Efficient Text-Guided Convolutional Adapter for the Diffusion Model

Supplementary Material

1. Additional Experiments

Table 1. Evaluation of models on the CUB-200 dataset for fine-grained image generation under *Sketch* and *Depth* conditions. FID (↓) and CLIP (↑) scores are reported under two configurations: *Without Finetuning (w/o FT)* and *With Finetuning (w FT)*.

Method	Model	Sketch		Depth	
		FID	CLIP	FID	CLIP
w/o FT	ControlNet [7]	28.18	24.03	28.81	24.68
	T2I-Adapter [4]	27.84	23.52	29.37	24.03
	ControlNet++ [1]	29.54	24.35	28.36	24.37
	CtrLoRA [6]	28.83	23.03	28.80	22.76
	UniCon [2]	27.73	24.41	28.79	24.13
	Nexus Slim Adapter	28.44	24.65	27.84	24.47
	Nexus Prime Adapter	27.14	24.91	26.31	24.39
w FT	ControlNet [7]	26.25	26.48	27.29	26.62
	T2I-Adapter [4]	25.88	26.14	27.55	23.93
	ControlNet++ [1]	27.47	26.88	26.70	24.37
	CtrLoRA [6]	26.98	25.64	27.29	22.08
	UniCon [2]	25.96	26.21	26.19	23.60
	Nexus Slim Adapter	26.81	26.14	25.79	24.15
	Nexus Prime Adapter	25.72	26.69	25.25	24.46

To validate whether the proposed Nexus Adapters can handle complex image generation in fine-grained tasks, we conduct two evaluations. First, we test our model on the CUB-200 dataset [5] and compare it against baselines. Second, we fine-tune all models for $1k$ steps to measure adaptability and efficacy. Prior work, such as CtrLoRA [6], suggests that ControlNet with a LoRA-based architecture adapts quickly with limited data (1k images). However, our observations show that fidelity often degrades in complex generations, with noticeable loss of color and spatial consistency. We fine-tune all architectures under two distinct conditions—*Sketch* and *Depth*—and report results in Table 1.

The results highlight that the adaptation capabilities of the proposed Nexus Adapters are largely on par with later state-of-the-art models such as ControlNet++ [1], CtrLoRA, and UniCon. Without fine-tuning, most models face challenges in maintaining detail and consistency, but fine-tuning consistently improves performance across both *Sketch* and *Depth* conditions. Nexus Slim and Prime Adapters demonstrate strong competitiveness, narrowing the gap with heavier baselines while retaining their lightweight design advantages. These observations suggest that adapter-based approaches can match the adaptability of established architectures while offering a more efficient path for fine-grained image generation tasks.

2. Training and Inference Details

Both Nexus Prime and Nexus Slim architectures are trained for 200,000 steps at an image resolution of 512×512 using mixed-precision (*fp16*) training, which significantly reduces memory footprint and speeds up computation without degrading performance. A batch size of 2 is employed, with gradient accumulation steps set to 4, effectively simulating a batch size of 8 to stabilize training under GPU memory constraints. The optimizer is AdamW [3], configured with a constant learning rate of 5×10^{-6}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, and a weight decay of 1×10^{-2}. A short linear warm-up of 500 steps ensures stability in the early training phase, after which a constant learning rate schedule is maintained.

Table 2. Estimated training and inference time across models. Training time is computed for $200k$ steps with batch size 2.

Model	Train Time (hrs)	Infer Time (ms/img)
ControlNet	≈ 129	≈ 38
T2I-Adapter	≈ 33	≈ 9
ControlNet++	≈ 129	≈ 38
CtrLoRA	≈ 150	≈ 45
UniCon	≈ 124	≈ 37
Nexus Slim Adapter	≈ 26	≈ 7
Nexus Prime Adapter	≈ 37	≈ 11

During inference, we adopt a standard classifier-free guidance strategy with a guidance scale of 7.5, which balances fidelity to the text prompt against structural adherence to the conditioning input. Sampling is performed with 35 denoising steps, offering a practical trade-off between image quality and computational efficiency. All experiments were conducted on a high-performance server equipped with an *AMD EPYC 7763 64-Core* CPU, 128 GB of RAM, and *2 NVIDIA A100 80 GB* GPUs, running *CUDA 12.2*. This configuration enables efficient large-scale training and reliable evaluation of the proposed architectures.

Table 2 summarizes the estimated training and inference costs across models, measured over 200K steps (batch size 2) and average per-image latency on a single A100 GPU. ControlNet and ControlNet++ are the most resource-intensive, each requiring ∼129 hours of training and 38 ms per image. CtrLoRA is even heavier, at 150 hours and 45 ms, due to the overhead of low-rank updates. UniCon slightly reduces cost (124 hrs, 37 ms), while T2I-Adapter is far more efficient, training in only 33 hours with 9 ms inference. Nexus Slim achieves the best efficiency (26 hrs, 7 ms), and Nexus Prime balances performance with moderate cost (37 hrs, 11 ms). For fairness, inference is reported with 35 steps across all models.

3. Additional Ablations

Number of Adapter Blocks: To evaluate whether adapter blocks in deeper UNet layers remain necessary—where structural information may diminish and the model relies more on the text prompt—we perform an ablation varying

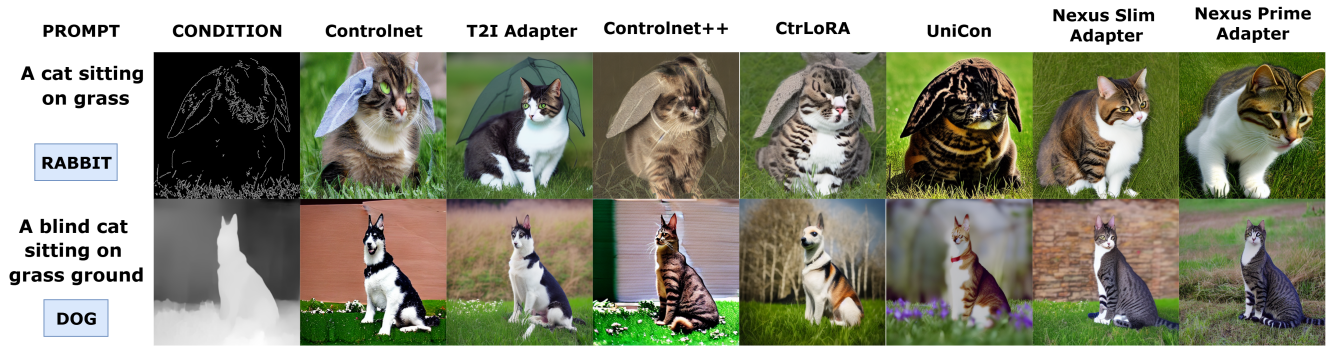

Figure 1. Qualitative ablation on conflicting prompts and conditional images. The actual object in the image-conditioning input is highlighted with a blue background, and a conflicting prompt is provided as test input. We can observe that the model preserve the image-conditioned structural input while generating the object based on the text prompt.

Table 3. Ablation on the number of adapter blocks. FID (↓) and CLIP Score (↑) for Depth and Sketch conditioning tasks. The 4-block configuration represents the full model.

Blocks	Model	Depth		Sketch	
		FID↓	CLIP↑	FID↓	CLIP↑
2	Nexus Slim Adapter	33.43	17.69	32.88	18.35
	Nexus Prime Adapter	28.52	21.98	30.50	20.49
3	Nexus Slim Adapter	30.38	20.75	29.95	20.86
	Nexus Prime Adapter	26.11	22.30	27.42	21.53
4 (ours)	**Nexus Slim Adapter**	**25.30**	**26.71**	**26.40**	**26.86**
	Nexus Prime Adapter	**23.91**	**27.68**	**24.73**	**27.66**

the number of adapter blocks (2, 3, and 4) in both Nexus Slim and Nexus Prime Adapters. Table 3 reports FID (↓) and CLIP Score (↑) for Depth and Sketch conditioning. Increasing the number of blocks consistently improves both FID and CLIP scores, with the full 4-block configuration (ours) achieving the best performance. This demonstrates that deeper adapter blocks effectively preserve fine-grained structural details while still integrating prompt guidance, validating that the final layers contribute meaningful improvements rather than being redundant. These results confirm that careful design of adapter depth is critical for maintaining structural fidelity and semantic alignment in conditional image generation.

Conflict through Prompt: In the Figure 1 we presents a qualitative ablation on conflicting prompts and conditional images, where the actual object in the conditioning input is highlighted in blue. This study validates that incorporating textual information in the conditioning is crucial for generating coherent images with proper alignment to both shape and structure, alongside accurate subject mapping. The Nexus Prime Adapter consistently produces outputs that respect the conditioning structure while adapting to the prompt, yielding coherent and realistic results. Nexus Slim preserves structural fidelity but slightly sacrifices fine details. ControlNet and ControlNet++ often overfit to the condition, ignoring the prompt and producing semantically inconsistent outputs. CtrLoRA captures basic structure but introduces blurring and artifacts, while UniCon generates plausible objects yet struggles with realism under conflicting prompts. Overall, these results emphasize the superior balance of prompt and condition integration achieved by the Nexus adapters compared to existing baselines.

References

[1] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. In Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VII, page 129–147, Berlin, Heidelberg, 2024. Springer-Verlag. 1

[2] Xirui Li, Charles Herrmann, Kelvin CK Chan, Yinxiao Li, Deqing Sun, and Ming-Hsuan Yang. A simple approach to unifying diffusion-based conditional generation. In The Thirteenth International Conference on Learning Representations, 2025. 1

[3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2017. 1

[4] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models. In Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. AAAI Press, 2024. 1

[5] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. Technical Report. 1

[6] Yifeng Xu, Zhenliang He, Shiguang Shan, and Xilin Chen. CtrloRA: An extensible and efficient framework for controllable image generation. In The Thirteenth International Conference on Learning Representations, 2025. 1

[7] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In IEEE International Conference on Computer Vision (ICCV), 2023. 1