SwiftPie: Lightning-fast Subject-driven Image Personalization via One step Diffusion

Supplementary Material

In this supplementary material, we present additional quantitative and qualitative results in Sec. 1. Furthermore, we provide extended ablation studies in Sec. 2 to examine the contribution of each component in SwiftPie. A user study is also included in Sec. 3 to assess user's preference of our personalization approach. Finally, we discuss the societal implications of our personalization tool in Sec. 4.

1. Additional Results

We report additional quantitative results on DreamBench++ [4] in Tab. 1. DreamBench++ extends the original DreamBench benchmark [5] by providing more reference images and personalized prompts (9 per subject) generated by GPT-4o. The reference images cover several sub-categories, including objects, animals, humans, and styles. Since our work focuses on subject-driven personalization, we exclude style references from the evaluation. Note that each subject is represented by a single reference image. In the object sub-category, SwiftPie outperforms other multi-step approaches in both subject preservation and text alignment. For animal and human sub-categories, SwiftPie delievers competitive performance relative to multi-step methods. We believe these results could be further improved by expanding SwiftPie's training data to include more examples from these sub-categories.

Additionally, we provide more qualitative results of SwiftPie on both DreamBench [5] (see Fig. 2, Fig. 3), and DreamBench++ [4] (see Fig. 4, Fig. 5). As illustrated, our one-step personalization approach consistently delivers high-fidelity subject preservation and strong text alignment, while maintaining significantly faster inference compared to multi-step methods. The qualitative examples also demonstrate that SwiftPie effectively handles diverse subject categories and complex prompts without sacrificing visual quality.

2. Additional Ablation Studies

2.1. Effect of losses

We investigate the impact of different training objective configurations used for SwiftPie. Specifically, we ablate the contributions of both perceptual losses and adversarial loss during training. As shown in Tab. 2, removing either perceptual loss (DIST and SSIM) or the adversarial loss results in a significant drop in identity preservation, whereas the full objective setting achieves the best identity preservation score.

2.2. Compatibility with another one-step backbone

As discussed in the main paper, SwiftPie can be further improved with stronger generative priors from more powerful one-step generative models. To demonstrate this, we adopt the same training strategy described in the main paper and train SwiftPie using an alternative one-step model, DMDv2 with an SDXL backbone. We report both qualitative and quantitative results in Fig. 1 and Tab. 3. As shown in Tab. 3, SwiftPie with the SDXL backbone achieves higher identity preservation and better text alignment score compared to SD1.5 backbone. Due to increase model's size, runtime increases from 0.17s to 0.41s; however, this remains significantly faster than multi-step approaches. In Fig. 1, SwiftPie can generate personalized images with good generalization and strong alignment in both subject fidelity and text prompt.

3. User Study

We conducted an additional user study to compare preferences between our one-step personalization results and those produced by other multi-step approaches. Specifically, we selected 5 subject identities and 4 prompts per subject from the DreamBooth dataset as inputs to both SwiftPie and several multi-step baselines. For each of the resulting 20 paired samples (one from SwiftPie and one from a multi-step method), approximately 20 participants rank their preferred output based on three criteria: identity preservation, prompt alignment, and overall image quality. As shown in Fig. 6, SwiftPie emerges as the preferred method across all criteria, while also being the fastest: 86% of users favored it for identity preservation, 87% for prompt alignment, and 91% for overall image quality.

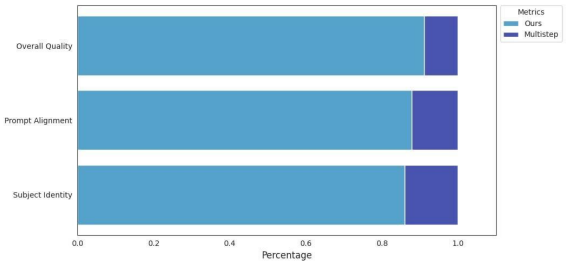

Figure 6. User study results

4. Societal Impact

SwiftPie is an AI-powered visual generation framework that can enable fast personalization with high-fidelity subject

Figure 1. Qualitative results of SwiftPie with DMDv2 one-step model (SDXL backbone) on DreamBench.

Table 1. Quantitative comparison with several multi-step personalization approaches on DreamBench++. Each color box object / animal / human indicates benchmark results on different sub-categories.

Type	Method	Subject Preservation			Text Alignment
		CLIP-I ↑	DINO ↑	Nexus ↑	CLIP-T ↑
Multi-step (50 steps)	DreamBooth [5]	0.586 / 0.587 / 0.488	0.201 / 0.156 / 0.204	0.293 / 0.302 / 0.212	0.287 / 0.275 / **0.320**
	IP-Adapter [7]	0.853 / 0.889 / **0.787**	0.743 / 0.790 / 0.584	0.771 / **0.819** / **0.622**	0.292 / 0.294 / 0.235
	ELITE [6]	0.788 / 0.849 / 0.705	0.578 / 0.693 / 0.489	0.603 / 0.704 / 0.492	0.310 / 0.311 / 0.270
	SSR-Encoder [9]	0.814 / 0.851 / 0.709	0.628 / 0.704 / 0.461	0.689 / 0.726 / 0.479	0.314 / 0.316 / 0.271
	BLIP-Diffusion [3]	0.828 / 0.867 / 0.713	0.639 / 0.707 / 0.511	0.721 / 0.759 / 0.533	0.286 / 0.306 / 0.237
	DisEnvisioner [2]	0.851 / **0.892** / 0.783	0.725 / 0.786 / **0.589**	0.767 / 0.808 / 0.605	0.315 / 0.320 / 0.258
One-step	SwiftPie (Ours)	**0.857** / 0.890 / 0.746	**0.770** / **0.815** / 0.540	**0.783** / 0.796 / 0.534	**0.316** / **0.321** / 0.272

preservation and strong text alignment. Its exceptional speed makes it well-suited for interactive content creation tasks. However, SwiftPie also raises potential societal concerns, as it could be misused for unethical purposes, such as generating sensitive or harmful content to spread misinformation. This necessitates future research works on detecting and localizing AI-manipulated images to alleviate such potential issues.

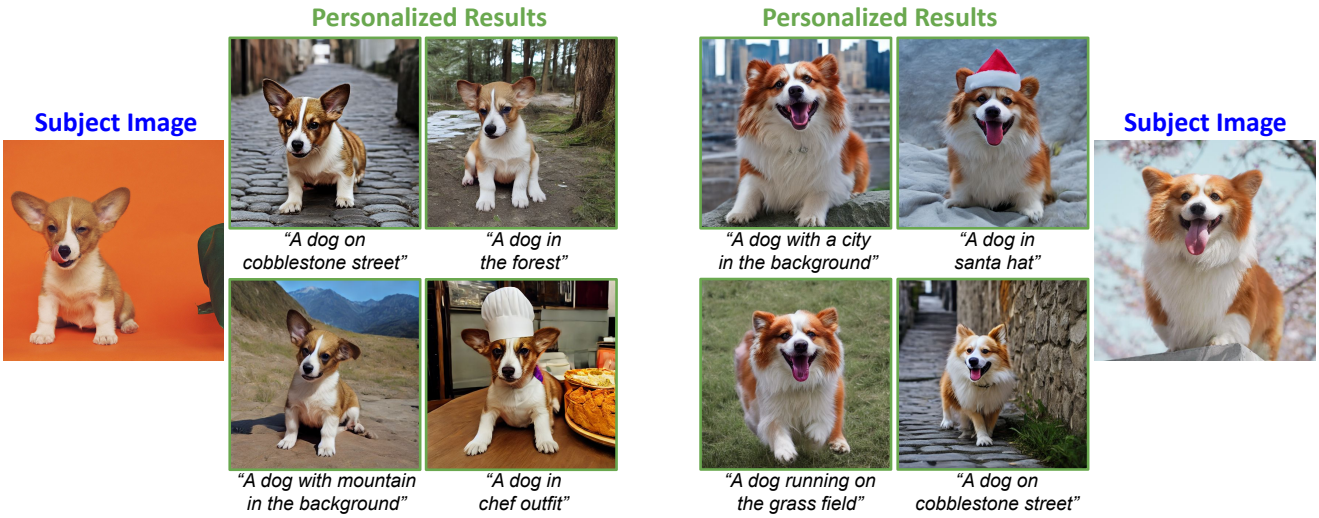

Figure 2. Additional personalization results on DreamBench.

Figure 3. Additional personalization results on DreamBench.

Table 2. Ablation studies on different training objectives.

Setting	$\mathcal{L}_{\text{SSIM}}$	$\mathcal{L}_{\text{DIST}}$	$\mathcal{L}_{\text{adv}}$	Subject Preservation			Text Alignment
				CLIP-I ↑	DINO ↑	Nexus ↑	CLIP-T ↑
Setting 1	✓	✗	✓	0.824	0.684	0.720	**0.313**
Setting 2	✗	✓	✓	0.842	0.729	0.744	0.310
Setting 3	✓	✓	✗	0.841	0.736	0.748	0.311
Setting 4	✗	✗	✓	0.824	0.683	0.719	0.310
Full Setting (SwiftPie)	✓	✓	✓	**0.862**	**0.777**	**0.778**	0.306

Table 3. Quantitative results of SwiftPie with different one-step backbone models on DreamBench

Type	One-step Model	Subject Preservation			Text Alignment	Runtime ↓ (seconds)
		CLIP-I ↑	DINO ↑	Nexus ↑	CLIP-T ↑	
SwiftPie	SwiftBrushv2 [1] (SD1.5 backbone)	0.862	0.777	0.778	0.306	**0.17s**
	DMDv2 [8] (SDXL backbone)	**0.868**	**0.787**	**0.795**	**0.310**	0.41s

Figure 4. Additional personalization results on DreamBench++.

Figure 5. Additional personalization results on DreamBench++.

References

[1] Trung Dao, Thuan Hoang Nguyen, Thanh Le, Duc Vu, Khoi Nguyen, Cuong Pham, and Anh Tran. Swiftbrush v2: Make your one-step diffusion model better than its teacher. In *Proceedings of the European Conference on Computer Vision*, 2025. 3

[2] Jing He, Haodong Li, huyongzhe, Guibao Shen, Yingjie CAI, Weichao Qiu, and Ying-Cong Chen. Disenvisioner: Disentangled and enriched visual prompt for customized image generation. In *Proceedings of International Conference on Learning and Representation*, 2025. 2

[3] Dongxu Li, Junnan Li, and Steven C. H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing, 2023. 2

[4] Yuang 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 *Proceedings of International Conference on Learning and Representation*, 2025. 1

[5] 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 Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2

[6] 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 International Conference on Computer Vision*, 2023. 2

[7] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. 2

[8] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, 2024. 3

[9] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, and Zhongliang Jing. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2