Unconditional Priors Matter! Improving Conditional Generation of Fine-Tuned Diffusion Models

Supplementary Material

In this supplementary material, we first provide additional evidence for the fine-tuned models' poor unconditional priors by quantitatively showing that the base model has better unconditional generation quality than the fine-tuned models in Sec. A. In Sec. B, we include more details about the experimental setups for Zero-1-to-3, Versatile Diffusion, DiT, DynamiCrafter, and InstructPix2Pix. We include more qualitative results in Sec. E and more ablation studies on the CFG scale in Sec. C. Finally, we provide details on the inference speed and memory cost of our method in Sec. D.

A. Quantitative Evaluation of Unconditional Samples

In the main paper, we argued that the poor unconditional priors from the fine-tuned models degrade the quality of the conditional generation. We qualitatively showed in Fig. 2 of the main paper that the fine-tuned models exhibit poor unconditional generation quality. In this section, we quantitatively show that the base models have better unconditional generation quality than the fine-tuned models. We unconditionally sample 5000 images from each of `SD1.4`, `SD2.1`, `PixArt-`α, Zero-1-to-3, Versatile Diffusion, and InstructPix2Pix, and evaluate the image quality using Inception Score (IS) [12]. The results are shown in Tab. 1. We observe that the fine-tuned models indeed have quantitatively worse unconditional generation than the base models. Thus, in the main paper, we proposed replacing the poor unconditional noise from the fine-tuned models with the good unconditional noise from the base model which improves the conditional generation quality.

Method	IS ↑
`SD1.4`	**14.085**
`SD2.1`	12.640
`PixArt-`α	9.224
Versatile Diffusion	2.704
Zero-1-to-3	9.140
InstructPix2Pix	5.852

Table 1. **Image Model Unconditional Generation.** We sample using the unconditional noise predictions from each model. The unconditional samples from `SD1.4`, `SD2.1`, and `PixArt-`α are higher quality than those of the fine-tuned models. (**bold** represents the best performance.)

B. Experiment Details

For all experiments, we use the DDIM [13] sampler. When applying our method to a base model with a different variance schedule, we use the fine-tuned model's variance schedule as the reference and choose the base model timestep that yields the closest available variance to the fine-tuned model's variance.

B.1. Zero-1-to-3 [10]

We evaulate our method using the Google Scanned Objects (GSO) dataset [4] which consists of over a thousand scanned objects. We render six views for each object at fixed radii and elevation with azimuths uniformly spaced 60° apart from each other. The first view is used as the reference image and Zero-1-to-3 is used to generate the remaining five images for evaluation. We use 50 steps of DDIM and a CFG scale of $\gamma = 5.0$.

B.2. Versatile Diffusion [16]

We use the COCO-Captions [9] 2014 validation set as the ground truth dataset. We randomly select 30,000 images from the validation set as input conditions to Versatile Diffusion and compute the FID and $\text{FD}_{\text{DINOv2}}$ against the *full* validation set. We use 50 steps of DDIM and a CFG scale of $\gamma = 2.0$.

B.3. DiT [11]

We sample the images using $\gamma = 1.5$ and 50 steps of DDIM. The base model used is DiT-XL/2 trained on ImageNet 256×256 [3]. The fine-tuning is done on each of the datasets using 20,000 steps with batch size 64 and learning rate 0.0001. To account for the impact of random variation, we compute the FID three times and report the minimum, as done by Karras et al. [8]. We provide additional details on each of the dataset below.

SUN397 [14] SUN397 [14] is a dataset used for testing algorithms for scene recognition consisting of 108,754 images distributed among 397 categories.

Food101 [1] Food101 [1] consists of 101,000 images split among 101 food categories. Each category contains 250 test images and 750 training images.

Caltech101 [5] Caltech101 [5] contains images of objects belonging to 101 classes, containing 9,145 images in total.

γ	3.0	4.0	5.0	6.0	7.0	8.0
Zero-1-to-3 [10]	0.192	0.170	0.182	0.179	0.178	0.178
Ours w/ `SD1.4`	0.170	0.165	0.163	0.163	0.161	0.161
Ours w/ `SD2.1`	**0.165**	**0.161**	**0.158**	**0.159**	**0.158**	**0.160**
Ours w/ `PixArt-`α	0.173	0.171	0.169	0.168	0.171	0.170

Table 2. **Zero-1-to-3 [10] (CFG Scales).** We report the LPIPS [17] (lower is better) of applying our method to Zero-1-to-3 using various CFG scales (**bold** represents the best, and underline represents the second best method).

Each class contains between 40 and 800 images with a typical edge length of between 200 and 300 pixels.

B.4. DynamiCrafter [15]

We sample 256×256 resolution videos using 50 steps of DDIM with a CFG scale of $\gamma_T = 7.5$ and $\gamma_I = 1.5$. Although the original paper uses a CFG scale of $\gamma_T = \gamma_I = 7.5$, we find that their choice of CFG scale results in mostly static images, as shown in their low dynamic degree of 40.57% in the VBench benchmark [7]. In contrast, the baseline DynamiCrafter with our choice of CFG scale has a higher dynamic degree of 59.59%.

B.5. InstructPix2Pix [2]

We evaluate the performance of InstructPix2Pix (IP2P) using the EditEvalv2 benchmark [6] which consists of 150 high quality images with edits from 7 categories.

IP2P uses a dual text-image CFG formulation:

$$\begin{aligned}\epsilon_\theta(\mathbf{x}_t, c_I, c_T) = \epsilon_\theta(\boldsymbol{x}_t, \emptyset, \emptyset) \\ + \gamma_I(\epsilon_\theta(\mathbf{x}_t, c_I, \emptyset) - \epsilon_\theta(\mathbf{x}_t, \emptyset, \emptyset)) \\ + \gamma_T(\epsilon_\theta(\mathbf{x}_t, c_I, c_T) - \epsilon_\theta(\mathbf{x}_t, c_I, \emptyset))\end{aligned} \quad (1)$$

For our method, we replace the IP2P *fully* unconditional score $\epsilon_\theta(x_t, \emptyset, \emptyset)$ with the unconditional score from `SD1.5` or `SD2.1`. We use 100 steps of DDIM with a CFG scale of $\gamma_I = 1.5$ and $\gamma_T = 7.5$.

C. Choice of CFG Scale

In this section, we provide an ablation study on the choice of CFG scale γ for Zero-1-to-3 [10] and Versatile Diffusion [16]. The results shown in Tab. 2 and 3 indicate that our method consistently yields improved results across CFG scales.

D. Memory and Inference Speed

As shown in Tab. 4, the inference speed is only slightly affected by our method.

E. Additional Qualitative Results

We provide additional qualitative results for Zero-1-to-3 (Fig. 1), Versatile Diffusion (Fig. 2), DiT (Fig. 3), DynamiCrafter (Fig. 4), and InstructPix2Pix (Fig. 5).

γ	2.5	3.0	4.0	5.0	7.5
Versatile Diffusion [16]	37.96	40.19	42.07	42.33	44.80
Ours w/ `SD1.4`	**35.67**	**35.24**	**35.45**	**35.60**	**36.07**
Ours w/ `SD2.1`	38.29	37.44	37.83	38.44	37.71
Ours w/ `PixArt-`α	37.55	39.03	39.62	40.24	40.89

Table 3. **Versatile Diffusion [16] (CFG Scales).** We report the FID-5k (lower is better) of applying our method to Versatile Diffusion using various CFG scales (**bold** represents the best, and underline represents the second best method).

Method	Memory (GB)		Speed (seconds/sample)	
	Baseline	**Ours**	Baseline	**Ours**
Zero-1-to-3 [10]	4.93	10.06	2.92	3.59
VD	5.68	10.80	7.20	8.17
DiT	3.11	5.65	4.24	4.96
IP2P	5.13	10.14	19.45	21.43
DynamiCrafter	19.17	29.03	125.15	142.84

Table 4. Memory and Inference Speed on an RTX3090 using float32 precision.

Figure 1. **Novel View Synthesis with Zero-1-to-3 [10]**. Zero-1-to-3 tends to produce views that have inaccurate lighting, coloring, or shape. Combining Zero-1-to-3 with the unconditional noise from `SD1.4`, `SD2.1`, or `PixArt-`α corrects these inaccuracies.

Figure 2. **Image Variations with Versatile Diffusion [16].** Images generated from Versatile Diffusion tend to be oversaturated and distorted. Combining Versatile Diffusion with the unconditional noise predictions from `SD1.4`, `SD2.1`, or `PixArt-`α corrects these artifacts.

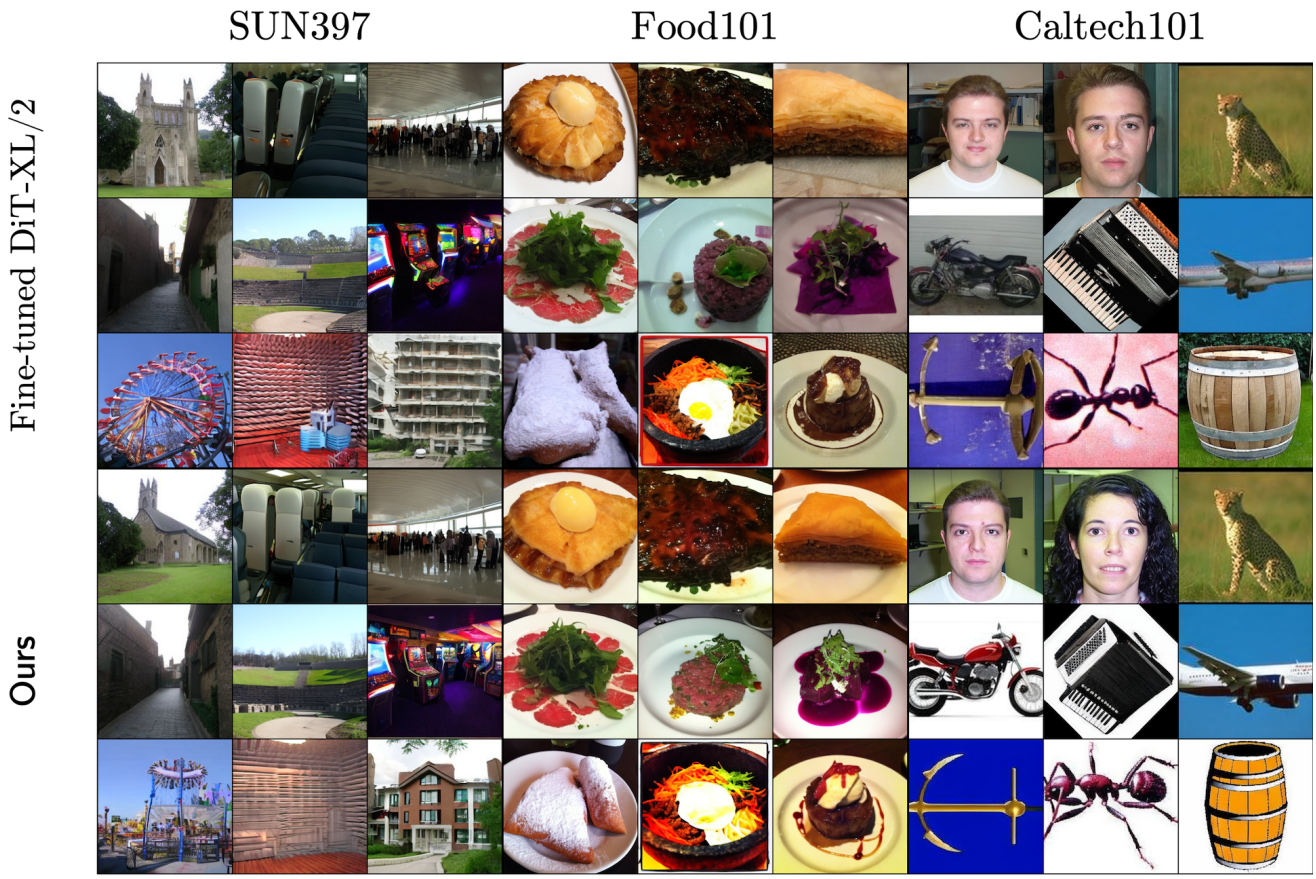

Figure 3. **Class-conditional generation with DiT [11]**. Class-conditional generation using DiT fine-tuned on SUN397 [14], Food101 [1], and Caltech101 [5].

Figure 4. **Image-to-Video Generation with DynamiCrafter [15].** Our method is more temporally consistent (number of horses in the first video and train color in the second video) and less distorted (hand and face in the last video).

Figure 5. **Image Editing with InstructPix2Pix (IP2P) [2].** InstructPix2Pix tends to produce distorted edits. Replacing the IP2P fully unconditional noise with the unconditional noise from `SD1.5`, `SD2.1`, or `PixArt-`α corrects these distortions and improves image quality.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 1, 5

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2, 7

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[4] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 1

[5] Gregory Griffin, Alex Holub, Pietro Perona, et al. Caltech-256 object category dataset. Technical report, Technical Report 7694, California Institute of Technology Pasadena, 2007. 1, 5

[6] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey. *arXiv preprint arXiv:2402.17525*, 2024. 2

[7] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 2

[8] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022. 1

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[10] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 1, 2, 3

[11] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1, 5

[12] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 1

[13] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1

[14] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 1, 5

[15] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 2, 6

[16] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023. 1, 2, 4

[17] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2