaluate personalization effectiveness. Participants compare outputs

Base	Adapter	$I_{\text{dino}} \uparrow$	$I_{\text{clip}} \uparrow$	$T_{\text{clip}} \uparrow$	PRES \downarrow	DIV \uparrow
w/o SLT	+ LoRA(4)	0.653	0.822	0.265	0.608	0.378
	+ LoRA(16)	0.770	0.852	0.264	0.383	0.351
w/o PD	+ LoRA(4)	0.753	0.852	0.265	0.676	0.376
	+ LoRA(16)	0.775	0.854	0.265	0.781	0.373
Ours (w/ SLT, PD)		0.786	0.853	0.267	0.709	0.272


Table 2. Quantitative comparison with LoRA.

Please select the best image among the generated images (a, b, c) according to the evaluation criteria.


Subject Image




(a)



(b)



(c)



Input prompt: a black cat on a cobblestone street

(a)

Subject Fidelity : ☐

Prompt Alignment : ☐

(b)

Subject Fidelity : ☐

Prompt Alignment : ☐

(c)

Subject Fidelity : ☐

Prompt Alignment : ☐

Figure 4. User study interface for evaluating personalization quality.

from our method against baselines, focusing on subject fidelity and prompt alignment, following the evaluation protocol from DreamBooth [6]. The user interface is shown in Fig. 4.

C. Additional Qualitative Comparisons

C.1. Comparison with FLUX-based Method

We further compare our VAR-based approach with Personalize Anything (PA) [1], which is based on the FLUX framework [3, 4]. As illustrated in Fig. 5, PA [1] struggles to generate dynamic poses from the subject image, whereas our approach successfully synthesizes diverse and dynamic poses, highlighting the efficacy of our VAR-based personalization.

C.2. Comparison with Diffusion-based Methods

Additional qualitative comparisons against diffusion-based baselines are provided in Fig. 6. These examples clearly demonstrate our method’s ability to better preserve subject identity, accurately capturing crucial attributes such as color



Figure 5. Qualitative comparison to Personalize Anything (PA) [1].

and shape. Furthermore, our method achieves notably improved alignment with the provided prompts.

References

- [1] Haoran Feng, Zehuan Huang, Lin Li, Hairong Lv, and Lu Sheng. Personalize anything for free with diffusion transformer. *arXiv preprint arXiv:2503.12590*, 2025. 2
- [2] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1
- [3] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 2
- [4] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2
- [5] Yang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. In *The Thirteenth International Conference on Learning Representations*. 1
- [6] 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
- [7] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2025. 1

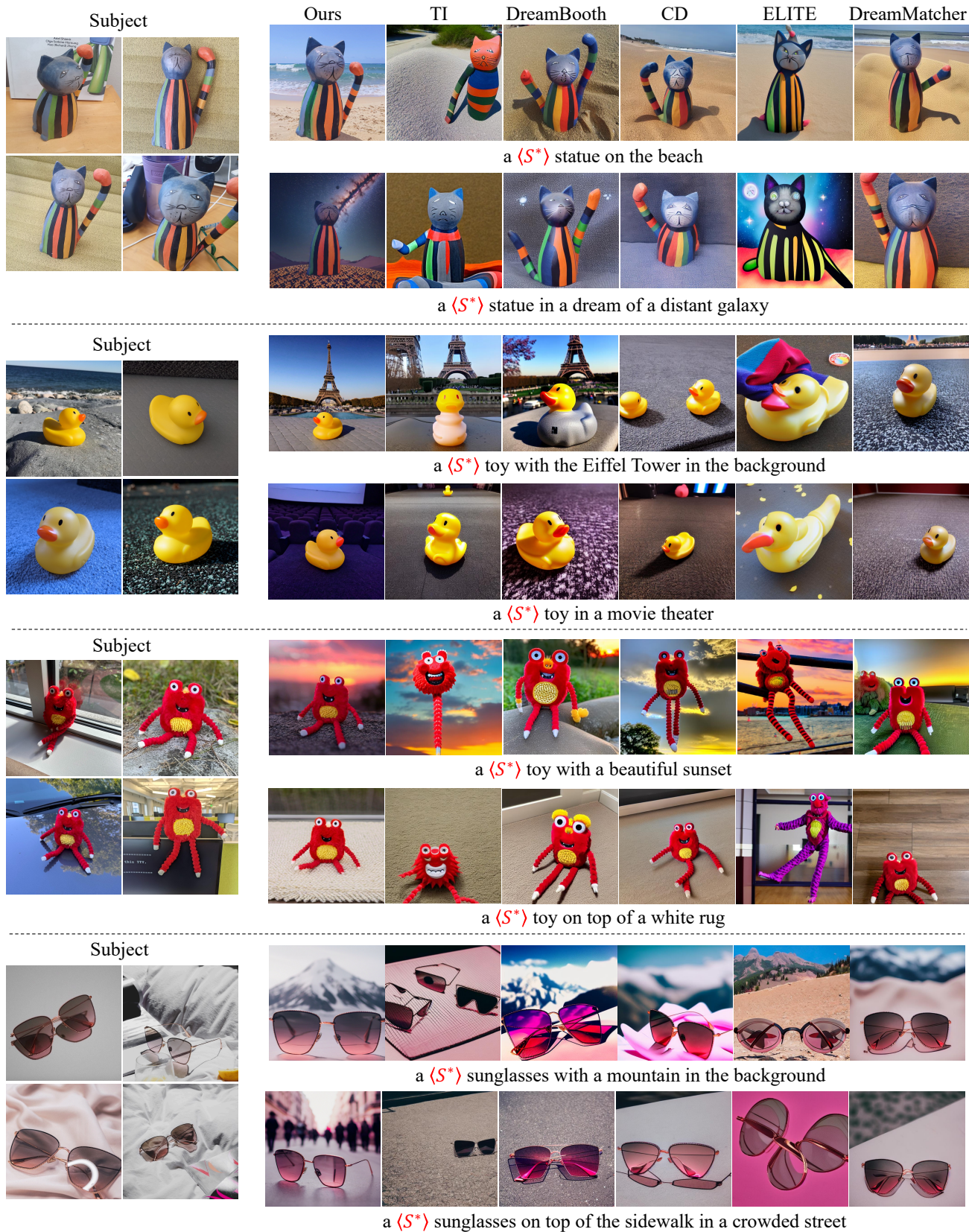


Figure 6. Extended qualitative comparisons to diffusion-based baselines.