NitroFusion: High-Fidelity Single-Step Diffusion through Dynamic Adversarial Training

Supplementary Material

A. Additional Implementation Details

Timestep Shift: Following prior works [7] and our base models, DMD2 [55] and Hyper-SD [39], we adopt the timestep shift technique, shifting the original $T = 1000$ to 500 and 250. NitroSD-Realism and -Vibrant are trained on timesteps $\{250, 188, 125, 63\}$ and $\{500, 375, 250, 125\}$, respectively, for multi-step generation. Both models were trained over approximately 20 NVIDIA A100 days.

User Study Details: We evaluate user preferences using 128 prompts from the LADD [45] subset of PartiPrompts [57], gathering 2,884 votes from 170 participants.

B. Additional Comparison

In Table 2, we provide additional quantitative comparison on FFHQ [22] and ImageNet [10] validation set, captioned with LLM for T2I evaluation. We further use PFID [25] to assess fine-grained texture fidelity. We find NitroSD-Realism achieves the best FID and Patch FID on both datasets, outperforming all base- lines. This implies high degree of photorealism in generated human faces, similar to high quality face images in FFHQ. Figure 9 demonstrates NitroSD-Realism to present high-frequency detail in face features, compared to prior methods' overly smooth textures.

Model	CLIP (↑)	FID (↓)	Patch FID (↓)	Aesthetic Score(↑)	Image Reward(↑)
SDXL-Turbo	0.327/0.327	41.38/32.67	30.52/47.06	5.50/5.40	1.340/0.664
SDXL-Lightning	0.312/0.315	66.44/34.60	47.05/59.25	6.02/5.70	0.912/0.382
Hyper-SDXL	**0.336**/0.324	73.62/34.44	53.18/46.96	**6.68/5.82**	**1.723/1.075**
DMD2	0.331/**0.329**	45.22/28.98	32.19/51.02	5.68/5.42	1.309/0.711
NitroSD-Realism	0.335/0.327	**36.88/28.59**	**23.79/45.25**	5.63/5.54	1.317/0.780
NitroSD-Vibrant	0.325/0.314	67.97/35.26	51.06/48.82	6.25/5.85	1.591/0.964

Table 2. Quantitative comparisons on FFHQ/ImageNet.

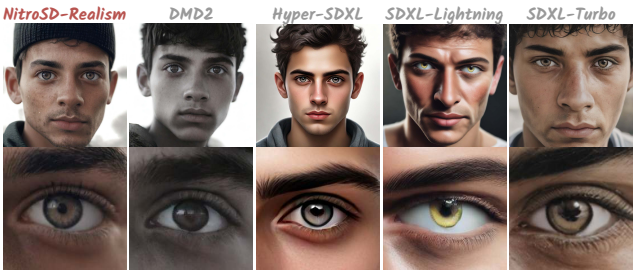

Figure 9. Visual comparison on human face.

C. Additional Ablation Study

The ablation study in Section 4.6 employs the 8-step Hyper-SDXL [39] as the teacher, with 30 hours of training. We conduct additional ablative study on Figure 10, using 4-step DMD2 [55] as the teacher. These results confirm that removing different components leads to artifacts (redundant body) and fine-grained fidelity degradation (clarity of stripes).

Table 3 presents the quantitative results. In particular, we introduce the Teacher PFID metric, which measures the FID score between 299×299 center-cropped patches from student and teacher samples [25], assessing how well high-resolution details are preserved. This metric serves as a critical index for evaluating the effectiveness of GAN training, as it emphasizes the generator's ability to represent fine-grained features and maintain fidelity to the teacher model. Table 3 shows that removing each component causes varying levels of degradation in Patch Teacher FID, highlighting the unique contributions of each to the overall performance of our Dynamic Adversarial framework.

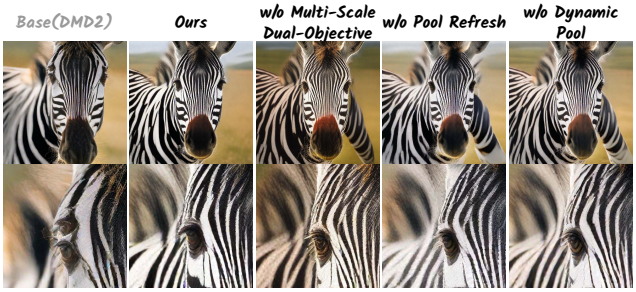

Figure 10. Ablation study (1 A100 training day, DMD2 as base)

Teacher	Ablative setting	CLIP (↑)	Teacher PFID (↓)	Aesthetic Score (↑)	Image Reward(↑)
DMD2	Our Full	0.323	**16.85**	5.58	0.847
	w/o M-S D-O GAN	0.324	22.01	5.63	0.863
	w/o Pool Refresh	0.323	21.48	5.47	0.801
	w/o Dynamic Pool	0.323	17.81	5.98	0.838
Hyper-SDXL	Our Full	0.315	**18.70**	5.87	1.020
	w/o M-S D-O GAN	0.316	18.99	5.83	1.035
	w/o Pool Refresh	0.316	18.78	5.98	1.054
	w/o Dynamic Pool	0.316	19.46	5.98	1.010

Table 3. Quantitative results of ablation study.

D. Discussion and Limitation

Classifier-Free Guidance (CFG): Like most few-step distillation methods [30, 39], our framework does not support CFG [11, 19]. While we achieve competitive results in one-

Figure 11. 1- to 4-step refinement process of our NitroSD-Realism and -Vibrant, illustrating the progressive enhancement of image quality and detail across steps.

step generation, incorporating CFG could enhance alignment with prompts, particularly for complex or ambiguous text. Future work could focus on integrating CFG into the adversarial framework to enhance controllability.

Training with Natural Images: Training on natural images offers the potential for improved quality by leveraging diverse, high-resolution data beyond teacher-generated samples. However, poorly aligned image-prompt pairs pose a significant risk of text-image misalignment, reducing adversarial training effectiveness. Future research will explore strategies for training with natural images while addressing image-prompt misalignment.

Training Efficiency: Our framework highlights the potential of adversarial training in one-step diffusion distillation, an area that remains underexplored. Future directions include optimizing adversarial strategies, such as more efficient adaptive learning schedules, to further boost training efficiency.

E. Additional Qualitative Results

We provide additional qualitative results in this section. Figure 11 showcases the 1- to 4-step refinement process of NitroSD, while Figure 12 presents further comparisons with baseline methods [25, 36, 39, 44, 55]. Additionally, Figure 13 and Figure 14 include more single-step samples generated by NitroSD-Realism and NitroSD-Vibrant, respectively.

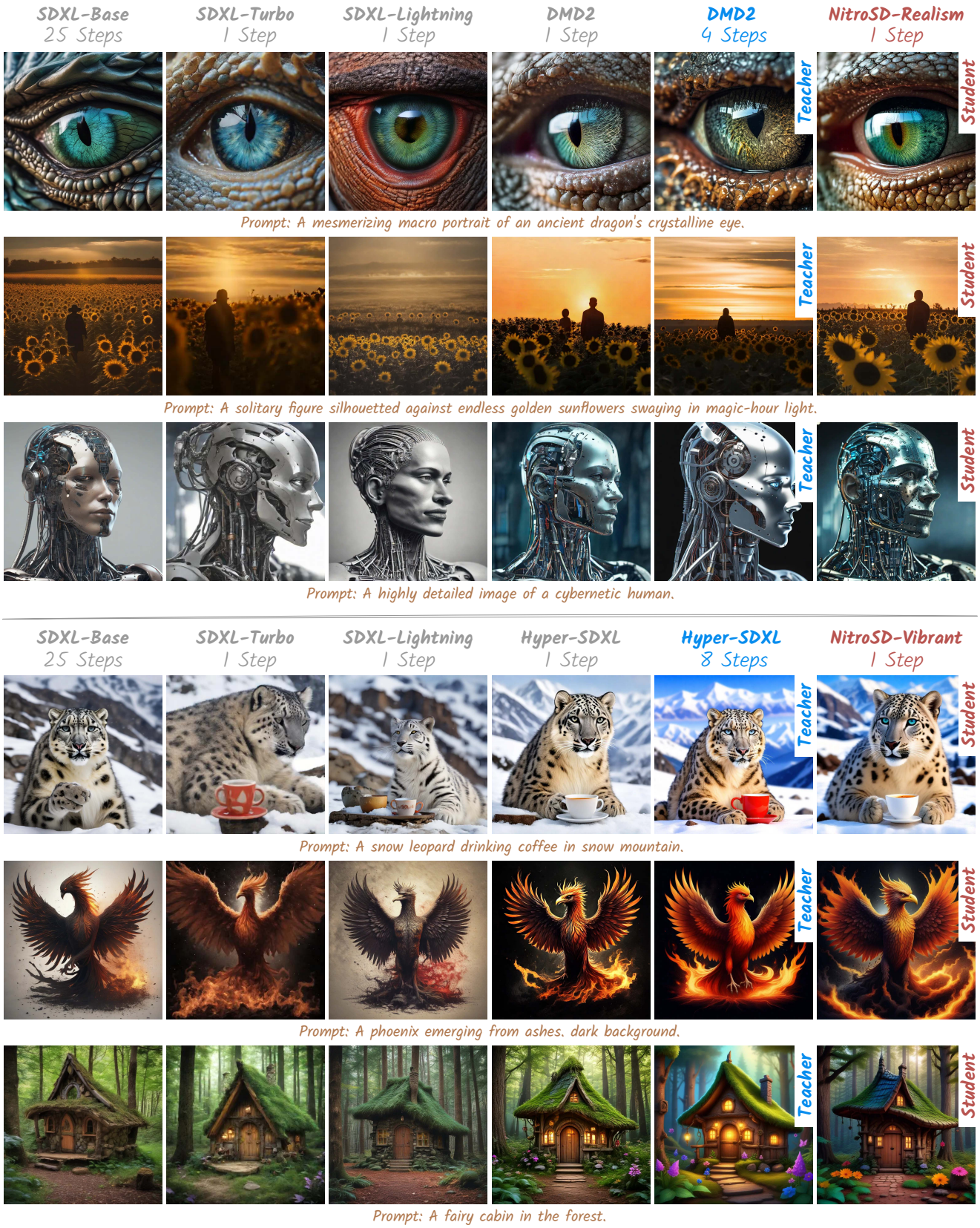

Figure 12. Additional visual comparison with state-of-the-art approaches.

Figure 13. Additional single-step samples from NitroSD-Realism.

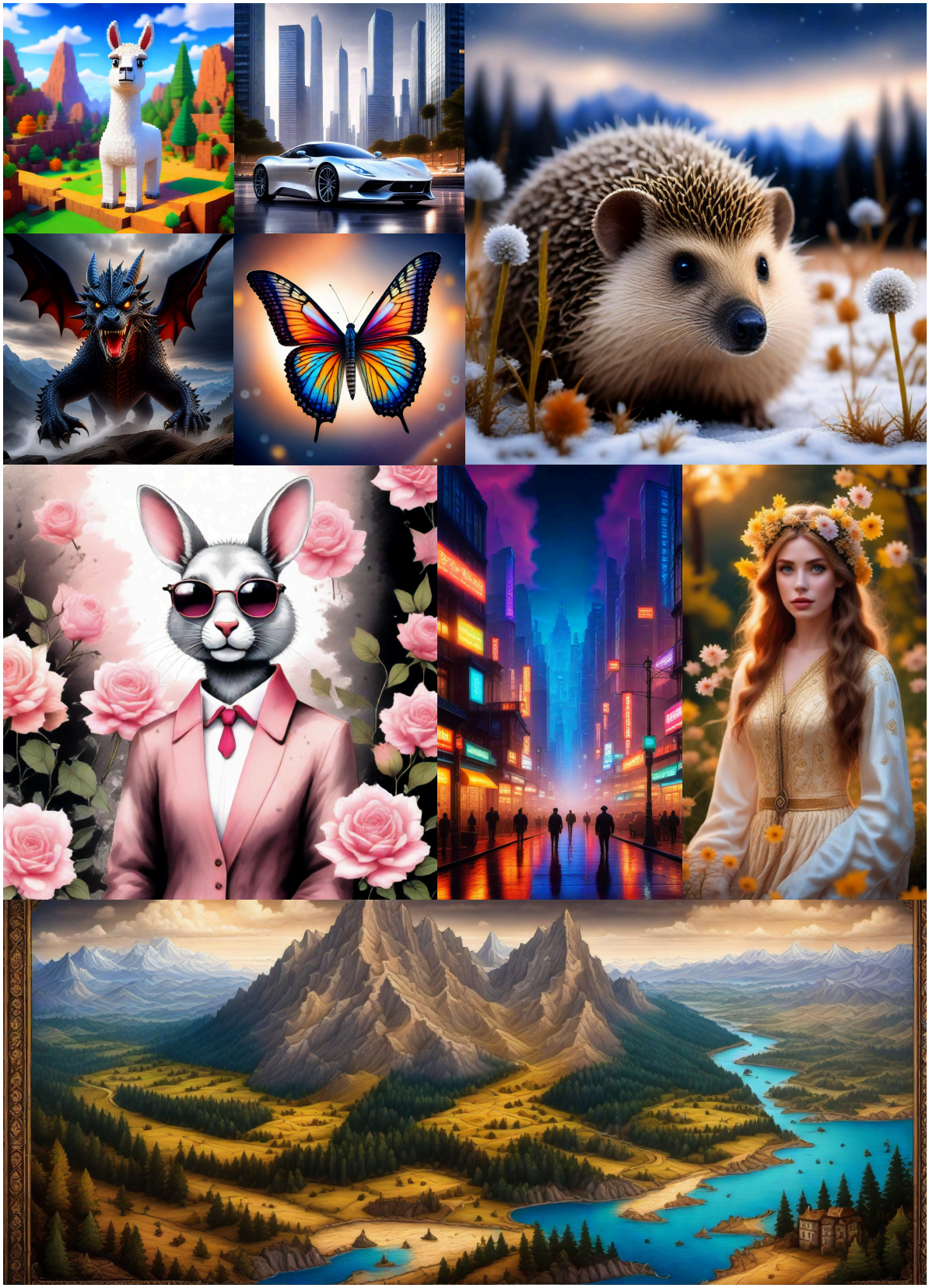

Figure 14. Additional single-step samples from NitroSD-Vibrant.