A BETTER CRITIC MATTERS FOR DISTRIBUTION MATCHING DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Distribution Matching Distillation (DMD) trains a few-step student using the difference between a teacher's score and an online critic's estimate of the student distribution's score. We investigate how critic capacity and conditioning affect distillation. Controlled experiments across seven SiT model sizes on ImageNet show that increasing teacher size while keeping a small critic fixed can degrade student quality and cause training collapse. Scaling the critic alongside the teacher, which we call joint scaling, allows the student to benefit from larger teachers. We also introduce a step-aware critic (STAC), which conditions on the student's generation step to estimate step-specific scores. Under a pure distribution-matching objective, joint scaling and STAC together reduce the FID of a 4-step SiT-B student from 13.91 to 6.90 on ImageNet without CFG. We further validate our findings on text-to-image and text-to-video generation with SDXL, FLUX.2, and Wan2.1. Additionally, we explore critic designs and training strategies, with comprehensive results and ablations reported in the appendix.

1 INTRODUCTION

Diffusion and flow matching models produce high-quality images and videos but need many denoising steps for each generation, which makes distilling them into few-step students central to their deployment. Distribution Matching Distillation (DMD) (Yin et al., 2024a;b) approaches this problem by matching distributions of the student and the teacher. Each student sample is re-noised, and the difference between the teacher and student-distribution scores provides a gradient for updating the student. Since the student distribution is unknown, DMD estimates it with a *critic* network. This teacher-critic recipe is now widely adopted in few-step image generation (Z-Image Team et al., 2026; Zhao et al., 2026a; Liu et al., 2026), real-time autoregressive video (Yin et al., 2025; Huang et al., 2025; Yang et al., 2025), interactive world models (He et al., 2026; Sun et al., 2026; Robbyant Team et al., 2026), and joint audio-video generation (Su et al., 2026; Zhang et al., 2026).

During DMD training, the critic must track the student's evolving distribution while providing the score estimates used to update the student. Prior work improves critic tracking through a two-time-scale update rule (Yin et al., 2024a). The critic is typically initialized from a pretrained diffusion model. When pretrained models of different sizes are available, it remains unclear which critic size is most effective for a given student model. In practice, compute and memory constraints further complicate this choice. For example, the DMD recipe of Self Forcing pairs a 14B teacher with a 1.3B critic and a 1.3B student (Huang et al., 2025), a configuration also used by subsequent methods (Zhu et al., 2026; Feng et al., 2026; Yu et al., 2026), while other studies pair a large teacher with a teacher-sized critic (Zheng et al., 2026a; Yang et al., 2026). These differing choices motivate us to revisit critic design: *what does the critic need, in capacity and in information, for DMD?*

We study teacher and critic scaling while holding the student architecture, the number of function evaluations (NFE) at inference, and the number of training iterations fixed. We extend the SiT (Ma et al., 2024) family to seven pretrained model sizes, from S to H, and systematically evaluate teacher–critic size combinations on ImageNet (Russakovsky et al., 2015) at 256×256 resolution. We also compare generation quality against training FLOPs to assess the computational efficiency of different scaling strategies. Figure 3 illustrates the resulting scaling behavior. With a small critic, increasing teacher size can initially improve student quality, but further scaling can degrade performance or lead to training collapse. Scaling only the critic helps little. In contrast, increasing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **Few-step students distilled with a critic as large as the teacher.** Top, samples of a 4-step FLUX.2 [klein] 4B student distilled with a 9B teacher and a 9B critic w/ STAC. Bottom, three frames each of six videos from a 4-step Wan2.1 1.3B student distilled with a 14B teacher and a 14B critic w/ STAC. (Settings: FLUX.2 512×512, Wan2.1 480p.)

both teacher and critic capacity together, which we call *joint scaling*, enables further improvements with larger teachers (Figure 4). Across this study, all observed training collapses occur with a critic *smaller* than its teacher.

Beyond network capacity, the critic's conditioning also matters. In multi-step DMD, critic training pools student predictions associated with different generation steps k. These predictions can follow different distributions. However, the critic is conditioned on the score-evaluation level t and the class or text prompt, but not on k. It therefore estimates the score of their mixture. We introduce a *step-aware critic* (STAC), which additionally conditions on k to estimate the score of each corresponding student distribution (Figure 2).

We conduct comprehensive experiments spanning class-conditional image generation with SiT on ImageNet, text-to-image generation with FLUX.2 (Black Forest Labs, 2025) and SDXL (Podell et al., 2023), and text-to-video generation with Wan2.1 (Wan et al., 2025). Keeping the FLUX.2 4B student fixed, we improve its GenEval (Ghosh et al., 2023) score from 70.9 with a small teacher and critic to 74.5 by combining joint scaling and STAC (Section 4). Additional experiments examining other critic designs as well as scaling in other distillation methods are also studied and reported in Appendices E and D.3, respectively.

Our contributions are summarized below:

- We systematically study teacher–critic scaling in DMD using seven SiT model sizes on ImageNet, examining both generation quality and training compute. We demonstrate the benefits of jointly scaling the teacher and critic at a fixed student model capacity.
- We introduce a *step-aware critic* (STAC), which conditions on the number of student backward-simulation steps to estimate the score of each corresponding student distribution.
- We validate our findings across image and video generation with SiT, FLUX.2, SDXL, and Wan2.1.

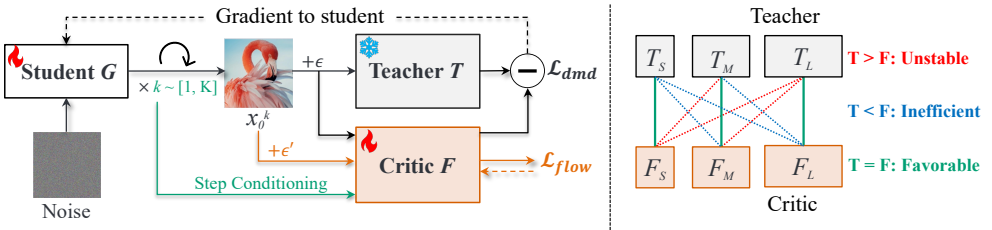


Figure 2: **Overview.** To improve critics, we propose: (a) an extra step conditioning as input to the critic and (b) a comprehensive analysis over the joint scaling behavior of the teacher and the critic.

2 Related Work

Distribution matching distillation. Distribution matching distillation or variational score distillation (Yin et al., 2024b; Nguyen & Tran, 2024; Wang et al., 2023; Luo et al., 2023b; Zhou et al., 2025) distills a 1-step student by matching its output distribution to the teacher's: it keeps a frozen teacher to estimate the real score and trains a learned fake-score *critic* to estimate the student's own distribution, then pushes the student along the difference between the two, with DMD2 extending it to few-step generation (Yin et al., 2024a). A growing body of follow-up work further improves it from various perspectives (Liu et al., 2025; Bai et al., 2026; Huang et al., 2026; Jiang et al., 2026; Kim et al., 2026; Lu et al., 2025; Ge et al., 2026; Chen et al., 2026; Zhao et al., 2026b; Zheng et al., 2026b). It is now widely used across generation tasks, including text-to-image (Z-Image Team et al., 2026; Zhao et al., 2026a; Liu et al., 2026), text-to-video and autoregressive video (Yin et al., 2025; Huang et al., 2025; Yang et al., 2025; Li et al., 2025), and interactive world and audio–video models (He et al., 2026; Su et al., 2026). In this paper, we carefully investigate the design choices of the DMD critic, including its size and input.

Few-step distribution modeling. A multi-step student substantially outperforms a 1-step one by refining its output over several steps, but each step generates from a different intermediate distribution, so the key to few-step quality is to model these intermediate distributions well. Within distribution matching, several methods split the steps to reduce the ambiguity of scoring a mixture of per-step distributions with one critic: score-level intervals in Trajectory Distribution Matching (TDM) (Luo et al., 2025), per-phase experts in Phased DMD (Fan et al., 2026), or step-specific matching along the trajectory (Zhou et al., 2025; Bandyopadhyay et al., 2026). Trajectory distillation instead conditions the student on explicit step information, namely the target step or step size, so the student itself disambiguates the mixture of few-step trajectories (Frans et al., 2025; Song et al., 2023; Salimans & Ho, 2022). In our work, we bring step conditioning into distribution matching: we model each per-step distribution separately with a single critic conditioned on the generation step (STAC), improving multi-step generation quality.

3 The Critic Matters

We study critic design using a 2-D toy problem (in Appendix C) and 256 × 256 class-conditional ImageNet generation with SiT (Ma et al., 2024). Here, we focus on ImageNet. We extend the SiT family to seven sizes, S, S-B, B, M, L, XL, and H, ranging from 33M to 951M parameters, and pretrain each model for 800k updates using a shared flow-matching recipe (Appendix B.1). We use these pretrained base models for DMD with only the distribution matching objective and a common training protocol detailed in Appendix B.2. Appendix A reviews DMD and our notation.

3.1 Scaling up critic capacity

Scaling the teacher. Since in standard diffusion models, generation quality generally improves with model size within an architecture family (Peebles & Xie, 2023; Ma et al., 2024), a natural expectation is that a larger teacher produces a better student in DMD. However, with the critic and student fixed at S, as shown in Figure 3a, FID reaches its lowest value of 28.5 with an M teacher, then worsens to 34.9 with L and 40.2 with XL, before the student collapses with H (FID 247). The same reversal appears

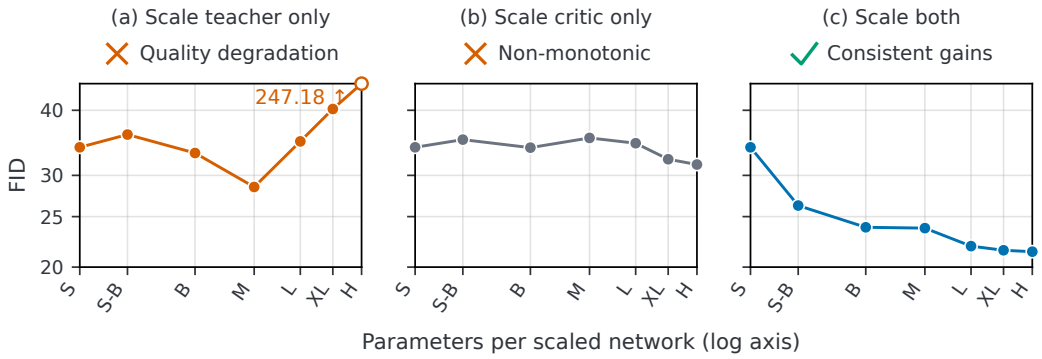


Figure 3: **The joint scaling phenomenon.** Starting from an S-sized teacher, critic, and student and fixing the student size, each panel scales the teacher, the critic, or both along the ladder, with the horizontal axis the parameter count of the scaled network. Scaling only the teacher first improves FID, then degrades it, and finally collapses the student at H, scaling only the critic yields marginal gains, whereas scaling both improves the student at every size. (Settings: 1-step S student, ImageNet 256×256, w/o CFG; best FID with the per-cell learning-rate selection of Appendix B.2.)

for a 1-step B student: the XL/B pair (teacher/critic) performs worse than B/B (Figure 5a). The XL/B run trails B/B from 30k updates onward and reaches FID 41.7 at 200k updates, compared with 16.8 for B/B. Thus, scaling the teacher while holding the student and critic fixed does not necessarily improve the student and can eventually degrade quality or cause collapse.

Scaling the critic. Since scaling the teacher alone does not consistently improve the student, we next scale the critic while holding the teacher and student at S. As shown in Figure 3b, training remains stable as the critic scales up, but FID changes little across the critic ladder. Similarly, for a 1-step B student with a B teacher, the B/XL pair reaches a best FID of 16.6, comparable to 16.8 for B/B (Figure 5a).

		FID↓					
T	F	lowest	100k	200k	IS↑	Precision↑	Recall↑
B	B	16.82	18.78	16.82	99.85±2.36	**0.762**	0.290
XL	B	18.56	21.91	41.75	107.19±1.97	0.735	0.224
B	XL	16.61	18.36	19.12	105.25±2.38	0.698	0.333
XL	XL	**13.50**	**14.72**	**13.50**	**110.14±2.49**	0.760	**0.388**

Table 1: **Training dynamics.** The four runs of Figure 5a, a 1-step B student under B and XL teachers (T) and critics (F) with one recipe: the lowest FID over training, the FID at 100k and 200k updates, and IS, precision, and recall at the lowest-FID checkpoint. XL/XL leads on every metric except precision, where it trails the best by 0.002, and only the two matched pairs reach their lowest FID at the end of training. (Settings: ImageNet 256×256, w/o CFG, default learning rates, 200k updates.)

Joint scaling. We therefore scale the teacher and critic together at matched sizes, which we call *joint scaling*. As shown in Figure 3c, for the 1-step S student, this improves FID at every size, from 34.0 to 21.4 along the ladder. For the 1-step B student, XL/XL achieves the best FID of 13.5, ahead of B/B (16.8) and the mismatched XL/B and B/XL pairs (Figure 5a). Figures 5b and c extend the comparison to a 1-step S student with L, XL, and H teachers and S-to-H critics. In both figure panels, the matched critic gives the lowest FID for each teacher. In Figure 3c, matched pairs remain close to their best FID at 100k/200k updates (also in Table 1), while most mismatched pairs have degraded. Figure 10 shows that every observed collapse occurs with a critic smaller than its teacher, and we further investigate the effect of the critic learning rate in Table 14. Although a critic larger than the teacher can sometimes outperform the matched pair, joint scaling remains among the best configurations across our experiments and is also favorable for compute efficiency (Figure 4).

Explore other settings. Unlike the frozen teacher, the critic must track the student's changing distribution throughout training. A natural remedy is to update the critic more often, as suggested in DMD2 (Yin et al., 2024a). Five critic updates per student update speed up early training, but the student-size critic still settles at an FID well above that of an XL critic (Figure 13 and Appendix E.1).

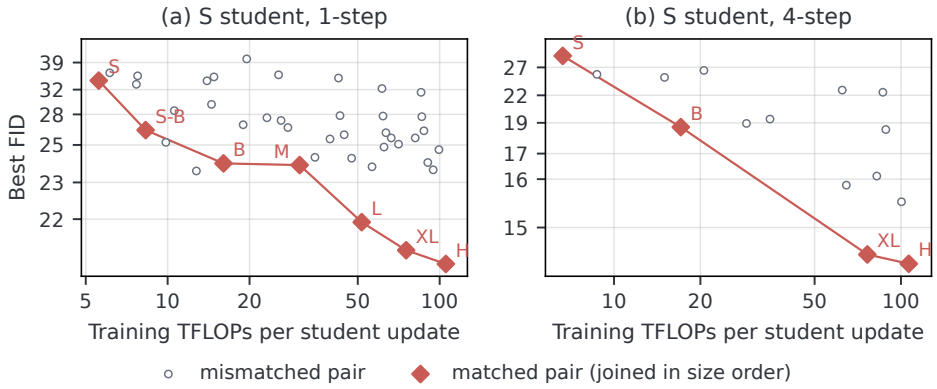


Figure 4: **Training compute versus FID.** Best FID against the training compute of one student update for every non-collapsed teacher/critic pair of the S-student grids at (a) 1 and (b) 4 steps. Matched pairs are labelled by size and joined in size order, and the FID axis is magnified toward the lowest values. For their compute, matched pairs sit at or near the lowest FID, and most mismatched pairs lie above them. (Settings: ImageNet 256 × 256, w/o CFG; learning rates as in Figure 10; FLOPs from Table 6.)

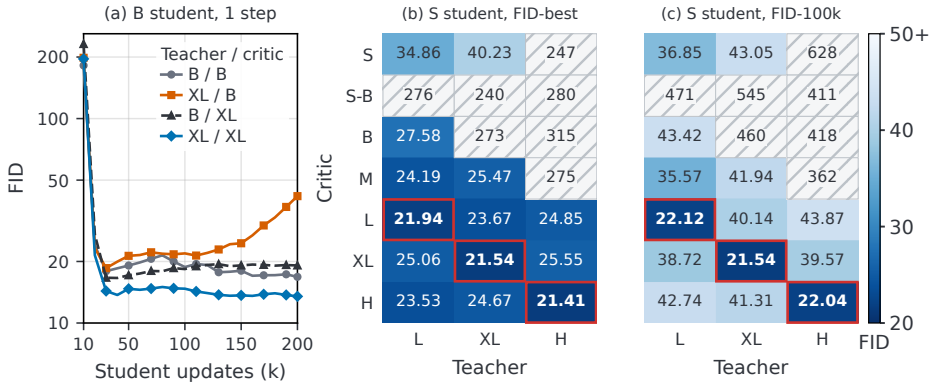


Figure 5: **Scaling the teacher and critic on ImageNet.** Every panel varies the teacher and critic sizes while the student is fixed. **(a)** FID over training for a 1-step B student under four teacher/critic pairs (B/B, XL/B, B/XL, XL/XL); the mismatched XL/B pair degrades only late, so (a) runs to 200k updates to expose it. **(b, c)** A 1-step S student over a grid of L, XL, and H teachers and S-to-H critics: (b) reports the best FID over training, and (c) the FID at a fixed 100k-update budget. Red boxes mark the *diagonal*, matched teacher/critic pairs. For both students, matched pairs give the lowest FID, while mismatched pairs often degrade and sometimes collapse (best FID of 100 or higher). (Settings: ImageNet 256 × 256, w/o CFG; (a) default learning rates; (b, c) the best learning-rate pair per cell by the lowest FID over training.)

Warming up a small critic on student samples before distillation also fails to prevent collapse (Figure 14 and Appendix E.2). A large critic need not be initialized from the teacher's weights. For a 1-step B student distilled from the XL teacher pretrained for 7M updates by Ma et al. (2024), our XL critic pretrained for 800k updates reaches FID 12.01, close to initializing the critic from the teacher, while a B critic reaches only 35.96 (Table 16 and Appendix E.3). Moreover, low-rank adapters on our frozen XL teacher, with under 1% of the trainable parameters of a student-size critic, already outperform that full critic and thus offer a cheap form of joint scaling (Table 17 and Appendix E.4). Velocity knowledge distillation and MeanFlow (Geng et al., 2025; 2026) distillation do not show the same collapse under larger teachers (Appendix D.3), suggesting that the collapse is tied to distribution matching. Together, these results suggest that in DMD, a larger critic helps track the student's evolving distribution when using a larger teacher, without requiring the teacher's exact pretrained weights.

3.2 STEP-AWARE CRITIC (STAC)



Figure 6: **Generation quality varies sharply across steps.** Clean predictions of a 4-step SDXL student along one generation path with a shared prompt and seed. The first step already fixes the overall color and layout but still carries significant artifacts; the second step refines the image substantially, yet artifacts remain in detailed regions such as the face and the wheel; the third step is near-perfect apart from residual artifacts around the mouth, which the last step removes.

Intuition. A K-step student produces clean x_0 predictions with different characteristics at each sampling step. Early predictions often contain blur and artifacts, while later predictions recover finer detail, as illustrated in Figure 6. During multi-step DMD training, these predictions are pooled to train a critic that receives the score-evaluation level t but not the generation step k. The critic therefore estimates the score of a mixture across steps, which can differ from the score of each step's distribution even at the same t.

Step-aware critic. We introduce a *step-aware critic* (STAC) that conditions on the generation step during both critic training and student updates. This allows it to estimate the score of each step's distribution. The loss retains its DMD2 form, while the teacher target, rollout, and sampler remain unchanged. STAC therefore adds no inference cost and little training overhead (detailed in Appendix F.4). The DMD algorithm with STAC is shown in Algorithm 1. The critic now estimates step-specific scores rather than a mixture score, so early and late predictions receive supervision matched to their respective distributions. On ImageNet, as a validation, we train a classifier to predict which generation step produced each student prediction (Figure 15 and Appendix F.2). Its accuracy decreases with STAC, suggesting that predictions from different steps become more similar. This is consistent with STAC providing more effective step-specific score estimates that guide all steps toward the same teacher distribution. We note that TDM also addresses the issue of sharing a critic across steps, but restricts each step to a disjoint range of score-evaluation levels (Luo et al., 2025), whereas STAC uses the same range for every step.

Under the SiT ImageNet setting used in Sec 3.1, we encode the generation noise level r_k of step k as four additional in-context tokens (Geng et al., 2026). As these tokens are learned from scratch rather than inherited from pretraining, we train them at a much higher rate than the critic (Appendix B.2). We evaluate STAC with a 4-step B student and B/B or XL/XL teacher-critic pairs, keeping all other settings fixed (Table 2, left). STAC lowers FID from 13.91 to 11.77 with B/B and from 8.08 to 6.90 with XL/XL, complementing joint scaling. IS and recall improve in both settings, while precision changes by less than 0.03. The gain also persists late in training (Figure 7). For a 4-step XL student with an XL/XL pair at CFG 1.3, FID without STAC worsens after 40k updates, whereas FID with STAC continues to improve and reaches its lowest value at the end of training. STAC also reduces fluctuations in the critic and distribution-matching losses (Figure 16 and Appendix F.2).

Ablation. To separate the effects of step information and learned tokens, we compare three controls for the 4-step B student with an XL/XL teacher-critic pair (Table 2, right). Learned tokens shared across all steps (*const tokens*) reach FID 7.95, compared with 6.90 for STAC, indicating that extra tokens alone explain little of the gain. Step-specific random tokens kept fixed during training (*frozen tokens*) reach 7.76. An *AdaLN branch* (Peebles & Xie, 2023) that injects the step through adaptive normalization reaches 7.68. Its step embedder uses the same learning rate as STAC's conditioning

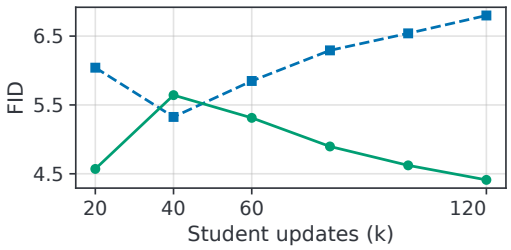


Figure 7: **Training dynamics of STAC.** FID during training of a 4-step student w/o STAC (dashed) and w/ STAC (solid). With STAC, FID improves steadily after 40k updates. (Settings: XL student, teacher, and critic, ImageNet 256×256, CFG $= 1.3$.)

parameters. Both step-aware controls improve on shared tokens, but learned in-context tokens perform best.

T / F	FID↓	IS↑	Prec.↑	Rec.↑
B / B	13.91	116.81±1.98	0.792	0.334
+ STAC	11.77	127.90±1.84	**0.815**	0.340
XL / XL	8.08	161.01±3.75	0.739	0.452
+ STAC	**6.90**	**165.16±3.13**	0.731	**0.470**

Step input to F	FID↓
In-context tokens (STAC)	**6.90**
Const tokens	7.95
Frozen tokens	7.76
AdaLN branch	7.68

Table 2: **Cumulative results and step-input ablations on ImageNet.** Left: a 4-step B student under B/B and XL/XL teacher–critic pairs, w/o and w/ STAC ("+ STAC"); joint scaling and STAC each lower FID, and together they give the best student. Right: the XL/XL student w/ STAC against three controls that replace its learned in-context step tokens, namely *const tokens* shared by all steps, step-specific but random and untrained *frozen tokens*, and an *AdaLN branch* that injects the step through adaptive normalization; every control gives up most of STAC's gain. Every run shares the recipe, and IS is given with its standard deviation over splits. (Settings: ImageNet 256×256, w/o CFG, default learning rates, FID at 100k updates.)

4 Additional Experiments

Beyond the controlled ImageNet setting, we run experiments on text-to-image and text-to-video generation. For text-to-image, we use FLUX.2 (Black Forest Labs, 2025), which provides 4B and 9B models in one family, so the teacher can be scaled alone or together with the critic; we also validate STAC on SDXL (Podell et al., 2023), a denoising UNet-based text-to-image model. For video, Wan2.1 (Wan et al., 2025) pairs a 14B teacher with a 1.3B student, so the critic can grow from the student's size to the teacher's. Within each comparison the student and its number of steps are fixed. We score images with FID, CLIP score (CLIP-S) (Radford et al., 2021), ImageReward (Xu et al., 2023), and HPSv3 (Ma et al., 2025), adding GenEval (Ghosh et al., 2023) for FLUX.2, and evaluate videos on VBench (Huang et al., 2023). Detailed training and evaluation protocols are in Appendices B.3–B.5.

4.1 Text-to-Image Experiments

FLUX.2 (Black Forest Labs, 2025). We distill a 4B FLUX.2 base student with the DMD2 recipe under the 4B/4B, 9B/4B, and 9B/9B teacher/critic pairs, and add STAC to the 9B/9B pair (Table 3, Figure 18). With the 4B critic, the 9B teacher yields a weaker student than the 4B/4B pair—most clearly in FID, ImageReward, and HPSv3—so a larger teacher alone does not help. Scaling the critic to 9B as well improves all five scores of the 9B-teacher student, with FID falling from 20.03 to 15.20, below the 17.71 of the 4B/4B pair. Adding STAC to this pair further improves every score, reaching an FID of 13.93. The bottom rows report the 50-step FLUX.2 4B and 9B teachers for reference. The CFG value used in DMD often differs from that used by the base model; we discuss its effect in Appendix D.2.

SDXL (Podell et al., 2023). We append the four step tokens to the critic's text sequence. Under pure DMD with a 4-step student, STAC improves FID from 34.71 to 32.73 at an unchanged CLIP score (Appendix F.3). We then add STAC to the full DMD2 recipe, with its GAN loss and backward simulation, by continuing training from the official DMD2 checkpoint (Table 4 and Figure 19). In this recipe, the step tokens also enter the critic's GAN branch, while the pure-DMD result above, which has no GAN loss, isolates their effect on the score estimate. With STAC, the continued student improves FID from 18.95 to 16.88 and ImageReward from 71.01 to 80.00, whereas an equally long continuation without STAC worsens all four metrics. This further supports the usefulness of

T	F	STAC	FID↓	CLIP-S↑	ImageReward↑	HPSv3↑	GenEval↑
4B	4B	w/o	17.71	28.24	78.6	**12.65**	70.9
9B	4B	w/o	20.03	27.85	55.1	8.11	70.5
9B	9B	w/o	15.20	28.23	79.2	11.64	71.4
9B	9B	w/	**13.93**	**28.56**	**86.8**	11.91	**74.5**
4B teacher			9.57 (14.00)	28.51 (29.25)	73.4 (101.3)	11.35 (12.74)	68.0 (76.8)
9B teacher			8.05 (10.30)	28.13 (28.78)	66.6 (95.9)	10.43 (12.42)	70.7 (81.5)

Table 3: **FLUX.2 benchmark at** 512×512**.** A 4B student is used with 4-step sampling. The bottom rows are 50-step teachers at CFG 2.0 (4.0 in parentheses). CLIP-S and ImageReward are multiplied by 100. Details in Appendix B.4.

conditioning the critic on the student's generation step. These results are consistent with improved step-specific supervision, although we do not measure score-estimation error.

Method	FID↓	CLIP-S↑	ImageReward↑	HPSv3↑
LCM-LoRA (Luo et al., 2023a)	22.27	31.71	39.56	6.45
SDXL-Turbo (Sauer et al., 2023)	27.27	32.16	46.09	9.09
SDXL-Lightning (Lin et al., 2024)	24.49	32.31	57.48	9.48
Flash (Chadebec et al., 2024)	22.96	31.84	19.04	6.49
PCM (Wang et al., 2024)	24.13	32.52	64.73	9.46
DMD2, official (Yin et al., 2024a)	18.95	33.14	71.01	9.64
Decoupled DMD (Liu et al., 2025)	17.80	**33.62**	78.61	9.79
DMD2, continued w/o STAC	19.49	33.04	70.10	9.15
DMD2 + STAC (Ours)	**16.88**	33.60	**80.00**	**9.84**

Table 4: **Comparison with distilled SDXL systems.** SDXL systems with 4-step sampling on 10k COCO prompts at 512×512. The first seven rows are the values reported by Liu et al. (2025), and our rows follow the same protocol. CLIP-S and ImageReward are multiplied by 100. Continued with STAC, DMD2 improves clearly over the official checkpoint on every metric, whereas the same continuation without STAC falls below it. (Settings: full DMD2 recipe continued from the official checkpoint; Appendix B.3.)

4.2 Text-to-Video Experiments

Wan2.1 (Wan et al., 2025) provides two model sizes, a 14B and a 1.3B model. We scale the teacher and the critic while keeping the 1.3B student fixed. We evaluate the distilled four-step 1.3B student models on VBench (Huang et al., 2023).

We first isolate the effect of STAC in the simplest setting (Figure 8). The teacher, critic, and student are all 1.3B, and training uses neither a GAN loss nor backward simulation, so the step conditioning of STAC is the only change. For bidirectional video distillation with standard multi-step DMD (Figure 8, top), STAC consistently improves the distillation quality: the multi-step DMD baseline peaks near 1k iterations and then degrades, whereas STAC keeps improving and attains a higher VBench total throughout training.

We turn to autoregressive video distillation (Figure 8, bottom). The Self Forcing (Huang et al., 2025) and Causal Forcing (Zhu et al., 2026) recipes pair a 14B teacher with a 1.3B critic and student; to keep the comparison clean, we again set the teacher, critic, and student all to 1.3B. Under the autoregressive Self Forcing rollout, STAC attains a higher VBench total at every evaluated iteration and degrades less after prolonged training than the multi-step DMD baseline.

Finally, we study joint scaling using the autoregressive ODE-initialized student of Causal Forcing with the Self Forcing rollout. Across the controlled configurations in Table 5, we hold the student initialization and the remaining training settings fixed while varying the teacher and critic sizes and adding STAC. The 1.3B teacher–critic pair yields the highest Dynamic Degree and the lowest semantic score, with excessive motion and visible artifacts (Figure 20). Scaling only the teacher to 14B leaves the VBench total almost unchanged. Joint scaling to a 14B teacher and critic raises the VBench total to 84.26. Compared with teacher-only scaling, it improves both semantic and quality scores and partially recovers motion. Adding STAC yields small further gains.

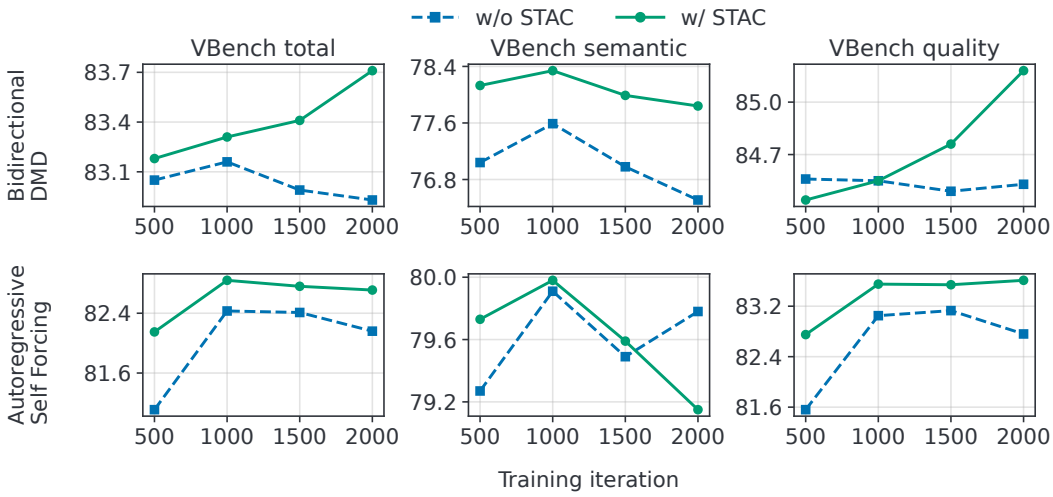


Figure 8: **Controlled comparison of STAC on Wan2.1.** VBench total, semantic, and quality scores over training for bidirectional DMD (top) and autoregressive Self Forcing (bottom), w/o and w/ STAC. In both regimes, the run w/ STAC leads in VBench total at every evaluated iteration. (Settings: 1.3B teacher, student, and critic; 4 steps, 480p VBench; no GAN loss. The bidirectional rows use their standard DMD construction, while the autoregressive rows use online Self Forcing backward simulation.)

Configuration	Total↑	Quality↑	Semantic↑	Dynamic↑
CausVid (Yin et al., 2025)	82.76	83.49	79.85	–
Self Forcing (Huang et al., 2025)	83.87	84.59	81.00	50.12
Causal Forcing (Zhu et al., 2026)				
$T = 14$B, $F = 1.3$B (Default)	83.94	84.73	80.79	52.06
$T = 1.3$B, $F = 1.3$B	83.96	85.14	79.26	81.39
$T = 14$B, $F = 14$B	84.26	85.04	81.16	57.52
+ STAC (Ours)	84.32	85.08	81.32	58.06

Table 5: **Critic size and STAC on Wan2.1.** Autoregressive Wan2.1-T2V-1.3B students. The controlled rows below the Self Forcing row share the student initialization of Causal Forcing (Zhu et al., 2026). The row marked Causal Forcing also uses its teacher and critic sizes, which follow the default DMD recipe of Self Forcing, and the last row adds STAC to the 14B critic. A critic matched to the 14B teacher recovers part of the motion and raises every score, and STAC brings a further gain. (Settings: 4 steps, 480p VBench; Appendix B.5.)

5 Discussion

Through extensive experiments, we show that few-step distillation using DMD can be improved by better designs on the critic. First, we study how to choose the critic's model size, given the differing design choices across current DMD practices. Controlled experiments on ImageNet and distillation results on production-level models suggest that setting the critic's model size to match the teacher's is a favorable and compute-efficient choice. Furthermore, we argue that the critic should model the conditional distribution of the student's output given the step information in a few-step distillation setting. This improves the distillation result as well as optimization stability since the critic does not have to model a generic joint distribution of the student's outputs at different steps.

Our scaling results are empirical comparisons under the tested architectures, training budgets, and learning-rate search space; they do not establish that matched teacher–critic sizes are universally optimal. The best-FID grids include learning-rate and checkpoint selection and should therefore be interpreted as comparisons within a common search protocol. It remains an open problem why a larger teacher alone does not provide enough signal to consistently improve the performance of the student under the current DMD setting. We provide a description of the problem, and a prescriptive solution in future work might result in a more efficient distillation algorithm. Furthermore, as step conditioning breaks the symmetry between the critic and the teacher, future work may investigate other asymmetric designs that could improve the current DMD framework.

AI Use Statement

The authors are responsible for the content and verification of the work. Generative AI tools assisted with experimental-design feedback, refinement of hypotheses, literature search, figure preparation, editing of the manuscript text, and internal review.

Ethics Statement

The work studies training methods for image and video generators. Faster generation can support creative and research applications, while also lowering the cost of producing deceptive or harmful synthetic media. We use existing model and dataset assets under their applicable terms. Deployment safeguards and detection of misleading media are outside the scope of this study.

Reproducibility Statement

Appendices B–F specify the model ladder, initialization, scaling configurations, the STAC objective, and the diagnostic protocols. Appendices B.3–B.5 give the SDXL, FLUX.2, and Wan2.1 protocols, and Appendix G presents matched visual comparisons.

References

Lichen Bai, Zikai Zhou, Shitong Shao, Wenliang Zhong, Shuo Yang, Shuo Chen, Bojun Chen, and Zeke Xie. Optimizing Few-Step Generation with Adaptive Matching Distillation, 2026. URL https://arxiv.org/abs/2602.07345.

Hmrishav Bandyopadhyay, Xuanchi Ren, Zijian Huang, Jay Zhangjie Wu, Tianshi Cao, Ruilong Li, Bryan Chu, Sanja Fidler, Yi-Zhe Song, and Zian Wang. Context-Matched Distillation: Teacher Causality for Autoregressive Video Distillation, 2026. URL https://arxiv.org/abs/2608.13391.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://github.com/black-forest-labs/flux2, 2025.

Clément Chadebec, Onur Tasar, Eyal Benaroche, and Benjamin Aubin. Flash Diffusion: Accelerating Any Conditional Diffusion Model for Few Steps Image Generation, 2024. URL https://arxiv.org/abs/2406.02347.

Siyi Chen, Shaowei Liu, Yixuan Jia, Zian Wang, Huan Ling, Qing Qu, and Jun Gao. Data-Forcing Distillation: Restoring Diversity and Fidelity in Few-Step Video Generation, 2026. URL https://arxiv.org/abs/2606.18478.

Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased DMD: Few-step Distribution Matching Distillation via Score Matching within Subintervals, 2026. URL https://arxiv.org/abs/2510.27684.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. FLUX-Reason-6M & PRISM-Bench: A Million-Scale Text-to-Image Reasoning Dataset and Comprehensive Benchmark, 2025. URL https://arxiv.org/abs/2509.09680.

Jiaqi Feng, Justin Cui, Yuanhao Ban, and Cho-Jui Hsieh. One-Forcing: Towards Stable One-Step Autoregressive Video Generation, 2026. URL https://arxiv.org/abs/2605.23458.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One Step Diffusion via Shortcut Models, 2025. URL https://arxiv.org/abs/2410.12557.

Xingtong Ge, Xin Zhang, Tongda Xu, Yi Zhang, Xinjie Zhang, Yan Wang, and Jun Zhang. SenseFlow: Scaling Distribution Matching for Flow-based Text-to-Image Distillation, 2026. URL https://arxiv.org/abs/2506.00523.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean Flows for One-step Generative Modeling, 2025. URL https://arxiv.org/abs/2505.13447.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved Mean Flows: On the Challenges of Fastforward Generative Models, 2026. URL https://arxiv.org/abs/2512.02012.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment, 2023. URL https://arxiv.org/abs/2310.11513.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-Game 2.0: An Open-Source Real-Time and Streaming Interactive World Model, 2026. URL https://arxiv.org/abs/2508.13009.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, 2017. URL https://arxiv.org/abs/1706.08500.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance, 2022. URL https://arxiv.org/abs/2207.12598.

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, 2021. URL https://arxiv.org/abs/2106.09685.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion, 2025. URL https://arxiv.org/abs/2506.08009.

Yushi Huang, Xiangxin Zhou, Ruoyu Wang, Chi Zhang, Jun Zhang, and Tianyu Pang. Reinforcing Few-step Generators via Reward-Tilted Distribution Matching, 2026. URL https://arxiv.org/abs/2605.26108.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models, 2023. URL https://arxiv.org/abs/2311.17982.

Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Liuzhuozheng Li, Hengzhuang Li, Xin Jin, David Liu, Changsheng Lu, Zhen Li, Bo Zhang, Mengmeng Wang, Steven Hoi, Peng Gao, and Harry Yang. Distribution Matching Distillation Meets Reinforcement Learning, 2026. URL https://arxiv.org/abs/2511.13649.

Damjan Kalajdzievski. A Rank Stabilization Scaling Factor for Fine-Tuning with LoRA, 2023. URL https://arxiv.org/abs/2312.03732.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a Diffusion Model with a Bad Version of Itself, 2024. URL https://arxiv.org/abs/2406.02507.

Youngjoong Kim, Deokyeong Lee, and Jaesik Park. Distribution Matching Distillation without Fake Score Network, 2026. URL https://arxiv.org/abs/2605.19256.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved Precision and Recall Metric for Assessing Generative Models, 2019. URL https://arxiv.org/abs/1904.06991.

Jiachen Li, Qian Long, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Yang Wang. T2V-Turbo-v2: Enhancing Video Generation Model Post-Training through Data, Reward, and Conditional Guidance Design, 2025. URL https://arxiv.org/abs/2410.05677.

Shanchuan Lin, Anran Wang, and Xiao Yang. SDXL-Lightning: Progressive Adversarial Diffusion Distillation, 2024. URL https://arxiv.org/abs/2402.13929.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, 2015. URL https://arxiv.org/abs/1405.0312.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Hongsheng Li, and Steven Hoi. Decoupled DMD: CFG Augmentation as the Spear, Distribution Matching as the Shield, 2025. URL https://arxiv.org/abs/2511.22677.

Jiaxiang Liu, Zhida Feng, Pengyu Zou, Zhenyu Qian, Tianrui Zhu, Jun Xia, Yuehu Dong, Yanzheng Lin, Honglin Xiong, Anqi Chen, Yunpeng Ding, Jinghui Duan, Lin Gao, Chao Han, Tiechao He, Jiakang Hu, Ranjun Hua, Xueming Jiang, Qingli Kong, Yuting Lei, Tianyu Li, Yunlin Liu, Changling Liu, Yaxin Liu, Yi Liu, Xuguang Liu, Xiaolong Ma, Yan Pan, Yiran Ren, Nan Sheng, Yu Sun, Siyang Sun, Yixiang Tu, Yang Wan, Huanai Wang, Siqi Wang, Yang Wu, Youzhi Yang, Xiaowen Yang, Jianwen Yang, Yehua Yang, Quanwen Zhang, Xinmin Zhang, Haoxin Zhang, Xiang Zhang, Jun Zhang, Qian Zhang, Qiao Zhao, and Qi Zhou. ERNIE-Image Technical Report, 2026. URL https://arxiv.org/abs/2605.25347.

Yanzuo Lu, Yuxi Ren, Xin Xia, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Andy J. Ma, Xiaohua Xie, and Jian-Huang Lai. Adversarial Distribution Matching for Diffusion Distillation Towards Efficient Image and Video Synthesis, 2025. URL https://arxiv.org/abs/2507.18569.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step Latent-free Image Generation with Pixel Mean Flows, 2026. URL https://arxiv.org/abs/2601.22158.

Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. LCM-LoRA: A Universal Stable-Diffusion Acceleration Module, 2023a. URL https://arxiv.org/abs/2311.05556.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-Instruct: A Universal Approach for Transferring Knowledge From Pre-trained Diffusion Models, 2023b. URL https://arxiv.org/abs/2305.18455.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning Few-Step Diffusion Models by Trajectory Distribution Matching, 2025. URL https://arxiv.org/abs/2503.06674.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring Flow and Diffusion-based Generative Models with Scalable Interpolant Transformers, 2024. URL https://arxiv.org/abs/2401.08740.

Yuhang Ma, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards Wide-Spectrum Human Preference Score, 2025. URL https://arxiv.org/abs/2508.03789.

Thuan Hoang Nguyen and Anh Tran. SwiftBrush: One-Step Text-to-Image Diffusion Model with Variational Score Distillation, 2024. URL https://arxiv.org/abs/2312.05239.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing Open-source World Models, 2026. URL https://arxiv.org/abs/2601.20540.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, 2022. URL https://arxiv.org/abs/2112.10752.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2015. URL https://arxiv.org/abs/1409.0575.

Tim Salimans and Jonathan Ho. Progressive Distillation for Fast Sampling of Diffusion Models, 2022. URL https://arxiv.org/abs/2202.00512.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs, 2016. URL https://arxiv.org/abs/1606.03498.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial Diffusion Distillation, 2023. URL https://arxiv.org/abs/2311.17042.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing, 2013. URL https://arxiv.org/abs/1207.6076.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency Models, 2023. URL https://arxiv.org/abs/2303.01469.

Yaofeng Su, Yuming Li, Zeyue Xue, Jie Huang, Siming Fu, Haoran Li, Ying Li, Zezhong Qian, Haoyang Huang, and Nan Duan. OmniForcing: Unleashing Real-time Joint Audio-Visual Generation, 2026. URL https://arxiv.org/abs/2603.11647.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. WorldPlay: Towards Long-Term Geometric Consistency for Real-Time Interactive World Modeling, 2026. URL https://arxiv.org/abs/2512.14614.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 2017. URL https://arxiv.org/abs/1706.03762.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models, 2025. URL https://arxiv.org/abs/2503.20314.

Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, Xiaogang Wang, and Hongsheng Li. Phased Consistency Models, 2024. URL https://arxiv.org/abs/2405.18407.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation, 2023. URL https://arxiv.org/abs/2305.16213.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation, 2023. URL https://arxiv.org/abs/2304.05977.

Lingxiao Yang, Liu Liu, Moran Li, Han Feng, Wenjian Cao, Jiangning Zhang, and Ye Shi. In-Context Forcing: Uncovering Context Effects in Autoregressive Video Diffusion, 2026. URL https://arxiv.org/abs/2608.05237.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time Interactive Long Video Generation, 2025. URL https://arxiv.org/abs/2509.22622.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved Distribution Matching Distillation for Fast Image Synthesis, 2024a. URL https://arxiv.org/abs/2405.14867.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step Diffusion with Distribution Matching Distillation, 2024b. URL https://arxiv.org/abs/2311.18828.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models, 2025. URL https://arxiv.org/abs/2412.07772.

Yonghao Yu, Lang Huang, Runyi Li, Zerun Wang, and Toshihiko Yamasaki. Video-Mirai: Autoregressive Video Diffusion Models Need Foresight, 2026. URL https://arxiv.org/abs/2606.03971.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Qilong Wu, Feng Yu, Zechao Zhan, Chi Zhang, Shifeng Zhang, Ruikai Zhou, and Shilin Zhou. Z-Image: An Efficient Image Generation Foundation Model with Single-Stream Diffusion Transformer, 2026. URL https://arxiv.org/abs/2511.22699.

Jintao Zhang, Kai Jiang, Jintao Chen, Xu Wang, Yang Luo, Yuji Wang, Dechuang Chen, Jungang Li, Chengyang Ye, Marco Chen, Hongzhou Zhu, Min Zhao, Yuxuan Jiang, Zhengkun Huang, Chendong Xiang, Kaiwen Zheng, Haoxu Wang, Xiaohang Wang, Qi Jia, Xin Chen, Yimin Chen, Youhe Jiang, Fangcheng Fu, Zhijie Deng, Fan Bao, Jianfei Chen, and Jun Zhu. Vidu S1: A Real-Time Interactive Video Generation Model, 2026. URL https://arxiv.org/abs/2607.03118.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, Kun Yan, Liang Peng, Lihan Jiang, Niantong Li, Ningyuan Tang,

Shengming Yin, Tianhe Wu, Xiao Xu, Xiaoyue Chen, Xihua Wang, Yan Shu, Yanran Zhang, Yi Wang, Yilei Chen, Ying Ba, Yixian Xu, Yujia Wu, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhendong Wang, Zihao Liu, Zikai Zhou, An Yang, Chen Cheng, Chenxu Lv, Dayiheng Liu, Fan Zhou, Hantian Xiong, Hongzhu Shi, Hu Wei, Huihong Zhao, Ivy Liu, Jianwei Zhang, Jiawei Zhang, Kai Chen, Kang He, Levon Xue, Lin Qu, Linhan Tang, Luwen Feng, Minggang Wu, Minmin Sun, Na Ni, Rui Men, Shuai Bai, Sishou Zheng, Tao Lan, Tianqi Zhang, Tingkun Wen, Wei Wang, Weixu Qiao, Weiyi Lu, Wenmeng Zhou, Xiaodong Deng, Xiaoxiao Xu, Xinlei Fang, Xionghui Chen, Yanan Wang, Yang Fan, Yichang Zhang, Yixuan Xu, Yu Wu, Zhiyuan Ma, and Zhizhi Cai. Qwen-Image-2.0 Technical Report, 2026a. URL https://arxiv.org/abs/2605.10730.

Zhuoran Zhao, Shengju Qian, Tongtong Liang, Xianghao Kong, Songchun Zhang, Junchao Huang, Guian Fang, Xin Wang, Pan Hui, and Anyi Rao. Mask Forcing: Improving Autoregressive Video Diffusion Distillation via Dual-Noise Masking Rollout, 2026b. URL https://arxiv.org/abs/2609.09123.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rCM: A Unified Teacher-Forcing and Self-Forcing Open Recipe for Autoregressive Diffusion Distillation in Streaming Video Generation and Interactive World Models, 2026a. URL https://arxiv.org/abs/2606.25473.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large Scale Diffusion Distillation via Score-Regularized Continuous-Time Consistency, 2026b. URL https://arxiv.org/abs/2510.08431.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity Distillation: Exponentially Fast Distillation of Pretrained Diffusion Models for One-Step Generation, 2024. URL https://arxiv.org/abs/2404.04057.

Mingyuan Zhou, Yi Gu, and Zhendong Wang. Few-Step Diffusion via Score identity Distillation, 2025. URL https://arxiv.org/abs/2505.12674.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal Forcing: Autoregressive Diffusion Distillation Done Right for High-Quality Real-Time Interactive Video Generation, 2026. URL https://arxiv.org/abs/2602.02214.

APPENDIX: TABLE OF CONTENTS

A Background **16**

B Implementation Details **17**

B.1 Architecture and Teacher Pretraining on ImageNet 17

B.2 Distillation Implementation on ImageNet 18

B.3 Distillation Experiments on SDXL 20

B.4 Distillation Experiments on FLUX.2 20

B.5 Distillation Experiments on Wan2.1 21

C A Two-dimensional Example **22**

D More Results on Joint Scaling **24**

D.1 The Full Teacher–Critic Grid 24

D.2 Classifier-free Guidance 25

D.3 Exploring Teacher Scaling on Other Distillation Methods 26

E Exploring the Critic **27**

E.1 Critic Update Ratio 27

E.2 Critic Warm-up 28

E.3 Critic Initialization 28

E.4 Low-rank Adaptation on Critics 29

F More Results on STAC **30**

F.1 Pooled and Step-wise Objectives 30

F.2 Step Information and Optimization Stability 30

F.3 Ablations on Step Conditioning for SDXL 31

F.4 Cost Analysis 32

G More Visualizations **34**

G.1 Qualitative Results on ImageNet 34

G.2 Qualitative Results on FLUX.2 35

G.3 Qualitative Results on SDXL 36

G.4 Qualitative Results on Wan2.1 37

A BACKGROUND

Distribution matching. Distribution Matching Distillation (DMD) trains an efficient student G_θ to match the target distribution learned by a frozen diffusion teacher s_T. Because the exact score of the student's own generated distribution is unknown, DMD introduces an auxiliary network—the fake score model, or the *critic*—to estimate it online. Throughout the paper, we denote the *sizes* (i.e., parameter counts) of the student, teacher, and critic as G, T, and F, respectively.

The training process operates as follows. First, the student generates a sample $x_0 = G_\theta(z, c)$ from noise z and condition c. To supervise this prediction, we corrupt it to a *score-evaluation level* t by adding noise: $x_t = \alpha_t x_0 + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. The student is then updated by moving its prediction along the difference between two scores at x_t: the teacher's score s_T and the student score s_q, which the critic estimates. With exact scores, this gradient minimizes a weighted KL divergence between the student distribution q_t and the teacher distribution p_t (Luo et al., 2023b; Yin et al., 2024b). Specifically, the student parameters θ are updated via:

$$\Delta\theta \;\propto\; \mathbb{E}_{z,c,t,\epsilon}\left[w(t)\,\alpha_t\, J_\theta(x_0)^\top \big(s_T(x_t, t, c) - s_q(x_t, t, c)\big)\right], \tag{1}$$

where $J_\theta(x_0) = \partial x_0 / \partial \theta$ is the Jacobian and $w(t)$ is a nonnegative weight. To obtain s_q, the critic acts as a denoiser $D_F(x_t, t, c)$ with parameters ϕ, trained online in alternation with the student to reconstruct the clean student sample x_0 from its corruption x_t. Its prediction is analytically converted to the score s_F, which substitutes s_q in practical DMD (Yin et al., 2024a;b):

$$\mathcal{L}_F = \mathbb{E}\big[\lambda(t)\|D_F(x_t, t, c) - \mathrm{sg}(x_0)\|^2\big], \qquad s_F = \frac{\alpha_t D_F - x_t}{\sigma_t^2}. \tag{2}$$

Here, sg stops gradients and $\lambda(t)$ weights the fitting loss. (Note that noise- and velocity-prediction networks use their corresponding targets and conversions). This two-score formulation is also foundational to Variational Score Distillation (VSD) and Score identity Distillation (SiD) (Wang et al., 2023; Zhou et al., 2024; 2025), extending the critic's relevance beyond DMD.

Multi-step students. Instead of generating an image in a single pass, a K-step student iteratively refines the sample over K distinct calls. During training (e.g., in DMD2), we simulate this sequential process. At step k, the student receives an input at noise level r_k and predicts a fully clean sample $x_0^k = G_\theta(x_{r_k}, r_k, c)$. To proceed to the next step, this clean prediction is analytically re-noised to a lower noise level r_{k+1}, providing the input for step $k+1$. This progressive rollout is known as *backward simulation*.

Importantly, the objective in Equation 1 can be used to supervise *every* step k. To do so, the step's clean prediction x_0^k is corrupted with fresh noise, $x_t^k = \alpha_t x_0^k + \sigma_t \epsilon$, inducing a distribution q_t^k. A crucial distinction here is that this score-evaluation level t is drawn *independently* of the simulation level r_k. Consequently, two distinct time variables operate simultaneously: r_k acts as a "schedule timestamp" that tracks the step's position within the generation rollout, whereas t determines the intensity of the noise used to evaluate the score-matching gradient from the teacher.

B Implementation Details

B.1 Architecture and Teacher Pretraining on ImageNet

We use seven sizes, the S, B, L, and XL models of SiT (Ma et al., 2024) and three added sizes, S-B between S and B, M between B and L, and H above XL. Table 6 gives the full architecture, and Table 7 gives the shared pretraining recipe. We conduct all ImageNet experiments at a resolution of 256×256. All models use patch size 2, MLP ratio 4, and $32 \times 32 \times 4$ latents from the SD-VAE latent space (Rombach et al., 2022). We evaluate metrics on exponential moving average (EMA) checkpoints. The 800k-step training budget is chosen to compare all sizes under one shared recipe rather than to reach state-of-the-art FID.

Model	Depth	Width	Heads	Head dim.	Params. (M)	GFLOPs Fwd.	GFLOPs Bwd.	FID↓
S	12	384	6	64	33.0	10.9	21.8	50.99
S-B	12	512	8	64	58.3	19.4	38.8	40.64
B	12	768	12	64	130.5	43.6	87.2	29.50
M	18	896	14	64	264.2	89.0	178.0	20.66
L	24	1024	16	64	458.1	155.0	309.9	15.91
XL	28	1152	16	72	675.1	228.8	457.6	14.43
H	32	1280	16	80	951.3	322.8	645.6	12.56

Table 6: **The teacher and critic ladder.** Architecture, forward and backward GFLOPs per sample, and teacher FID of the seven SiT sizes after 800k pretraining updates. Teacher FID falls steadily from S to H. (Settings: ImageNet 256×256, w/o CFG; teacher sampling as in Table 7.)

Setting	Value
Training objective	Flow Matching (Lipman et al., 2023)
Optimizer	AdamW; $\beta = (0.9, 0.999)$; weight decay 0
Learning rate	10^{-4}, constant; no warmup or schedule
Batch size	256
Training length	800k updates
EMA decay	0.9999
Class dropout	0.1
VAE type	`sd-vae-ft-ema`
Teacher sampling	Dormand–Prince (dopri5), 250 steps

Table 7: **Hyperparameter choices of ImageNet teacher pretraining.**

B.2 Distillation Implementation on ImageNet

Pseudocode for one training iteration. Algorithm 1 lists one training iteration of our method, with a detached DMD2-style rollout over the generation schedule $r_1, \ldots, r_K$ with step weights π. On the conditioning side, the critic receives the step k in both updates, whereas the teacher sees the same corrupted prediction without it. The critic update reuses the detached student predictions of the same iteration. Notably, the training procedure decouples the architectures of the student, critic, and teacher, which enables the study of different scaling strategies.

Algorithm 1 One DMD training iteration with STAC.

Require: student G_θ; frozen teacher denoiser D_T; critic D_F with parameters ϕ; schedule $r_1, \ldots, r_K$; step weights π

1: Sample condition c, step $k \sim \pi$, and initial noise
2: Roll the student out through steps $1, \ldots, k-1$ to obtain x_{r_k}
3: $x_0^k \leftarrow G_\theta(x_{r_k}, r_k, c)$
4: Sample the score-evaluation level t and $\epsilon \sim \mathcal{N}(0, I)$; $x_t^k \leftarrow \alpha_t x_0^k + \sigma_t \epsilon$
5: Query $D_T(x_t^k, t, c)$ and $D_F(x_t^k, t, k, c)$; form u by Equation 5
6: Update θ with $\mathcal{L}_G$ of Equation 5
7: Update ϕ with $\mathcal{L}_F$ of Equation 2, fitting $D_F(x_t^k, t, k, c)$ to $\mathrm{sg}(x_0^k)$

A pure DMD baseline. Table 8 shows the two recipe changes that make the distribution matching (DM) loss of Equation 5 alone a strong baseline on ImageNet. We raise the student learning rate to eight times the critic rate and draw the score-evaluation level t from a logit-normal distribution (Esser et al., 2024), $t = \mathrm{sigmoid}(n)$ with $n \sim \mathcal{N}(0, 1)$, instead of a uniform one. Each change lowers FID at every size tested, and we keep both as the ImageNet default, listed with the other shared settings in Table 9.

Learning rates. Unless stated otherwise, ImageNet runs use the default pair $(\eta_G, \eta_F) = (8 \times 10^{-6}, 10^{-6})$ (Table 8). The teacher–critic grids of Figures 5b,c and 10, and the summaries drawn from them in Figures 3 and 4, are the exception. For each of their cells separately, we search

t-sampler	η_G	η_F	FID↓ S	FID↓ B	FID↓ L
Uniform	10^{-6}	10^{-6}	51.16	30.77	12.28
Uniform	8×10^{-6}	10^{-6}	37.13	19.11	8.91
Logit-normal	8×10^{-6}	10^{-6}	**33.96**	**16.82**	**8.28**

Table 8: **The pure DMD baseline.** FID of 1-step students after 200k updates at three sizes, with student, critic, and teacher of equal size, for the uniform and logit-normal samplers of the score-evaluation level and two student learning rates. Each change lowers FID at every size, and the last row is our default ImageNet recipe. (Settings: ImageNet 256×256, w/o CFG, one critic update per student update.)

$\eta_G \in \{4 \times 10^{-6}, 8 \times 10^{-6}, 1.2 \times 10^{-5}\}$ and $\eta_F \in \{5 \times 10^{-7}, 10^{-6}, 2 \times 10^{-6}\}$ and keep the pair with the lowest FID over training. The step tokens of STAC, and the step embedder of the AdaLN-branch control in Table 2, are always trained at *20 times* the critic rate η_F.

Prediction-space implementation. As shown in Equation 1, DMD moves the student along the difference between the teacher's and the critic's scores. Substituting the critic's score s_F for s_q and converting both scores to clean predictions by Equation 2, $s_T - s_F = \alpha_t(D_T - D_F)/\sigma_t^2$, the update becomes

$$\Delta\theta \;\propto\; \mathbb{E}\left[w(t)\,\frac{\alpha_t^2}{\sigma_t^2}\,J_\theta(x_0)^\top\big(D_T - D_F\big)\right]. \tag{3}$$

This is, up to a constant factor, the negative gradient of a regression of the student prediction onto a detached target,

$$\mathcal{L}_G = \tfrac{1}{2}\,\mathrm{MSE}\big(x_0, \mathrm{sg}(x_0 - \tilde{u})\big), \qquad \tilde{u} = w(t)\,\frac{\alpha_t^2}{\sigma_t^2}\,(D_F - D_T), \tag{4}$$

since $\nabla_\theta \mathcal{L}_G \propto J_\theta(x_0)^\top \tilde{u}$. This $\mathcal{L}_G$ is the actual code implementation of DMD. In this form the weight $w(t)$ does nothing but rescale the residual $D_F - D_T$ of each sample, so choosing it is essentially choosing a loss weight. Following DMD (Yin et al., 2024a;b), we set $w(t) = \sigma_t^2/(\alpha_t^2 \cdot \mathrm{avg}\,|x_0 - D_T|)$, which divides the residual by the teacher's own average absolute residual on the same sample across dimensions and thereby normalizes the update magnitude across score-evaluation levels and stabilizes optimization. The implemented update is therefore

$$u = \frac{D_F - D_T}{\max(\mathrm{avg}\,|x_0 - D_T|, 10^{-6})}, \qquad \mathcal{L}_G = \tfrac{1}{2}\,\mathrm{MSE}\left(x_0, \mathrm{sg}(x_0 - u)\right), \tag{5}$$

where avg averages over the coordinates of each sample, and importantly every quantity forming u is detached. On ImageNet the networks predict velocity on the linear path $x_\tau = \tau x_0 + (1-\tau)\epsilon$ with $\tau = 1 - t$, and $D = x_\tau + (1-\tau)v$ converts each velocity output to the clean prediction used above.

Setting	Value
Objective	Equation 5; no auxiliary regression or GAN loss
Initialization	Pretrained 800k EMA checkpoint of each network's own size
Batch size	64
Optimizer	AdamW with $\beta = (0.9, 0.999)$
Learning rates (η_G / η_F)	8×10^{-6} / 10^{-6} unless stated otherwise
Weight decay	0.01
Gradient clipping norm	10
Critic updates per student update	1
Score-evaluation level	$t = \mathrm{sigmoid}(n)$, $n \sim \mathcal{N}(0, 1)$
EMA decay	0.9999
Step weights π	Uniform over the K steps
Step schedules	2 steps: 999, 499; 4 steps: 999, 666, 333, 124

Table 9: **Hyperparameter choices of ImageNet distillation experiments.** Sizes, learning rates, and budgets of individual comparisons are given with each experiment.

Evaluation metrics. We report Fréchet Inception Distance (FID) (Heusel et al., 2017) between 50k generated and 50k real images and, where stated, Inception Score (IS) (Salimans et al., 2016) and the improved precision and recall (Kynkäänniemi et al., 2019) of the same samples (Tables 1 and 2). Unless noted otherwise, we take either the lowest FID over training or the FID at a fixed number of updates, as indicated for each experiment, and compute the other metrics at the same checkpoint.

Model selection and interpretation. For the teacher–critic grids, each configuration is evaluated using the same nine learning-rate pairs and a nominal budget of 200k student updates per run. Results labelled "best FID" include selection over both learning rates and evaluated checkpoints. They therefore describe the best performance found within this prescribed search, rather than performance under a learning rate and checkpoint fixed independently of evaluation. For fixed-update summaries derived from these grids, we retain the learning-rate pair selected by the best-FID criterion and report its FID at the indicated update.

B.3 Distillation Experiments on SDXL

We conduct our experiments on SDXL 1.0 (Podell et al., 2023) in two settings. The first is a pure DMD ablation (Table 18) that compares how the critic receives the step. The second is our main comparison (Table 4), which continues training the official DMD2 checkpoint under DMD2's full recipe with and without STAC. Both settings train on LAION prompts with aesthetic score above 6.25 (Schuhmann et al., 2022) at 1024×1024 and are evaluated on 10k COCO 2014 prompts (Lin et al., 2015) at 512×512. Table 10 lists the hyperparameter settings.

Setting	Pure DMD ablation	**Main results**
Step schedule	$\{999, 749, 499, 249\}$	
Step input to critic	See Table 18	4 in-context tokens (w/ STAC only)
Learning rates	Student and critic 5×10^{-7}; step tokens 10^{-4}	
Optimizer	AdamW, $\beta = (0.9, 0.999)$; gradient clipping at norm 10	
Batch size	128	
Critic updates per student update	5	
EMA	None	
Teacher / critic CFG	8.0 / 1.0	
Maximum score-evaluation level	0.98	
GAN loss weights (student / critic)	None	5×10^{-3} / 10^{-2}
Backward simulation	No	Yes
Training length	8k updates	20k updates

Table 10: **Hyperparameter choices of SDXL experiments.** Our core comparison table (Table 4) uses the configurations on the right, while ablations in Table 18 use a more lightweight setting on the left.

B.4 Distillation Experiments on FLUX.2

FLUX.2 [klein] base (Black Forest Labs, 2025) provides checkpoints of two sizes, 4B and 9B, which we use directly as model initializations. Training prompts and real images are from the FLUX-Reason-6M dataset (Fang et al., 2025). We evaluate the students at 512×512 by FID, CLIP-S, ImageReward, and HPSv3 on 10k COCO 2014 validation images (Lin et al., 2015), each recaptioned by a vision-language model to a length close to that of the training captions. They are also evaluated by GenEval (Ghosh et al., 2023) on its prompts with *four seeds each*.

In our experiments in Table 3, training follows the SDXL recipe of DMD2 (Yin et al., 2024a), with a diffusion GAN whose discriminator head sits on the critic's backbone. Notably, the step condition enters *only* the critic's score estimate branch without affecting the discriminator branch. In other words, the discriminator sees real and generated images under the same conditioning.

Table 11 lists our detailed settings.

Setting	Value
Initialization	Base model of each network's own size, full fine-tuning
Step schedule	$\{1000, 750, 500, 250\}$
Step input to critic	None (w/o STAC) or 4 in-context tokens (w/ STAC)
Learning rates	Student and critic 5×10^{-7}, step tokens 5×10^{-5}
Optimizer	AdamW, $\beta = (0.9, 0.999)$
Weight decay	0.01
Gradient clipping norm	10
Batch size	8
Critic updates per student update	5
Teacher / critic CFG	2.0 / 1.0
Score-evaluation level (DM loss)	$t \sim \mathcal{U}[0.02, 0.98]$
Score-evaluation level (critic loss)	$t = \mathrm{sigmoid}(n)$, $n \sim \mathcal{N}(0, 1)$, clipped to $[0.02, 0.98]$
GAN loss weights (student / critic)	5×10^{-3} / 10^{-2}
Training length	10k updates

Table 11: **Hyperparameter choices of FLUX.2 experiments.** The settings are shared by the four runs of Table 3.

B.5 Distillation Experiments on Wan2.1

We run two controlled experiment families on Wan2.1-T2V-1.3B (Wan et al., 2025), both with a 4-step student generating 81 frames at 480×832 and evaluated on VBench (Huang et al., 2023). The first is the cleanest comparison of STAC, adding it to a 1.3B teacher, student, and critic under both bidirectional DMD and autoregressive Self Forcing (Figure 8); these runs last 2k training iterations. The second fixes the 1.3B student, varies the teacher and critic sizes, and finally adds STAC to the 14B critic (Table 5). Table 12 gives the full protocol for this central comparison. Figure 8 uses the model sizes, rollout families, and 2k budget stated above.

Training prompts and objectives. We sample text prompts from `vidprom_filtered_extended.txt`, the filtered and extended VidProM prompt list released with Self Forcing (Huang et al., 2025). No real videos or image latents are used: both the student and critic examples are generated online from Gaussian noise. The frozen Wan text encoder embeds the sampled prompt, while the unconditional teacher branch uses the standard Wan2.1 negative prompt. The student is trained only with the sample-normalized DMD objective of Equation 5, with unit coefficient. Independently generated student samples train the critic with the unweighted flow-matching MSE, $\|v_F - (\epsilon - x_0)\|_2^2$. We use no GAN, reconstruction, perceptual, or other auxiliary loss; STAC changes only the critic conditioning.

Rollout, score time, and guidance. Let $r = (1000, 750, 500, 250)$ be the four canonical generation rungs and set $r_5 = 0$. At each training iteration we sample one rung k uniformly and use it for every temporal block. Starting from Gaussian noise, earlier rungs are evaluated without gradient and their clean predictions are re-noised to the next rung; gradient is retained only through the prediction at rung k. The DMD and critic losses then independently sample a raw integer score-evaluation time $u \sim \mathcal{U}\{r_{k+1}, \ldots, 999\}$, apply the Wan flow-time shift

$$t = \mathrm{clip}\left(1000 \frac{5(u/1000)}{1 + 4(u/1000)}, 20, 980\right), \tag{6}$$

and add fresh Gaussian noise at t. Thus the score time is bounded below by the end of the selected generation interval but is not bounded above by r_k. The configuration's teacher guidance coefficient is 3.0: the code forms $D_T^{\mathrm{cond}} + 3(D_T^{\mathrm{cond}} - D_T^{\mathrm{uncond}})$, equivalent to conventional CFG scale 4.0. The critic uses its conditional prediction only, i.e., CFG scale 1.0, and the distilled student itself is sampled without CFG.

What "Causal Forcing" denotes. All controlled autoregressive rows use the online Self Forcing rollout mechanism (Huang et al., 2025): the 1.3B causal student generates blocks of three latent frames in temporal order, and each block attends through the KV cache to its own previously generated clean blocks. For the rows grouped under Causal Forcing (Zhu et al., 2026), we initialize that student from

the official chunk-wise Causal ODE checkpoint and inherit the associated Stage-3 prompt list and chunk-wise four-rung DMD recipe. Therefore, "Causal Forcing" in Table 5 identifies the initialization and surrounding recipe, whereas "Self Forcing rollout" identifies the online backward-simulation operator used during DMD training; they are not two competing labels for the same component.

Setting	**Main results (Table 5)**
Training prompts / real data	`vidprom_filtered_extended.txt`; text only, no real videos or cached real latents
Student	Causal Wan2.1-T2V-1.3B; full fine-tuning
Teacher / critic	Pretrained bidirectional Wan2.1 checkpoints of the sizes in Table 5; teacher frozen, critic fully trainable
Student initialization	Official chunk-wise Causal Forcing Causal ODE checkpoint (Stage 2A)
Rollout / backward simulation	Online autoregressive Self Forcing from Gaussian noise; 21 latent frames, 3 frames per causal block, clean self-generated KV-cache context
Student objective	Sample-normalized DMD, Equation 5, weight 1
Critic objective	Flow-matching denoising MSE on detached student samples, weight 1
Other losses	None (no GAN, reconstruction, or perceptual loss)
Step input to critic	None (w/o STAC) or 32 in-context tokens encoding the selected generation rung r_k (w/ STAC); teacher receives no rung input
Step-token initialization	Token vectors $\mathcal{N}(0, 1/d)$; two-layer SiLU MLP weights $\mathcal{N}(0, 1/d_{\text{in}})$ and zero biases
Generation schedule	$\{1000, 750, 500, 250\}$, warped with Wan flow-time shift 5
Rung sampling	Uniform over four rungs; one shared rung for all temporal blocks in an iteration
Score-evaluation time	Equation 6, sampled independently for the student DMD and critic losses
Teacher / critic guidance	Coefficient 3.0: $D_T^{\text{cond}} + 3(D_T^{\text{cond}} - D_T^{\text{uncond}})$ (conventional CFG 4.0) / conditional only (CFG 1.0)
Learning rates	Student 2×10^{-6} and critic 4×10^{-7}, constant; STAC parameters 2×10^{-4} through iteration 100, cosine-decayed to 4×10^{-7} by iteration 500
Optimizer	AdamW, $\beta = (0.0, 0.999)$; weight decay 0.01; gradient clipping at norm 10; bfloat16 training
Global batch size	64 prompts
Update ratio	One critic update per iteration; one student update every five iterations
EMA	Student decay 0.99; starts at iteration 200 for the size controls and at iteration 0 for the final STAC run; evaluation uses EMA weights
Training / checkpoint selection	Maximum 1.5k iterations for the final STAC configuration; checkpoints every 100 iterations; Table 5 reports the selected 700–800-iteration checkpoints
Inference	Four autoregressive consistency steps per 3-frame block on the shifted schedule; predict x_0, re-noise to the next rung with fresh noise, and cache the final clean block; no external ODE solver and no student CFG
Evaluation	81 frames at 480×832, 16 fps; official VBench prompts, five samples per prompt

Table 12: **Training and inference protocol for the main Wan2.1 experiments.** "Causal Forcing" specifies the student initialization and recipe, while the training rollout itself is the online autoregressive Self Forcing procedure described above.

C A Two-dimensional Example

To provide an intuitive visual understanding of our two core findings, the benefit of joint scaling and the effectiveness of the step-aware critic (STAC), we construct a two-dimensional toy experiment, as shown in Figure 9.

Setup. The target distribution is a single Swiss roll (2.5 turns, maximum radius 3.0) with Gaussian noise ($\sigma = 0.03$). All networks are MLPs, and a *depth-d* network has d residual linear-SiLU blocks. The student has depth 3, and the teacher/critic depths are $3/3$, $6/3$, and $6/6$. For STAC, we add a separate sinusoidal step embedder to the critic. Table 13 lists the remaining settings.

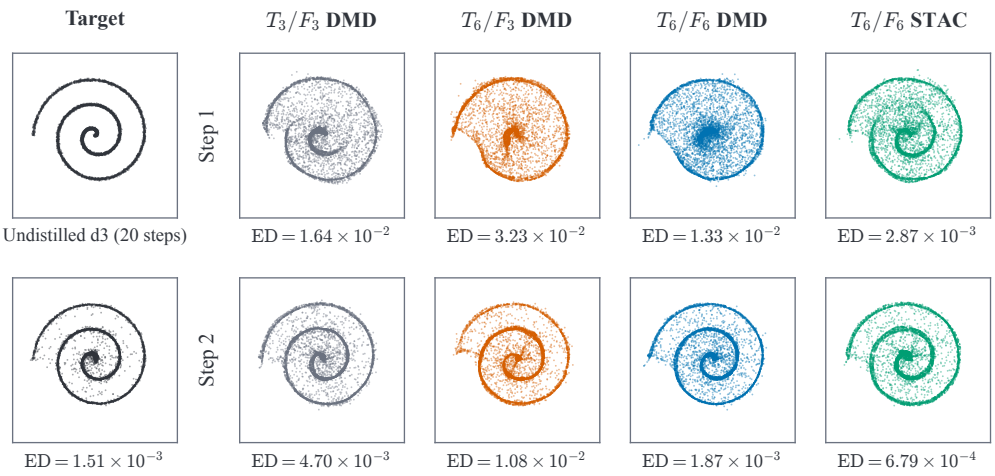


Figure 9: **Joint scaling and STAC on a two-dimensional example.** Samples of a 2-step, depth-3 student at each sampling step (rows), with energy distance (ED) to the target. The first column shows the single-roll target and an undistilled depth-3 teacher sampled with 20 Euler steps. The next three columns vary the teacher (T) and critic (F) depths, and the last adds STAC to the depth-6 pair. Deepening only the teacher makes both steps worse, deepening the critic as well reverses this, and STAC brings both steps closer to the single roll. (Settings: 4k DMD updates, five critic updates per student update; Table 13.)

Setting	Value
MLP width	256
Teacher training steps	12k with batch size 1,024
DMD training steps	4k with batch size 256
Critic updates per student update	5
Optimizer	Adam with $\beta = (0.9, 0.999)$
Teacher learning rate	10^{-3} with cosine decay
DMD learning rates	Critic 10^{-3}, student 10^{-4}, constant
Step noise levels r_k	$\{1, 0.5\}$
Score-evaluation level	$t \sim \mathcal{U}(0, 1)$

Table 13: **Hyperparameter choices of the two-dimensional toy experiment.**

Evaluation. We evaluate all checkpoints with the same initial noise, re-noising streams, and reference set. Generation quality is measured by the energy distance (ED) (Sejdinovic et al., 2013) to the target distribution,

$$\mathrm{ED}(X, Y) = 2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\|, \tag{7}$$

where X, X' are independent generated samples and Y, Y' are independent target samples. Lower ED indicates closer agreement with the target.

Results. The samples reflect our two main findings. First, regarding network capacity (Figure 9, columns 2–4), starting from a matched $3/3$ pair, deepening only the teacher to $6/3$ makes both steps of the student *worse*. However, deepening the critic as well ($6/6$ pair) reverses this degradation.

Second, the samples also expose the step-mixture issue. For all three pairs without STAC, the first step lies farther from the target than the second, suggesting that the two steps follow different distributions. Adding STAC to the $6/6$ pair (Figure 9, column 5) makes the critic step-aware, and both steps then follow the single Swiss roll more closely, with the weaker first step improving the most.

D More Results on Joint Scaling

D.1 The Full Teacher–Critic Grid

Extending the FID grid. Figure 10 extends the teacher–critic scaling grid of Figure 5 to encompass more combinations and student configurations (S and B students, one and four sampling steps). As shown in the figure, a clear *asymmetry* emerges: cells above the diagonal, where the critic is smaller than the teacher, often degrade and sometimes collapse. Conversely, in cells below the diagonal, where the critic is larger than the teacher, no run collapses in our experiments. Empirically, increasing the number of student sampling steps from one to four removes most collapses of the mismatched pairs. These experiments suggest that the critic should *not* be smaller than the teacher, which makes joint scaling a safe default.

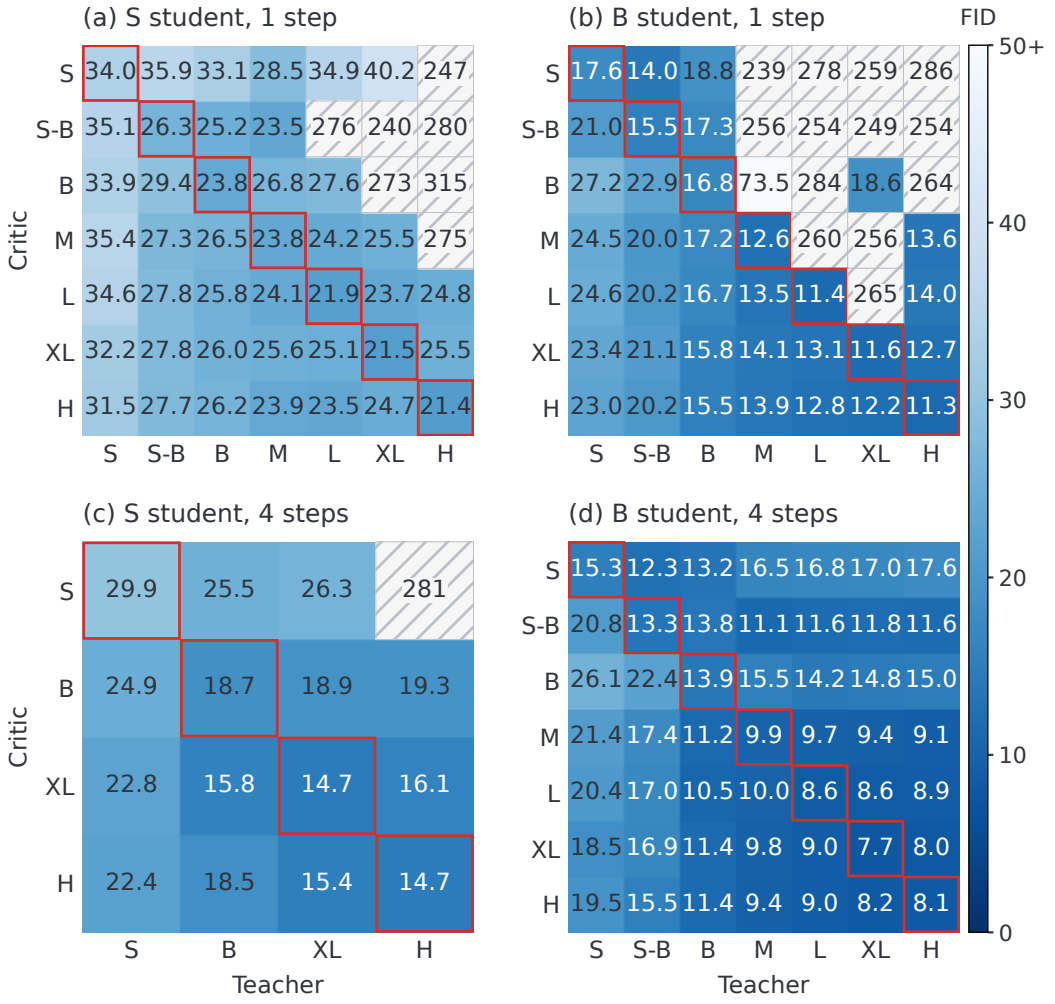


Figure 10: **The expanded teacher–critic scaling grid.** Best FID per teacher/critic pair for S and B students at 1 and 4 steps. Red boxes mark the matched pairs, and hatched cells mark collapsed runs, whose best FID is 100 or higher under all nine learning-rate pairs. Every collapse occurs with a critic smaller than its teacher. (Settings: ImageNet 256×256, w/o CFG, 200k updates; learning-rate selection as in Figure 5b.)

More evaluations on quality and coverage. To confirm that the benefits of joint scaling extend beyond FID, we evaluate the four cornerstone runs of Figure 5a (a 1-step B student under B and XL teachers and critics) using Inception Score (IS), precision, and recall. As shown in Table 1, the jointly scaled XL/XL pair leads on every metric except precision, suggesting better generation quality and coverage.

Training stability. Furthermore, tracking the training dynamics reveals a critical stability advantage. Comparing the lowest FID over training against the FID at 100k and 200k updates shows that only the matched pairs reach their lowest FID at the end of training, whereas the mismatched pairs reach it earlier and then degrade. Figure 17 in Appendix G.1 shows the samples at 200k updates, where the XL/B student in particular loses diversity.

Are the collapses merely optimization artifacts? A natural question is whether the training collapses observed in mismatched pairs are capacity failures or simply artifacts of suboptimal learning rates. In Figure 10, every marked cell fails to achieve an FID below 100 across all nine learning-rate pairs in our search (Appendix B.2). Thus, these failures are not specific to a single learning-rate choice within the tested range, even though this does not establish an intrinsic capacity limit or rule out stabilization through other optimization settings.

To isolate capacity from optimization, we increased the critic's learning rate by a factor of 16 for the H/S pair with an S student, so that the small critic can track the student faster. As Table 14 shows, while this larger learning rate avoids the collapse, the resulting student still performs poorly compared to the matched teacher–critic pairs. Thus, increasing the critic learning rate can prevent collapse in this setting, but does not recover the quality achieved by the matched teacher–critic configuration.

Teacher / critic Critic learning rate	S / S searched	H / S searched	H / S 16× default	H / H searched
S student, 1 step	<u>33.96</u>	247.18	52.53	**21.41**
S student, 4 steps	<u>29.91</u>	281.41	42.89	**14.65**

Table 14: **Forcing a small critic with an extreme learning rate.** Best FID of 1- and 4-step S students under the H teacher with an S critic. We compare the searched learning rates against an aggressively scaled critic learning rate (16× default), alongside the matched S/S and H/H baselines. Although the larger learning rate prevents the collapse, the generation quality stays far behind that of the matched H/H pair. (Settings: ImageNet 256×256, w/o CFG, best FID over training; 16× runs use $(\eta_G, \eta_F) = (8 \times 10^{-6}, 1.6 \times 10^{-5})$; other cells use the best of the nine searched pairs, as in Figure 10.)

D.2 Classifier-free Guidance

In this section, we investigate how classifier-free guidance (CFG) (Ho & Salimans, 2022) interacts with the DMD framework.

DMD prefers smaller guidance scales. We investigate the optimal CFG value for a 1-step B student and for direct sampling from its teacher. We use M, XL, and H teacher–critic pairs, with $F = T$ throughout. As shown in Figure 11, the optimal CFG value for the students across sizes is $\sim$1.3, which is smaller than the optimal value for the teachers ($\sim$1.5). We attribute this shift to the *mode-seeking* nature of the DMD objective. Distribution matching already encourages the student to concentrate on high-density regions of the teacher's distribution, effectively absorbing a portion of the CFG effect (Wang et al., 2023; Luo et al., 2023b).

Note that in DMD each student guidance scale requires a separate distillation run. A natural remedy, successfully employed in recent MeanFlow models (Geng et al., 2026), is to explicitly condition the student on the guidance scale during distillation. We leave this continuous guidance conditioning to future work.

Investigating autoguidance in DMD. As an alternative to standard CFG, autoguidance replaces the unconditional reference with a weaker model evaluated under the same class condition (Karras et al., 2024). In DMD, we use the B model as the weak reference for the M, XL, and H teachers. As shown in Figure 11c, autoguidance yields no measurable improvement over standard CFG in this distillation setting.

Joint scaling under CFG. It is natural to ask whether the benefit of joint scaling, shown without CFG in the main text, persists under CFG. To study this, we fix the CFG scale at 1.2 and the critic at XL, and vary the teacher across the full scaling ladder, for S students at 1, 2, and 4 steps and a B student at 4 steps.

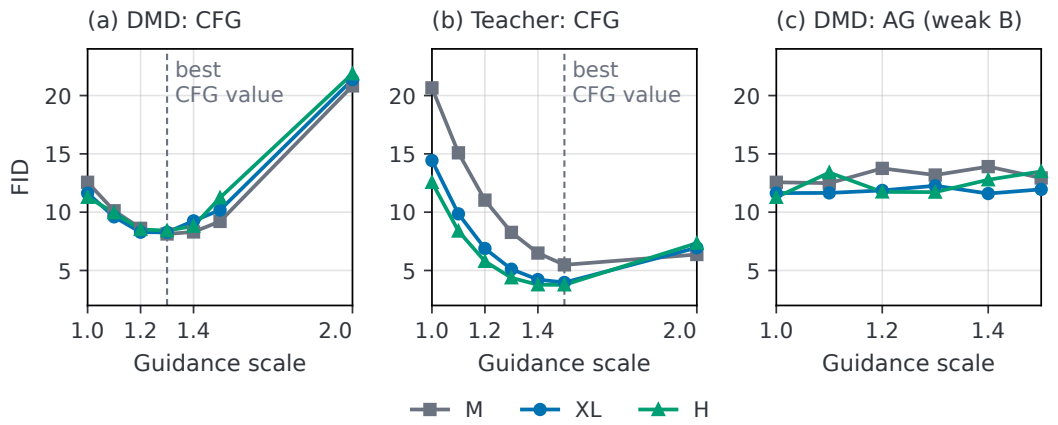


Figure 11: **Guidance for DMD students and their teachers.** FID against guidance scale for (a) 1-step B students distilled under CFG, (b) their teachers sampled with CFG, and (c) students distilled with a weak B model as the autoguidance reference. The students' best CFG scale, about 1.3, lies below the teachers' 1.5, and autoguidance does not improve on CFG. (Settings: ImageNet 256×256, $F = T$ for M, XL, and H; students report their best FID over training.)

As shown in Table 15, in every student configuration the XL teacher, which matches the critic, gives the best student. Scaling the teacher further to H degrades every row, suggesting that joint scaling also holds under CFG.

G	Steps	S	S-B	B	M	L	XL	H
S	1	24.07	20.75	18.55	16.98	19.10	**16.19**	20.17
S	2	18.17	15.85	12.83	10.92	11.69	**10.34**	11.91
S	4	17.47	14.98	11.09	9.72	9.95	**8.64**	10.54
B	4	13.81	11.31	7.05	6.40	6.03	**5.52**	5.76

Table 15: **Joint scaling with classifier-free guidance.** Best FID for each teacher (columns) with the critic fixed at XL. Rows vary the student and its step count. Every row is best at the XL teacher, the one matching the critic, and worse again at the larger H teacher. (Settings: ImageNet 256×256, CFG $= 1.2$; other settings follow Appendix B.2.)

D.3 Exploring Teacher Scaling on Other Distillation Methods

To put our scaling experiments in a broader context, we investigate teacher scaling under two alternative distillation methods: velocity knowledge distillation (KD) and MeanFlow distillation (MFD).

Knowledge Distillation. The most straightforward approach to distill a diffusion teacher is velocity KD, which directly regresses the student's velocity prediction onto that of the frozen teacher. We train an S student using this method from every teacher in our ladder (except S itself) without CFG. As shown in Figure 12a, the student's FID improves substantially as the teacher scales up to the M size, but strictly plateaus thereafter. While the student initially benefits from a stronger teacher, this advantage saturates early.

MeanFlow Distillation. MFD is a few-step distillation framework that learns jumps along the teacher's diffusion trajectory. A MeanFlow model (Geng et al., 2025; 2026; Lu et al., 2026) predicts the average velocity $u(x_t, r, t)$ over an interval $[r, t]$ and is trained via the identity $u(x_t, r, t) = v(x_t, t) - (t - r)\,\frac{\mathrm{d}}{\mathrm{d}t} u(x_t, r, t)$, tying the average velocity to the instantaneous one. During distillation, this instantaneous velocity v is supplied by the frozen teacher rather than the dataset.

We adopt the standard hyperparameter configurations from the official MeanFlow recipe (Geng et al., 2025). Figure 12b illustrates the scaling behavior across S and B student configurations, evaluated

at 1 and 4 steps. The pattern closely mirrors velocity KD: an oversized teacher never collapses the student, and the marginal gains from scaling the teacher quickly shrink as the teacher grows.

Discussion. Both KD and MFD exhibit standard scaling behaviors characterized by diminishing returns, contrasting with DMD, where an overly capable teacher can induce training collapse (Section 3.1). We attribute this discrepancy to *the nature of the objectives*. KD strictly matches the teacher's velocity, and MFD learns its integral; both are *trajectory-based* methods regressing onto fixed, deterministic targets. Conversely, DMD is governed by *adversarial* training dynamics, relying entirely on an online critic to track the shifting student distribution (Yin et al., 2024b). Consequently, DMD is acutely sensitive to the capacity of the critic—and the resulting score difference—making joint network scaling critical in a way that trajectory-based methods simply bypass.

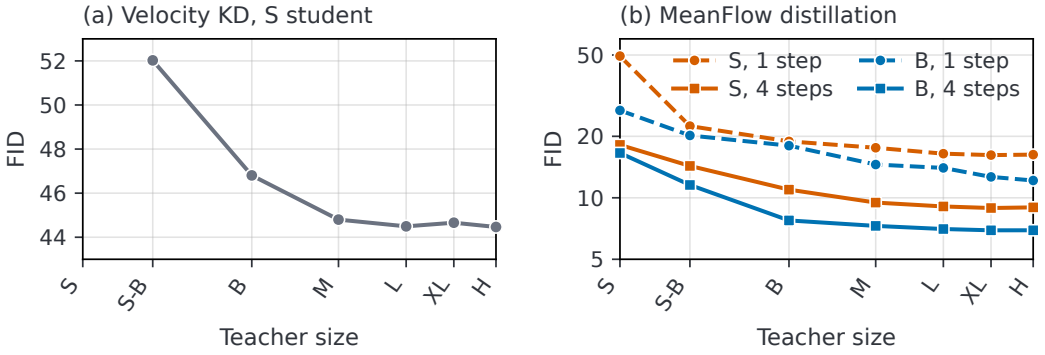


Figure 12: **Teacher scaling under velocity KD and MeanFlow distillation.** Best FID against teacher size for (a) an S student regressed onto each teacher's velocity (the S/S pair was not run) and (b) S and B students distilled with MeanFlow at 1 and 4 steps. No teacher collapses these students, and the gain from a larger teacher shrinks as it grows. (Settings: ImageNet 256×256; KD: learning rate 10^{-4}, 200k updates; MFD: official MeanFlow recipe with default guidance coefficients, 200k updates for S and 60k for B students.)

E Exploring the Critic

E.1 Critic Update Ratio

A common and effective strategy to enhance the critic's ability to track the student distribution is to increase its update frequency, as successfully demonstrated in DMD2 (Yin et al., 2024a) with five critic updates per student update. To investigate how this temporal remedy interacts with network capacity, we apply five critic updates per student update to a student-sized critic and compare it with a larger (XL) critic updated only once per step (Figure 13).

As shown in the figure, increasing the update ratio clearly accelerates early training and helps stabilize mismatched pairs (e.g., avoiding the collapse of a B student under an L teacher). However, while more frequent updates allow a small critic to maximize its potential and reach its performance plateau much faster, they do not fully close the capacity gap. In the later stages of training, the larger XL critic, even with a single update, consistently converges to a lower FID. This suggests that while increasing the update ratio is a beneficial optimization strategy, scaling the critic's capacity provides complementary advantages that are crucial for pushing the limits of final generation quality.

Implementation details. The student EMA (with decay 0.9999) is updated at every training iteration. Consequently, performing five critic updates per student update shortens the effective EMA horizon from $\sim$10k to $\sim$2k student updates, which shifts the moving average dynamics. To ensure a fair comparison, we swept the EMA decay over $\{0.9999, 0.99998, 0.99999\}$. Furthermore, the learning rates of the student and the critic must be carefully re-balanced. Empirically, raising the critic update ratio from 1 to 5 favors a reduced ratio between the student and critic learning rates (from $\sim$8 down to $\sim$1). For the five-update runs in Figure 13, we use $\eta_G = \eta_F = 2 \times 10^{-6}$, which aligns with the default setting of DMD2 (Yin et al., 2024a), and the one-update runs use the learning rates of the corresponding grid cells.

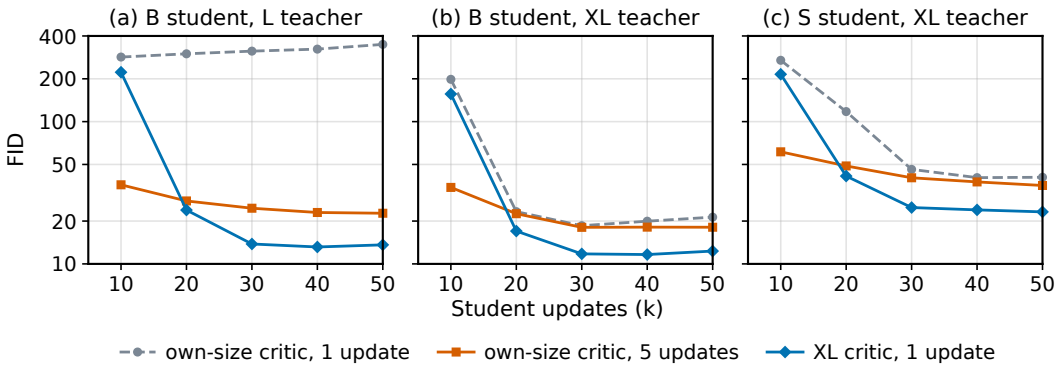


Figure 13: **More critic updates for a small critic.** FID during training with the student-size critic at one and five updates per student update, and with an XL critic at one update, for (a) a B student with an L teacher, (b) a B student with an XL teacher, and (c) an S student with an XL teacher. Five updates speed up early training and avoid the collapse in (a), though the student-size critic stays at a higher FID than the single-update XL critic. (Settings: ImageNet 256×256, 1-step students, w/o CFG.)

E.2 Critic Warm-up

Another intuitive hypothesis for why a small critic collapses under a large teacher (Figure 5b) is a poor initial fit. Because the critic is initialized from a pretrained checkpoint rather than tailored to the student, early gradient inaccuracies might push the student irreversibly off course. If true, fitting the critic *before* updating the student should prevent the collapse.

To test this, we freeze the student at initialization and "warm up" the critic alone for 10k or 40k updates before resuming the standard alternating recipe (Figure 14b tracks the validation loss of the fitting). We apply this intervention to the previously unstable configuration of a 1-step S student, an H teacher, and an S critic. As shown in Figure 14a, neither warm-up duration stabilizes the training.

Implementation details. The four runs in Figure 14a share the setting of the S-student grid: 1-step S student, H teacher, $(\eta_G, \eta_F) = (1.2 \times 10^{-5}, 5 \times 10^{-7})$, batch size 64, one critic update per student update, w/o CFG. The S critic starts from its pretrained checkpoint and trains for 40k updates at a learning rate of 10^{-6}. The validation loss is measured as the denoising loss on a fixed set of the frozen student's outputs at eight score-evaluation levels.

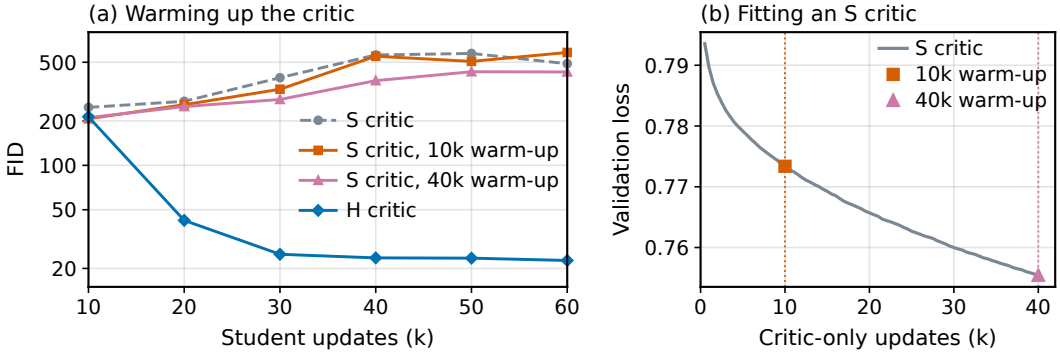


Figure 14: **Critic warm-up.** (a) FID of a 1-step S student under an H teacher with an S critic. The S critic is either trained concurrently from the start or warmed up for 10k/40k updates before student training begins. Both warm-ups fail to prevent collapse compared to the matched H critic. (b) Validation loss of an S critic fitted to a frozen student. (Settings: ImageNet 256×256, w/o CFG.)

E.3 Critic Initialization

In our primary experiments, a matched critic always starts as an *exact copy* of the teacher. Consequently, it inherits the teacher's exact initial weights, its level of pretraining convergence, and its intrinsic architectural capacity. To isolate which of these factors is responsible for avoiding the degradation of mismatched pairs, we systematically decouple them.

Do the teacher and critic need to share exact weights? We first hypothesize that the critic might rely on sharing the teacher's exact starting weights to effectively track the target distribution. To test this, we distill a 1-step S student using two *independently pretrained* H checkpoints (Model 1 and Model 2). We cross-test them as teachers and critics, removing the shared-weight advantage:

	Critic	
Teacher	Model 1	Model 2
Model 1	21.41	21.14
Model 2	21.47	20.97

As shown above (with shaded cells indicating shared initial weights), the off-diagonal pairs with independent weights perform practically identically to the diagonal pairs. This confirms that the critic does *not* need to share the teacher's specific initialization to provide comparable performance.

Does extended pretraining act like larger capacity? If intrinsic capacity is the key, how does pretraining duration factor in? One might suspect that a teacher trained for an extraordinarily long time, reaching near-perfect convergence, would act equivalently to a "larger" model, thereby crushing a matched-size critic. To investigate this, we use the official SiT-XL model heavily pretrained for 7M updates (Ma et al., 2024) (far longer than our 800k updates, with FID 9.35 against 14.43 for ours) and cross-test it against our standard B and XL checkpoints (Table 16). The results strictly isolate capacity from convergence: under the converged 7M XL teacher, our 800k XL critic successfully stabilizes the student and achieves highly competitive performance. Conversely, a B critic degrades under the 7M XL teacher (FID 35.96).

These results tell a cohesive story: a critic's ability to track the student without degrading is largely *not determined by shared initial weights or the teacher's pretraining duration*. It is instead greatly influenced by the critic's *intrinsic architectural capacity*.

	Critic		
Teacher	B	XL	XL, 7M
B	17.94	16.61	17.64
XL	18.56	13.71	11.49
XL, 7M	35.96	12.01	11.90

Table 16: **Teachers and critics from different checkpoints.** Best FID of a 1-step B student. Teachers (rows) and critics (columns) come from our B and XL models pretrained for 800k updates, alongside the highly converged XL model from Ma et al. (2024) trained for 7M updates (XL, 7M). Shaded cells denote shared initial weights. Regardless of the teacher's training duration, matching the critic's architectural size (XL) gives stable and strong performance, whereas a B critic degrades under both XL teachers. (Settings: ImageNet 256×256, w/o CFG, default learning rates, 100k updates.)

E.4 Low-rank Adaptation on Critics

When an undersized critic falls behind a large teacher, it suffers from two separate deficits: a smaller architectural backbone and fewer trainable parameters. This raises a question: is this degradation driven by insufficient *trainable parameters* to track the moving student, or by the limited *expressive capacity* of the smaller backbone?

To decouple parameter count from architectural capacity, we build the critic directly on the frozen XL teacher and train only Low-Rank Adapters (LoRA) (Hu et al., 2021). We inject adapters into the attention projections, MLPs, and AdaLN layers across all blocks, paired with rank-stabilized scaling (Kalajdzievski, 2023). Initializing the adapters to zero ensures the critic starts identically to the teacher, while the rank explicitly constrains how far it can deviate. Crucially, an $r = 1$ critic introduces less than 1% of the trainable parameters of a full B-sized critic, yet it operates entirely within the high-capacity representation space of the XL backbone.

As shown in Table 17, evaluating these configurations on a 1-step B student under an XL teacher (the setting where a full B critic degrades in Figure 5a) yields two findings: First, *backbone capacity*

matters more than the number of trainable parameters. Every LoRA variant avoids this degradation and outperforms the full B critic, and even at $r = 1$, with few trainable parameters, the adapter-based critic remains stable. This suggests that the degradation of small critics stems mainly from their smaller backbones. Second, *LoRA provides a practical, memory-efficient surrogate for joint scaling*. While a fully fine-tuned XL critic still achieves the lowest FID, low-rank adaptation recovers about 70% of the joint-scaling gain at a fraction of the optimizer memory. Adapting the frozen teacher thus offers an attractive, parameter-efficient recipe for scaling critics under constrained academic budgets.

Critic	Trainable params (M)	Critic learning rate 3×10^{-5}	10^{-4}	3×10^{-4}
LoRA on XL, $r = 1$	0.74	14.61	13.78	14.99
LoRA on XL, $r = 4$	2.97	<u>13.66</u>	14.97	16.34
LoRA on XL, $r = 16$	11.87	14.82	16.92	17.28
LoRA on XL, $r = 64$	47.48	13.84	17.68	17.75
Full B (student size)	130.5		18.56	
Full XL (matched)	675.1		**11.63**	
Full H	951.3		12.23	

Table 17: **Low-rank critics on the frozen teacher.** Best FID of a 1-step B student under an XL teacher. LoRA critics adapt the frozen XL teacher, and the full-parameter critics are the corresponding cells of Figure 10b, with the learning rates selected per cell. The best LoRA result is underlined. Even at rank 1, every LoRA critic beats the full B critic, though none reaches the full XL critic. (Settings: ImageNet 256×256, w/o CFG; LoRA runs: $\eta_G = 1.2 \times 10^{-5}$, 100k updates, batch size 64, scaling $\alpha/\sqrt{r}$ with $\alpha = 8$, zero-initialized up-projection.)

F More Results on STAC

F.1 Pooled and Step-wise Objectives

In standard multi-step DMD, the critic estimates the score of a mixture over student generation steps. STAC instead conditions the critic on the generation step. To motivate this design, we compare idealized pooled and step-wise KL objectives. STAC turns the resulting pooled objective into a *step-wise* objective by letting the critic estimate the score of each step's distribution.

For a fixed condition c and score-evaluation level t, let $q_k = q_t^k$ be the distribution of step k's re-noised prediction, $p = p_t$ the target, and $\bar{q} = \sum_k \pi_k q_k$ their pool, where π_k are the fixed sampling probabilities. Assuming densities exist, the step-wise and pooled KL divergences are related by

$$\begin{aligned}\sum_k \pi_k \,\mathrm{KL}(q_k \| p) &= \sum_k \pi_k \int q_k(x) \log \frac{q_k(x)}{\bar{q}(x)}\, dx + \sum_k \pi_k \int q_k(x) \log \frac{\bar{q}(x)}{p(x)}\, dx \\ &= I(K; X_t) + \mathrm{KL}(\bar{q} \| p), \end{aligned} \tag{8}$$

where $K \sim \pi$ and $X_t \mid K = k \sim q_k$. The step-wise objective (left side) matches every step to the target individually. It differs from the pooled objective (last term) by the mutual information $I(K; X_t)$ between the step index and the re-noised prediction. STAC prevents the critic from minimizing the loss via "a good pool of bad steps" by explicitly penalizing this mutual information gap.

F.2 Step Information and Optimization Stability

Equation 8 implies that the step-wise and pooled objectives differ only when the predictions from different steps are distinguishable, that is, when $I(K; X_t) > 0$. To measure this, we trained a classifier to predict the step index k from the re-noised prediction x_t^k of two 4-step B students, one without and one with STAC.

As shown in Figure 15, the classifier identifies the step well above chance at low and moderate score-evaluation levels, suggesting a clear distribution shift across generation steps. Moreover, STAC lowers this step information at nearly every level. This is consistent with Equation 8, since minimizing the step-wise objective also minimizes $I(K; X_t)$ and thus pushes each step's distribution toward the shared target.

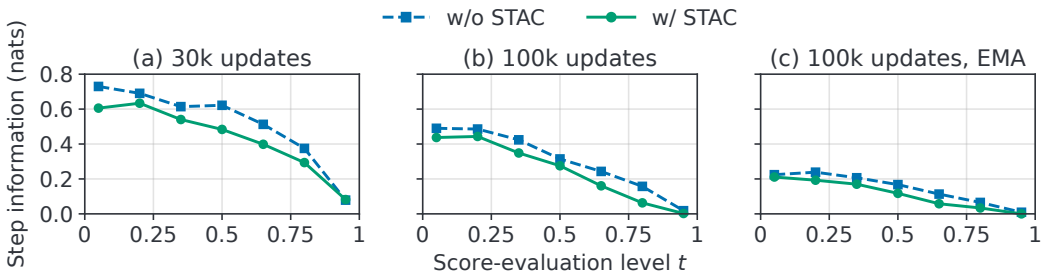


Figure 15: **Step information in re-noised predictions.** Lower bound on the mutual information $I(K; X_t)$ from a classifier predicting the step index of a 4-step B student, evaluated at (a) 30k and (b) 100k updates with online weights, and (c) at 100k updates with EMA weights. The step is readable at low and moderate score-evaluation levels, and STAC lowers this information at nearly every level. (Settings: ImageNet 256×256, XL teacher and critic, w/o CFG; 10k fixed noise–class pairs, 8k/2k training/held-out split.)

Loss variability. A single step-agnostic critic has to fit a mixture of step distributions whose composition shifts as the student changes. STAC gives the critic the step index, so it can model each step's distribution separately. Consistent with this, STAC lowers the coefficient of variation (CV) of both the critic loss and the distribution matching (DM) loss at every evaluation after the first (Figure 16). Together with its FID gains in the same runs (Figure 7), this suggests that a step-aware critic also makes training smoother.

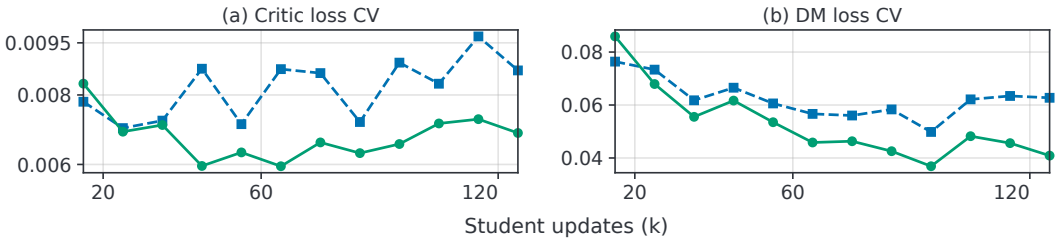


Figure 16: **Loss variability with and without STAC.** Coefficients of variation of (a) the critic loss and (b) the DM loss during the training of a 4-step student. STAC (solid) has a lower CV than the step-agnostic baseline (dashed) at every evaluation after the first. (Settings: XL student, teacher, and critic, ImageNet 256×256, CFG $= 1.3$.)

F.3 Ablations on Step Conditioning for SDXL

Our ImageNet experiments (Table 2) suggest that the way the step enters the critic matters, since in-context tokens outperform an AdaLN branch (Peebles & Xie, 2023). In this section, we further compare step conditioning methods on the SDXL UNet backbone.

The SDXL UNet reads text prompts via cross-attention (Rombach et al., 2022). We therefore compare two ways to inject the step index: adaptive group normalization (AdaGN) (Dhariwal & Nichol, 2021), the UNet counterpart of AdaLN, and learned in-context step tokens appended to the cross-attention text sequence. The detailed hyperparameters are listed in Appendix B.3.

The results in Table 18 mirror the ImageNet findings. Four in-context step tokens give the best FID, whereas AdaGN improves only slightly over the step-agnostic baseline. Furthermore, 8 and 16 tokens give smaller gains than four.

Conditioning	Channel	Tokens	FID↓	CLIP-S↑	IS↑
None	None	0	34.71	32.65	34.78
Step index	AdaGN	0	34.41	32.65	36.25
Step tokens	Cross-attention	4	**32.73**	32.64	35.29
Step tokens	Cross-attention	8	33.72	32.59	37.15
Step tokens	Cross-attention	16	34.01	32.60	36.51

Table 18: **Step conditioning of the SDXL critic.** FID, CLIP-S ($\times 100$), and IS for each conditioning channel and token count. Four cross-attention tokens give the best FID, ahead of AdaGN and of 8 or 16 tokens. (Settings: SDXL student and critic, 4 steps, 8k updates, 10k COCO prompts; pure DMD without GAN loss.)

Token construction. Across backbones, the N step tokens are built from one shared template. The scalar step index (either k or the associated noise level r_k) is first mapped to a sinusoidal frequency embedding of dimension 256, using standard sin/cos features (Vaswani et al., 2017). This 256-dim vector is passed through a small two-layer projection MLP, $\text{Linear}(256, D) \to \text{SiLU} \to \text{Linear}(D, D)$, where D is the backbone's context-conditioning embedding dimension ($D = 2048$ for the SDXL UNet's cross-attention text dimension; D is the DiT hidden dimension for Wan2.1). Both linear layers are initialized with $\mathcal{N}(0, 1/\text{fan_in})$ weights and zero bias. Separately, N learnable *type tokens* of dimension D are initialized as $\mathcal{N}(0, 1) \cdot D^{-1/2}$ and are otherwise independent of the step index. The final step tokens are formed by broadcasting the scalar's MLP embedding across all N type tokens and adding it: token n is $\mathbf{t}_n = \mathbf{e}_n + \text{MLP}(\text{sinu}(r_k))$, where $\mathbf{e}_n$ is the n-th type token. The N tokens are therefore distinct from one another (through $\mathbf{e}_n$) but share the same step-dependent additive term, rather than each encoding an independent projection of r_k. These tokens are appended to the end of the backbone's text/context conditioning sequence (for classifier-free guidance, to both the conditional and unconditional branch); no positional embedding is added to them, since none of the backbones apply positional encoding to this conditioning sequence. N is 4 for the SiT (ImageNet), SDXL, and FLUX.2 critics and 32 for the Wan2.1 critic (Appendix B.3, B.4, B.5).

Learning rates. On every backbone, the step tokens' projection MLP and type-token parameters are trained at a substantially higher rate than the critic backbone itself, so the newly initialized conditioning parameters converge faster than the pretrained critic: $20\times$ the critic rate on ImageNet (Appendix B.1); an explicit absolute rate of 10^{-4} against a critic rate of 5×10^{-7} on SDXL, i.e., $200\times$ (Appendix B.3); and 5×10^{-5} against the same 5×10^{-7} critic rate on FLUX.2, i.e., $100\times$ (Appendix B.4). On Wan2.1, the token parameters instead follow a warm absolute schedule, starting at 2×10^{-4} and cosine-decaying to the critic's own rate over the early training steps (Appendix B.5), so that the newly initialized tokens are not slowed down by the much lower rate needed by the pretrained critic.

F.4 Cost Analysis

Zero inference overhead. Because STAC modifies only the critic, which is discarded after training, the deployed student keeps its architecture and sampling procedure. STAC therefore adds *no computational cost at inference*.

Negligible training overhead. During training, the step tokens and their small projection MLP add little overhead. In the SDXL and Wan2.1 experiments (Sections 4.1 and 4.2), STAC adds less than 0.2% to the critic's parameters and at most 0.7% to peak memory (Table 19), with no measurable change in iteration time.

Separate critics per step. A direct alternative to STAC is to train K separate step-specific critics. However, this multiplies the critic's parameters and optimizer state by K. Assuming roughly 16 bytes per parameter for weights, gradients, and optimizer moments, four step-specific SDXL critics would need about 166 GB of training memory, compared with 42 GB for a single critic with STAC. For the 14B Wan2.1 critic, the requirement grows to 0.9 TB against 224 GB, which makes this design far more costly to train.

System	Added parameters	Iteration time	Peak memory
SDXL	$+0.18\%$	$\approx 0\%$	$+0.70\%$
Wan2.1, autoregressive	$+0.19\%$	$\approx 0\%$	$+0.15\%$

Table 19: **Cost analysis of STAC on SDXL and Wan2.1.** Changes relative to the w/o STAC baseline of each system, with the inference architecture unchanged. STAC adds a negligible amount of memory and iteration time.

G More Visualizations

G.1 Qualitative Results on ImageNet

Figure 17 shows samples of the four 1-step B students of Figure 5a and Table 1 at 200k updates, next to the B teacher, whose checkpoint also initializes every student. Compared with the B/B pair, scaling both the teacher and the critic to XL gives more coherent objects, whereas scaling only the critic leaves the sample quality largely unchanged. In contrast, the XL/B student produces nearly a single image per class, sometimes with distorted colors or textures.

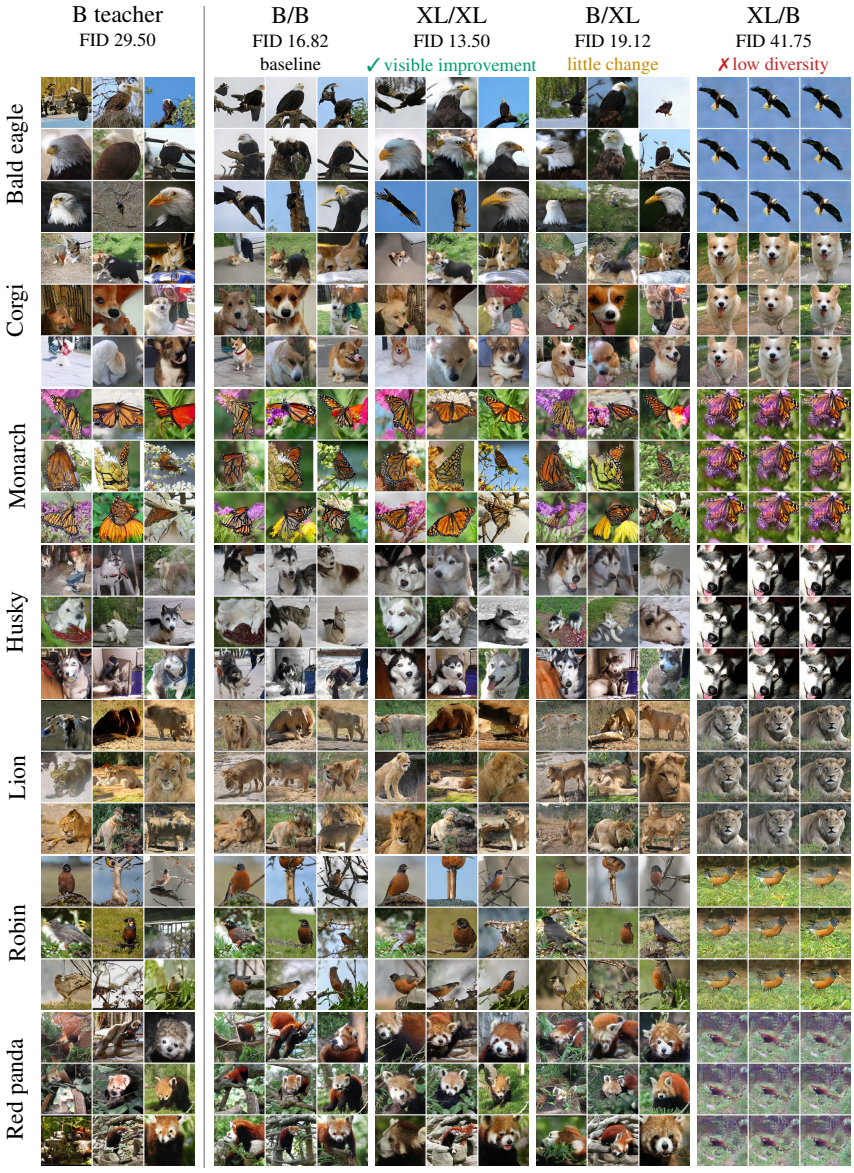


Figure 17: ***Uncurated* ImageNet 256×256 samples of 1-step B students.** Each 3 × 3 grid shows nine samples of one selected class. The first column is the B teacher without guidance, with its FID from Table 6, and the other columns are the students of Figure 5a at 200k updates, labelled by their teacher/critic pair, their FID (Table 1), and the change from B/B. Scaling both networks (XL/XL) gives more coherent objects than B/B, and scaling only the critic (B/XL) changes little, whereas scaling only the teacher (XL/B) leaves nearly one image per class. (Settings: w/o CFG.)

G.2 Qualitative Results on FLUX.2

Figure 18 provides visual comparisons of 4-step FLUX.2 4B students across different distillation configurations. We note that, because DMD matches the teacher's output distribution without following its sampling trajectory, the layouts of the student samples differ from those of the teacher samples.

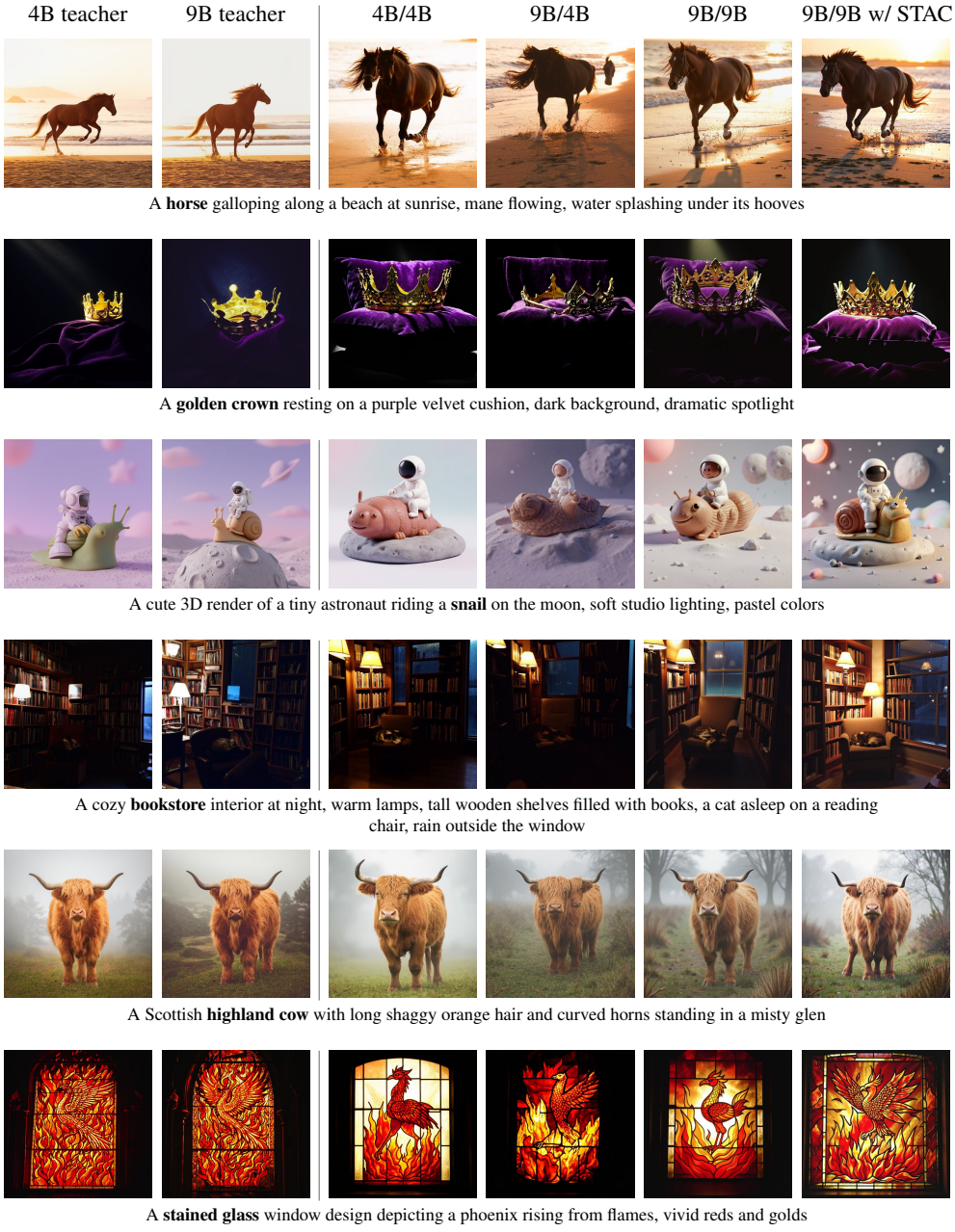


Figure 18: **FLUX.2 samples of 4-step 4B students across teacher/critic pairs.** Each row shares a prompt (listed below) and initial noise. The first two columns display the 50-step 4B and 9B teachers for reference. The subsequent columns show 4-step students distilled under the 4B/4B, 9B/4B, and 9B/9B teacher/critic pairs, alongside the 9B/9B pair combined with STAC. (Settings: students initialized from FLUX.2 [klein] base 4B, 4 steps; teachers sampled with 50 Euler steps at CFG 2.0; resolution 512×512.)

G.3 Qualitative Results on SDXL

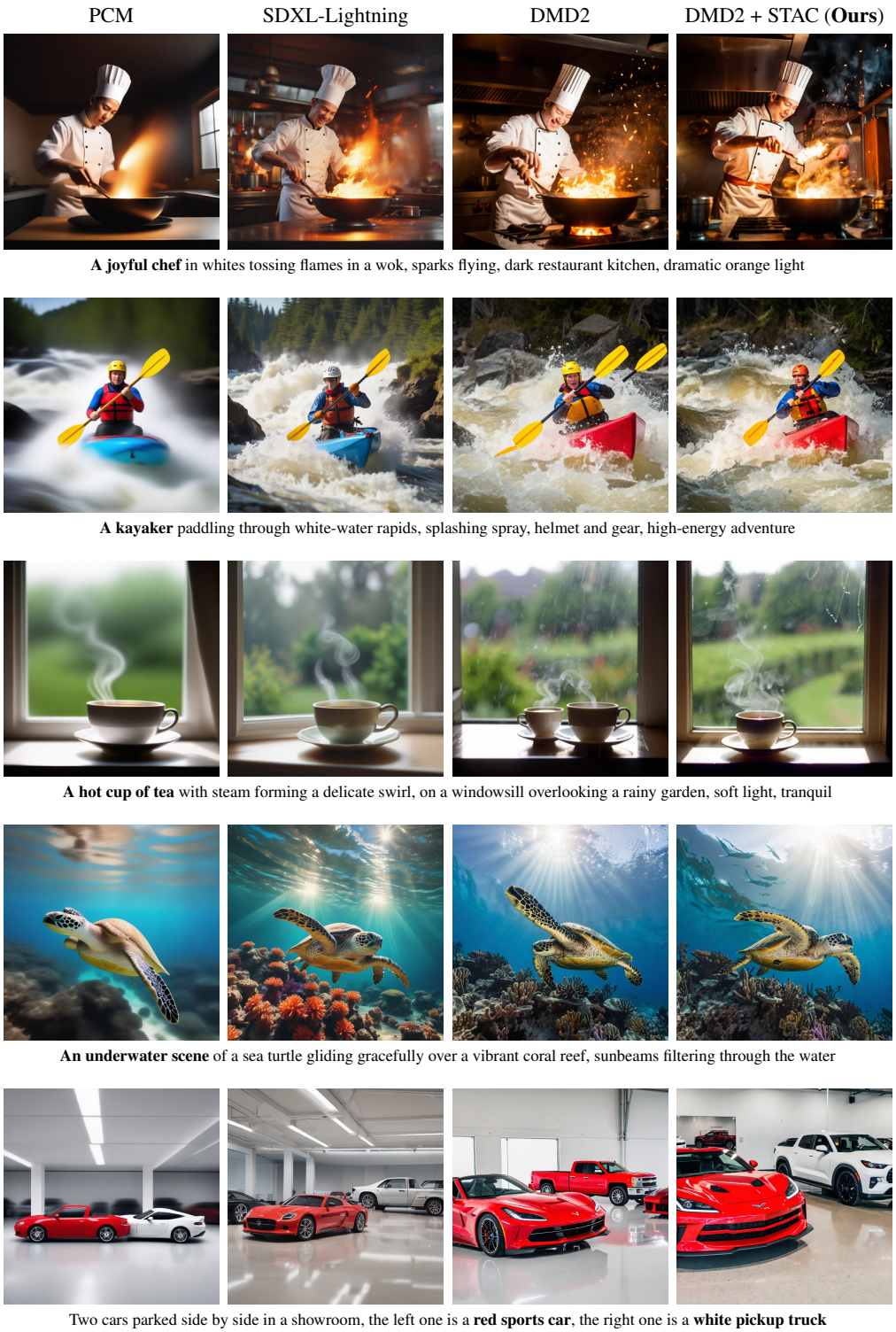


Figure 19: **SDXL samples of 4-step students across distillation methods.** Each row shares a prompt and seed across PCM, SDXL-Lightning, DMD2, and its STAC continuation, with the prompt given below it. Compared with the other methods in this figure, DMD2 + STAC gives sharp images with coherent objects and faithful attributes. (Settings: 4 steps, 1024×1024, shared seed per prompt.)

G.4 Qualitative Results on Wan2.1

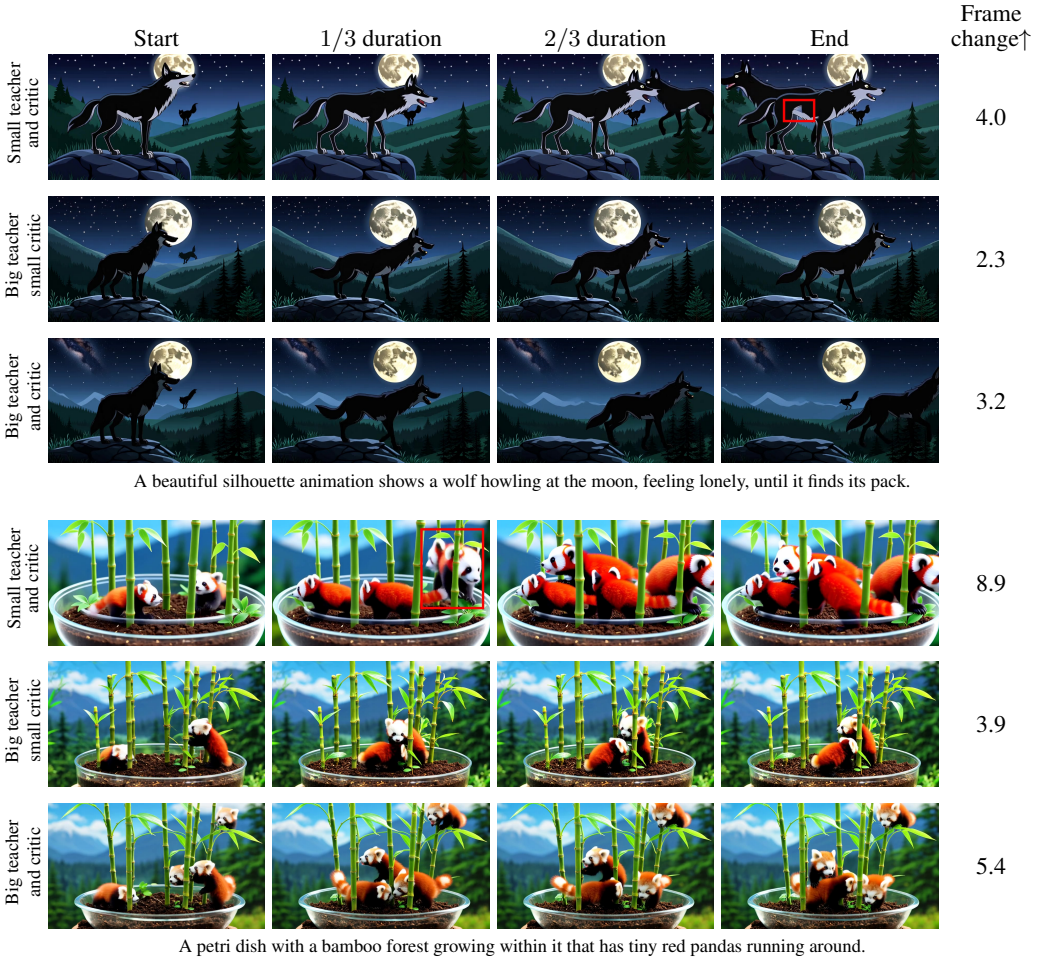


Figure 20: **Wan2.1 samples of 4-step 1.3B students across teacher/critic pairs.** Each row shows four frames of one video, and the three rows of a block share the prompt given below them. *Big* and *small* denote the 14B and 1.3B models. Frame change is the mean absolute luminance difference between consecutive shown frames, as a percentage of the intensity range, and higher values indicate more motion. Red boxes mark artifacts, and the boxed frame in the second block is taken at 0.94 s (19% of the duration). With the big teacher, the small critic reduces the motion of the student, and scaling the critic with the teacher recovers part of this motion and gives more plausible videos. (Settings: autoregressive Wan2.1-T2V-1.3B students with the Self Forcing rollout, 4 steps, 480p; teacher/critic pairs of Table 5.)