Supplementary Materials:
Steering Guidance for Personalized Text-to-Image Diffusion Models

Sunghyun Park* Seokeon Choi* Hyoungwoo Park Sungrack Yun
Qualcomm AI Research†
{sunpar, seokchoi, hwoopark, sungrack}@qti.qualcomm.com

A. Implementation Details

Hyperparameter Details. Fine-tuning was conducted with a batch size of 1 over 500 iterations (except for ClassDiffusion [6]). For experiments involving Stable Diffusion versions 1.5 and 2.1 [8], both training and inference utilized images at a 512×512 resolution, whereas SANA [12] employed a 1024×1024 resolution. In addition, the maximum diffusion timestep is set to 1,000. During inference, the DDIM scheduler [10] was applied to Stable Diffusion 1.5 and 2.1 with 50 inference steps. Conversely, SANA followed its original implementation by employing the Flow-DPM-Solver [11] with 20 inference steps.

For ω that is a scale for weight interpolation, a trade-off exists between subject and text fidelity. Thus, selecting optimal ω should align with the intended objective: values in the range of 0–0.3 are preferable when maximizing subject fidelity, while values around 0.4–0.6 offer a balanced improvement with better preservation of text fidelity.

Details of Personalization Methods. Following the implementation codes provided by Diffusers or the official SANA repository, DreamBooth-LoRA [5, 9] was configured with a rank of 4 and a learning rate of 1e-4 with prior preservation loss. For the combined DreamBooth-LoRA and Textual Inversion [3] approach, the same rank and learning rate were maintained, with the number of learnable textual embeddings set to 2. We utilized the original implementation code of ClassDiffusion [6], with a training batch size of 2, a learning rate of 1e-5, and applied augmentation to the training images. Additionally, Table 9 shows the prompt list we utilized to generate images for evaluation.

Details of Guidances (Baselines). For classifier-free guidance [4], experiments were conducted with the commonly used guidance scale of 7.5. In the case of autoguidance [7], based on Stable Diffusion 2.1, we experimented with guidance scales ranging from 2.0 to 5.0 during inference, reporting the best-performing value (λ = 2.0). Additionally, subject-agnostic guidance [1] was re-implemented, referencing the original paper, specifically for methods employing textual inversion.

User Study. We assessed user preferences between Stable Diffusion 2.1 combined with ClassDiffusion and SANA with DreamBooth-LoRA. Participants were instructed to select images exhibiting the highest subject fidelity and text fidelity. If a clear preference was indiscernible across all image results, they could choose 'undecided.' Each model was evaluated over 15 sets, totaling 30 sets for user preference assessment. The final results were calculated by averaging the preferences across all participants.

B. Details of Application Experiemnts.

(a) Diffusion-DPO We used a Diffusion-DPO model that was fine-tuned from Stable Diffusion 1.5 (SD 1.5). Since Diffusion-DPO is a fine-tuned version of SD 1.5, we constructed the weak model by performing weight interpolation between SD 1.5 and Diffusion-DPO. The interpolation coefficient ω was set to 0.4.

(b) PairCustom For the PairCustom setting, we used a style LoRA model trained using the official repository. Since LoRA is also used in subject personalization tasks, we adopted the same experimental setup. In this case, we set $\omega = 0.0$, meaning that the weak model corresponds to the base SD 1.5 model.

(c) InstructPix2Pix InstructPix2Pix is also a model fine-tuned from SD 1.5 for image editing based on natural language instructions. Therefore, we constructed the weak model by interpolating the weights of InstructPix2Pix and SD 1.5. We set $\omega = 0.0$, which effectively uses SD 1.5 as the weak model. In addition, InstructPix2Pix employs a separate classifier-free guidance (CFG) mechanism. For generating the unconditional output (i.e., using a null text prompt), we used the SD 1.5 model as the weak model.

*These two authors contributed equally to this work.
†Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

	Method	λ	DINO	CLIP-I	CLIP-T
SD 1.5	DB-LoRA	-	0.3741	0.6797	0.2834
	+ CFG	7.5	0.4701	0.7349	**0.3345**
	+ Ours	7.5	**0.4985**	**0.7516**	0.3260
	+ CFG++	0.4	0.4738	0.7352	**0.3321**
	+ Ours++	0.4	**0.5059**	**0.7510**	0.3250
SD 2.1	DB-LoRA	-	0.4202	0.7076	0.2929
	+ CFG	0.4	0.4976	0.7519	**0.3323**
	+ Ours	0.4	**0.5248**	**0.7655**	0.3254
	+ CFG++	0.4	0.5105	0.7531	**0.3304**
	+ Ours++	0.4	**0.5385**	**0.7680**	0.3245

Table 1. Integration with CFG++ and our method.

Method	Max Memory	Inference Time
CFG	6,591 MB	14.5 secs
CFG + SAG	6,595 MB	21.2 secs
CFG + AG	6,595 MB	21.2 secs
Ours	6,591 MB	14.5 secs

Table 2. Comparison of computational cost. Tested using SD 2.1 models with DB-LoRA + TI.

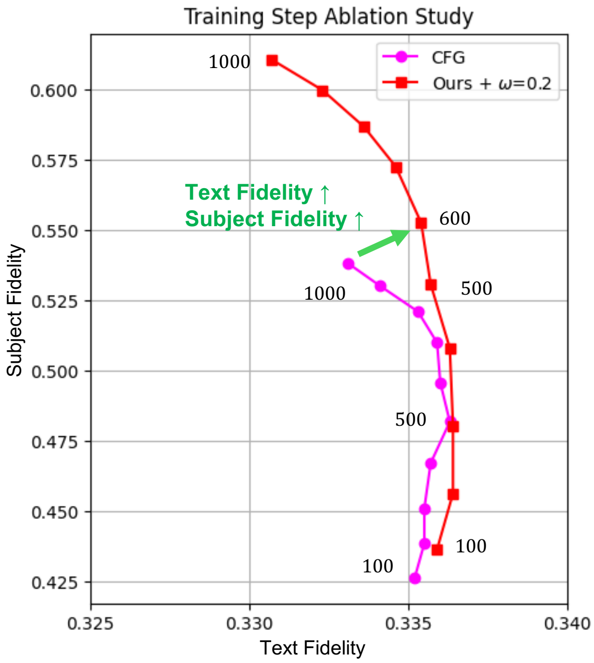

Figure 1. Ablation study on the number of training steps. Notably, our method achieves superior subject fidelity with fewer steps, demonstrating enhanced efficiency.

C. Additional Results

Integration with CFG++ [2]. To address mode collapse issues in classifier-free guidance (CFG), CFG++ [2] has been introduced as an enhancement to the CFG method. CFG++ offers a refined approach to guidance in diffusion models, overcoming several drawbacks of traditional CFG and leading to more reliable and higher-quality text-to-image generation. To demonstrate the applicability of our method, we conducted experiments integrating it with CFG++. Table 1 illustrates that our method significantly enhances subject fidelity not only when combined with CFG but also when integrated with CFG++. In this experiment, we adopted a simple approach by setting $\omega = 0$.

Computational Cost Analysis. We evaluate the computational efficiency of our method by reporting the maximum allocated memory and average inference time. As shown in Table 2, our approach introduces no additional computational overhead during guidance. All measurements were conducted using an A5000 GPU.

Further Study on DB-LoRA. We conducted additional experiments to enhance subject fidelity by increasing the number of training steps and the LoRA rank. Fig. 1 presents the results of DB-LoRA based on SANA with varying numbers of training steps. Notably, DB-LoRA with Ours for 600 steps outperforms DB-LoRA with CFG for 1000 steps in both subject and text fidelity, highlighting the superior efficiency and effectiveness of our approach. Table 3 compares DB-LoRA based on SD 1.5 and SD 2.1 with varying LoRA ranks. Across both models, our method outperforms CFG in terms of subject fidelity. In some cases, even with half the rank (*i.e.*, fewer parameters), our method achieves better subject fidelity.

Detailed Results of Ablation Study on ω. As shown in Fig. 4 in the main paper, we summarize the performance variations in subject fidelity and text fidelity based on different ω values in Table 4, 5, 6. While the graph visualizes only the DINO score, we also include the CLIP-I score in the table for a more comprehensive evaluation. In addition, we report the results on DB-LoRA+TI in Table 7, 8.

Additional Qualitative Results. Fig. 2 and Fig. 3 present additional qualitative results, which demonstrate our guidance's capability of improving subject and text fidelity. These results are generated using the same weights and the same noise seed.

D. Discussion

Dependency on Fine-tuned Models. Our approach relies on the pre-trained model as a weak model to guide the generation toward the direction of the fine-tuned model. As a result, the effectiveness of our method is inherently tied to the quality and learning direction of the fine-tuned model. Specifically, our guidance is most effective when the fine-tuned model has learned a more desirable distribution than the pre-trained model. This dependency highlights the importance of optimizing fine-tuning strategies to ensure robust and meaningful personalization.

Method	Rank	SD 1.5			SD 2.1		
		DINO	**CLIP-I**	**CLIP-T**	**DINO**	**CLIP-I**	**CLIP-T**
+CFG	8	0.4900	0.7445	**0.3335**	0.5154	0.7571	**0.3315**
+Ours	8	**0.5211**	**0.7584**	**0.3335**	**0.5396**	**0.7690**	0.3280
+CFG	16	0.5105	0.7519	**0.3337**	0.5288	0.7637	**0.3302**
+Ours	16	**0.5349**	**0.7633**	**0.3337**	**0.5574**	**0.7781**	0.3285
+CFG	32	0.5468	0.7644	**0.3309**	0.5551	0.7730	**0.3277**
+Ours	32	**0.5763**	**0.7794**	0.3272	**0.5845**	**0.7872**	0.3251

Table 3. Comparison of DB-LoRA with varying ranks on SD 1.5 and SD 2.1. Our method consistently outperforms CFG in subject fidelity.

Base Model	ω	**DINO**	**CLIP-I**	**CLIP-T**
SD 1.5 (DB-LoRA)	0.0	0.4985	0.7516	0.3260
	0.1	0.4987	0.7514	0.3261
	0.2	0.4991	0.7516	0.3264
	0.3	0.5000	**0.7519**	0.3265
	0.4	0.5011	0.7514	0.3271
	0.5	0.5021	0.7512	0.3280
	0.6	**0.5024**	0.7509	0.3288
	0.7	0.5017	0.7495	0.3302
	0.8	0.4979	0.7473	0.3320
	0.9	0.4881	0.7425	0.3333
	1.0	0.4701	0.7349	**0.3345**

Table 4. Ablation study on ω using SD 1.5 with DB-LoRA.

Base Model	ω	**DINO**	**CLIP-I**	**CLIP-T**
SD 2.1 (DB-LoRA)	0.0	0.5248	0.7655	0.3254
	0.1	0.5249	0.7655	0.3253
	0.2	0.5253	0.7656	0.3254
	0.3	0.5260	0.7656	0.3259
	0.4	0.5268	0.7658	0.3263
	0.5	0.5275	0.7658	0.3270
	0.6	**0.5281**	**0.7660**	0.3277
	0.7	0.5272	0.7648	0.3288
	0.8	0.5240	0.7626	0.3299
	0.9	0.5160	0.7591	0.3311
	1.0	0.4976	0.7519	**0.3323**

Table 5. Ablation study on ω using SD 2.1 with DB-LoRA.

Base Model	ω	**DINO**	**CLIP-I**	**CLIP-T**
SANA (DB-LoRA)	0.0	**0.5366**	**0.7522**	0.3357
	0.1	0.5341	0.7512	0.3359
	0.2	0.5308	0.7498	0.3362
	0.3	0.5270	0.7485	0.3364
	0.4	0.5225	0.7474	0.3365
	0.5	0.5178	0.7455	0.3365
	0.6	0.5120	0.7439	0.3367
	0.7	0.5057	0.7420	**0.3368**
	0.8	0.4986	0.7396	**0.3368**
	0.9	0.4906	0.7369	0.3367
	1.0	0.4819	0.7291	0.3363

Table 6. Ablation study on ω using SANA with DB-LoRA.

Base Model	ω	**DINO**	**CLIP-I**	**CLIP-T**
SD 1.5 (DB-LoRA + TI)	0.0	0.4814	0.7459	0.3233
	0.1	0.4815	0.7459	0.3236
	0.2	0.4821	0.7457	0.3239
	0.3	0.4832	0.7456	0.3242
	0.4	0.4847	0.7460	0.3247
	0.5	0.4860	0.7461	0.3254
	0.6	0.4874	0.7454	0.3260
	0.7	**0.4880**	**0.7461**	0.3272
	0.8	0.4867	0.7416	0.3287
	0.9	0.4794	0.7375	0.3306
	1.0	0.4618	0.7292	**0.3316**

Table 7. Ablation study on ω using SD 1.5 with DB-LoRA + TI.

Base Model	ω	**DINO**	**CLIP-I**	**CLIP-T**
SD 2.1 (DB-LoRA + TI)	0.0	0.5683	0.7918	0.3163
	0.1	0.5683	0.7917	0.3164
	0.2	0.5684	0.7913	0.3164
	0.3	0.5688	0.7915	0.3167
	0.4	0.5688	0.7916	0.3172
	0.5	**0.5689**	0.7913	0.3177
	0.6	**0.5689**	**0.7919**	0.3185
	0.7	0.5669	0.7902	0.3194
	0.8	0.5622	0.7877	0.3207
	0.9	0.5523	0.7832	0.3229
	1.0	0.5291	0.7749	**0.3244**

Table 8. Ablation study on ω using SD 2.1 with DB-LoRA + TI.

LIVE Prompt List	**Non-LIVE Prompt List**
’a <*> <subject> in the jungle’	’a <*> <subject> in the jungle’
’a <*> <subject> in the snow’	’a <*> <subject> in the snow’
’a <*> <subject> on the beach’	’a <*> <subject> on the beach’
’a <*> <subject> on a cobblestone street’	’a <*> <subject> on a cobblestone street’
’a <*> <subject> on top of pink fabric’	’a <*> <subject> on top of pink fabric’
’a <*> <subject> on top of a wooden floor’	’a <*> <subject> on top of a wooden floor’
’a <*> <subject> with a city in the background’	’a <*> <subject> with a city in the background’
’a <*> <subject> with a mountain in the background’	’a <*> <subject> with a mountain in the background’
’a <*> <subject> with a blue house in the background’	’a <*> <subject> with a blue house in the background’
’a <*> <subject> on top of a purple rug in a forest’	’a <*> <subject> on top of a purple rug in a forest’
’a <*> <subject> wearing a red hat’	’a <*> <subject> with a wheat field in the background’
’a <*> <subject> wearing a santa hat’	’a <*> <subject> with a tree and autumn leaves in the background’
’a <*> <subject> wearing a rainbow scarf’	’a <*> <subject> with the Eiffel Tower in the background’
’a <*> <subject> wearing a black top hat and a monocle’	’a <*> <subject> floating on top of water’
’a <*> <subject> in a chef outfit’	’a <*> <subject> floating in an ocean of milk’
’a <*> <subject> in a firefighter outfit’	’a <*> <subject> on top of green grass with sunflowers around it’
’a <*> <subject> in a police outfit’	’a <*> <subject> on top of a mirror’
’a <*> <subject> wearing pink glasses’	’a <*> <subject> on top of the sidewalk in a crowded street’
’a <*> <subject> wearing a yellow shirt’	’a <*> <subject> on top of a dirt road’
’a <*> <subject> in a purple wizard outfit’	’a <*> <subject> on top of a white rug’
’a red <*> <subject>’	’a red <*> <subject>’
’a purple <*> <subject>’	’a purple <*> <subject>’
’a shiny <*> <subject>’	’a shiny <*> <subject>’
’a wet <*> <subject>’	’a wet <*> <subject>’
’a <*> <subject> with Japanese modern city street in the background’	’a <*> <subject> with Japanese modern city street in the background’
’a <*> <subject> with a landscape from the Moon’	’a <*> <subject> with a landscape from the Moon’
’a <*> <subject> among the skyscrapers in New York city’	’a <*> <subject> among the skyscrapers in New York city’
’a <*> <subject> with a beautiful sunset’	’a <*> <subject> with a beautiful sunset’
’a <*> <subject> in a movie theater’	’a <*> <subject> in a movie theater’
’a <*> <subject> in a luxurious interior living room’	’a <*> <subject> in a luxurious interior living room’
’a <*> <subject> in a dream of a distant galaxy’	’a <*> <subject> in a dream of a distant galaxy’

Table 9. Evaluation prompt list we used.

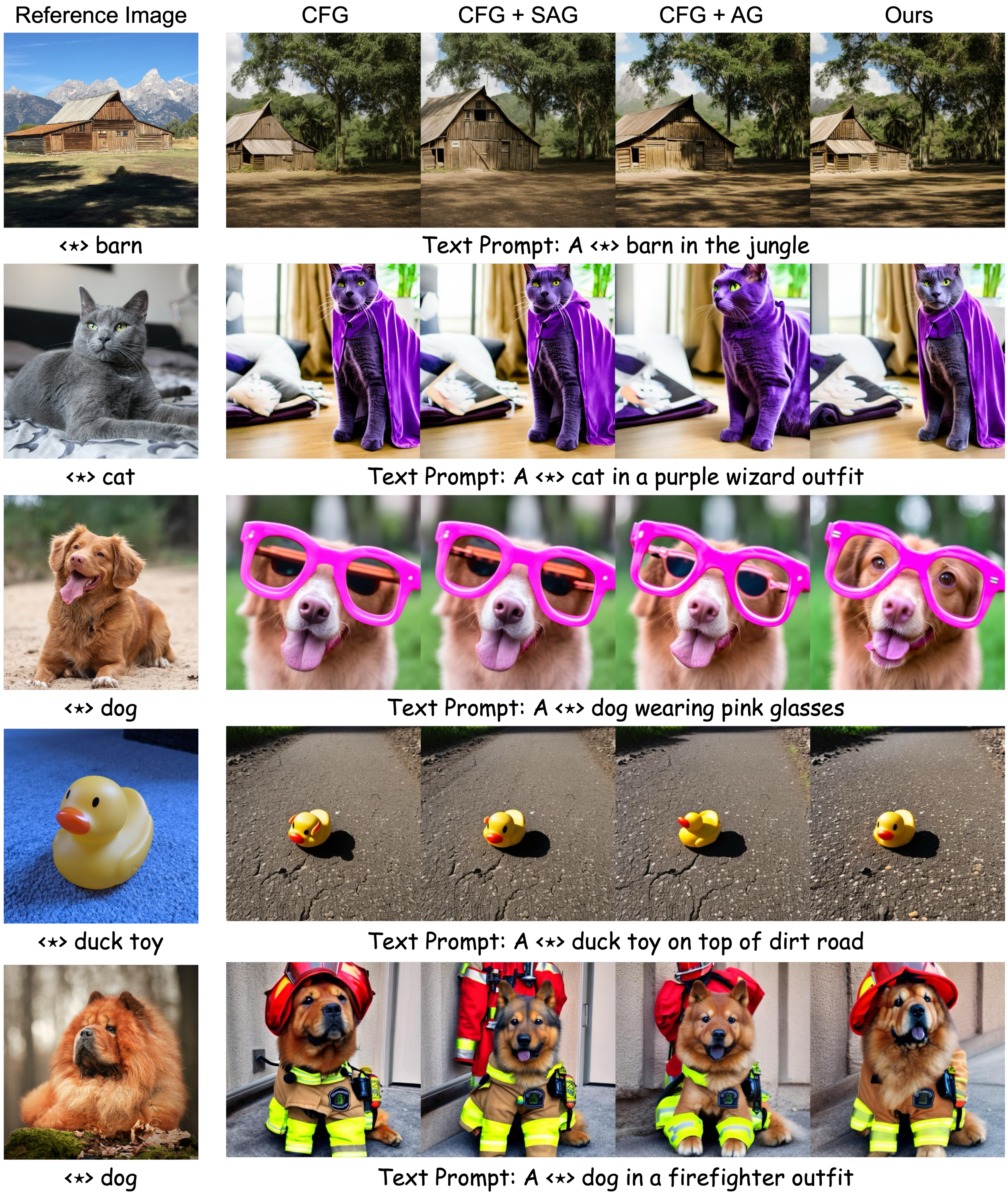
Reference Image
CFG
CFG + SAG
CFG + AG
Ours
<*> barn
Text Prompt: A <*> barn in the jungle
<*> cat
Text Prompt: A <*> cat in a purple wizard outfit
<*> dog
Text Prompt: A <*> dog wearing pink glasses
<*> duck toy
Text Prompt: A <*> duck toy on top of dirt road
<*> dog
Text Prompt: A <*> dog in a firefighter outfit

Figure 2. Comparison with other guidance techniques. These images are generated by fine-tuned SD 2.1 [8] using ClassDiffusion.

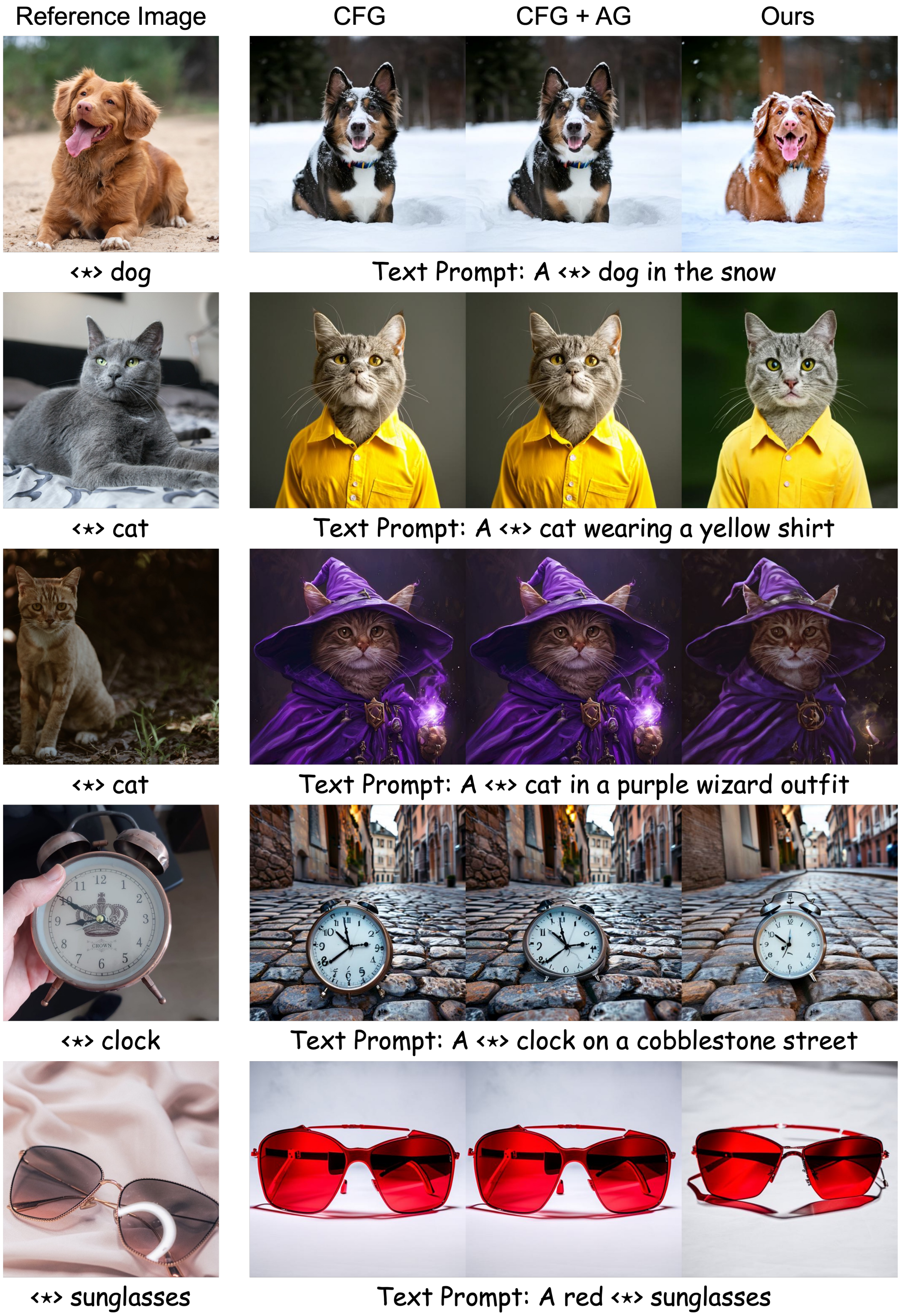
Reference Image
CFG
CFG + AG
Ours
<*> dog
Text Prompt: A <*> dog in the snow
<*> cat
Text Prompt: A <*> cat wearing a yellow shirt
<*> cat
Text Prompt: A <*> cat in a purple wizard outfit
<*> clock
Text Prompt: A <*> clock on a cobblestone street
<*> sunglasses
Text Prompt: A red <*> sunglasses

Figure 3. Comparison with other guidance techniques. These images are generated by fine-tuned SANA [12] using DB-LoRA.

References

[1] Kelvin CK Chan, Yang Zhao, Xuhui Jia, Ming-Hsuan Yang, and Huisheng Wang. Improving subject-driven image synthesis with subject-agnostic guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6733–6742, 2024. 1

[2] Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024. 2

[3] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*. 1

[4] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1

[5] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 1

[6] Jiannan Huang, Jun Hao Liew, Hanshu Yan, Yuyang Yin, Yao Zhao, and Yunchao Wei. Classdiffusion: More aligned personalization tuning with explicit class guidance. *arXiv preprint arXiv:2405.17532*, 2024. 1

[7] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2025. 1

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 5

[9] 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 *CVPR*, pages 22500–22510, 2023. 1

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[11] Kaiyu Song and Hanjiang Lai. Leveraging previous steps: A training-free fast solver for flow diffusion. *arXiv preprint arXiv:2411.07627*, 2024. 1

[12] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. 1, 6