DMAD: DISTRIBUTION MATCHING AS ADVERSARIAL DISTILLATION FOR FAST VISUAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Distribution Matching Distillation (DMD) trains a few-step student from the difference between separately estimated target and student scores, so it must keep an auxiliary diffusion model fitted to the student's evolving distribution at extra memory and computation cost. We introduce DMAD, **D**istribution **M**atching as **A**dversarial **D**istillation, which recasts distribution matching as classification and learns the required log-density ratios directly. Two discriminator heads on a shared backbone distinguish real data and teacher samples from the student's, and linear losses on their logits train the student without auxiliary score fitting. We prove that at the discriminator optimum these losses recover the distribution-matching gradient underlying DMD, through the classical identity linking discriminator logits to log-density ratios. We further introduce gap-based reweighting, which adapts teacher supervision across noise levels from the real-data head's empirical logit gap between real and teacher samples. DMAD reaches a Fréchet Inception Distance (FID) of 1.04 with one-step generation on ImageNet-64×64, 14.47 with four-step SDXL on COCO-10K, and a VBench total score of 85.15 with four-step Wan2.1-T2V-14B, the best values among the compared few-step methods and the multi-step teachers. On MiniMax-H3-33B, our four-step student achieves overall human preference rates of 79.1% over DMD2 and 84.6% over rCM for joint audio-video generation, excluding ties.

1 INTRODUCTION

Diffusion models generate high-quality images and videos (Ho et al., 2020; Song et al., 2020; Podell et al., 2023; Wan et al., 2025), but sampling requires many sequential denoising steps. Step distillation reduces this cost by training a few-step generator from a pretrained teacher. Existing approaches supervise the student along the teacher's sampling trajectories (Salimans & Ho, 2022; Song et al., 2023; Kim et al., 2024) or match distributions (Yin et al., 2024b;a). Distribution matching lets the student learn its own mapping from noise to samples while preserving the target distribution.

Distribution Matching Distillation (DMD) (Yin et al., 2024b) trains the student on the difference between target and student scores. A frozen teacher supplies the target score, while an auxiliary diffusion model, the critic, tracks the evolving student distribution. When the critic lags behind, its score estimate carries error and training destabilizes. DMD2 (Yin et al., 2024a) mitigates this with a two time-scale update rule (TTUR) that runs five critic updates per student update. Stability improves, but the student score is still only approximated, and each student update costs five critic passes. Beyond the critic, DMD2 evaluates the teacher online during training, adding further memory and computation cost. Its denoising task also constrains the architecture of the discriminator.

We introduce DMAD, Distribution Matching as Adversarial Distillation, which replaces separate score estimation with two classification tasks, one against teacher samples and one against real data. Two discriminator heads distinguish these targets from student samples, and linear losses on their logits drive the updates. At the discriminator optimum, this loss yields the distribution-matching gradient underlying DMD (Proposition 1). DMAD eliminates auxiliary score fitting. In our experiments, we use one discriminator update per student update. Freed from the denoising task, the discriminator can build on diffusion encoders or pretrained classification features. The teacher now only supplies samples, so we generate them once and cache them, removing online teacher evaluation from training.

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.

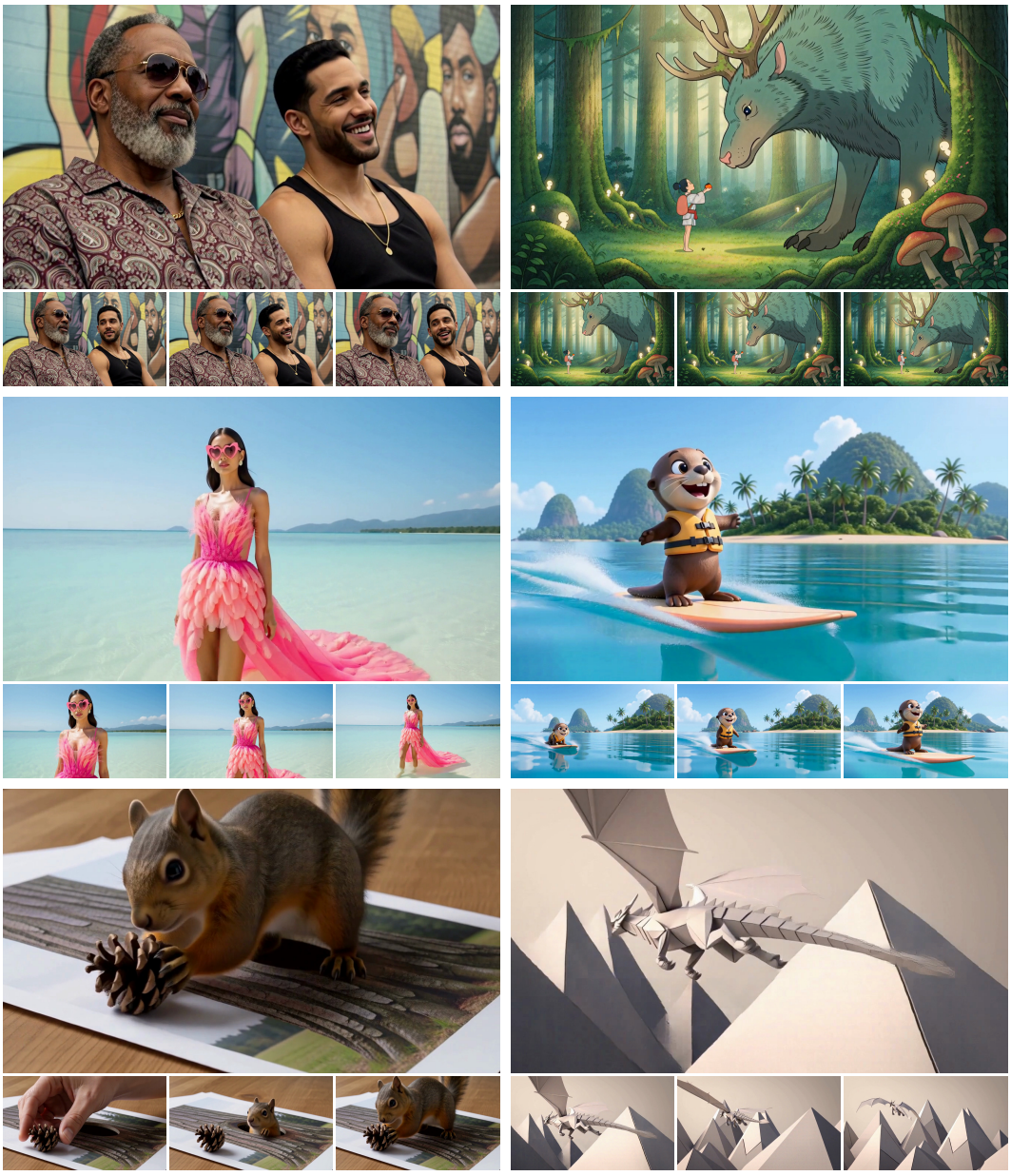


Figure 1: **Samples of our four-step student distilled from MiniMax-H3.** Zoom in for details.

Errors in separately estimated scores can be large relative to their difference, which drives the student update. We examine this in a controlled experiment on two nearby Gaussian mixtures with known gradients, holding the student distribution fixed and giving both estimators the same teacher pretraining and seed. Figure 2 compares DMD's score difference with our single-head logit gradient. In this setting, our estimate follows the ground-truth field more closely, with normalized mean squared error of 0.144, compared with 1.539 for DMD. Section B.2 gives the setup and the analytic gradients.

DMAD draws on two targets, the teacher's samples and real data, which play complementary roles. Previous methods, including DMD2 (Yin et al., 2024a), SiDA (Zhou et al., 2024a), and SenseFlow (Ge et al., 2025), add adversarial supervision on top of a score-based objective. In contrast, our discriminator heads carry the distribution-matching gradient themselves. Each head compares student samples with one target, and a linear loss on their logits trains the generator. The real-data head supplies supervision beyond the teacher's sampled distribution. Teacher samples alone would bound the student by the teacher's own fidelity; training against real data lifts that bound, and our students outperform their teachers on both ImageNet and COCO. Our ablations in Table 4 demonstrate the complementary roles of the two targets: removing the real-data head raises four-step FID from

14.47 to 21.37 on COCO-10K, while removing teacher supervision reduces semantic fidelity and visual quality, dropping VBench total score from 84.70 to 82.20 and semantic score from 79.55 to 75.36.

The balance between the two targets can vary across noise levels. During training, the real-data head assigns different mean logits to real and teacher samples (Figure 7(a)). This gap varies across noise levels and depends on the current student and discriminator. We use it as an empirical signal for *gap-based reweighting*, assigning larger teacher weights to noise bands with lower gaps. This adaptive allocation improves performance for both image and video generation (Table 4).

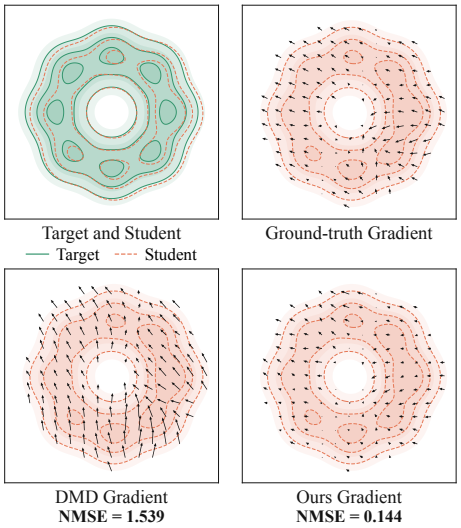


Figure 2: **Gradient estimation on two Gaussian mixtures with a fixed student.** Our logit gradient more closely matches the ground-truth field than DMD's score difference.

Our contributions are:

- We introduce DMAD, a distribution-matching distillation framework that jointly learns from teacher-generated samples and real data through two binary classification tasks, eliminating auxiliary score fitting and online teacher evaluation.
- We develop gap-based reweighting, which uses the real-data head's mean responses to teacher and real samples within each noise band to adapt the teacher's supervision as training proceeds.
- DMAD achieves 1.04 FID with one-step generation on ImageNet-64×64 and 14.47 with four-step SDXL on COCO-10K, where one step reaches 16.11. Four-step Wan2.1-T2V-1.3B and -14B reach VBench total scores of 84.70 and 85.15. These are the best values among the compared few-step methods and the multi-step teachers. On the 33B MiniMax-H3 (MiniMax, 2026), human raters prefer our four-step student to DMD2 in 79.1% and to rCM (Zheng et al., 2026) in 84.6%, excluding ties.

2 Related Work

Trajectory and consistency methods. Early work distills a teacher either by regressing to its outputs (Luhman & Luhman, 2021; Meng et al., 2023) or by progressively reducing the number of sampling steps (Salimans & Ho, 2022). PDD (Shaul et al., 2026) extends this idea to video using multi-interval mean velocities. Consistency and flow-map methods impose consistency along a common trajectory, either by mapping different states to the same endpoint or by relating states at different points on that trajectory (Song et al., 2023; Song & Dhariwal, 2024; Kim et al., 2024). sCM (Lu & Song, 2025) extends consistency modeling to continuous time, while MeanFlow (Geng et al., 2025) enables one-step generation from scratch. These methods all constrain how the student follows a particular trajectory. Distribution matching instead constrains the distribution of its outputs.

Score-based distribution matching. Diff-Instruct (Luo et al., 2023b), VSD (Wang et al., 2023), and DMD (Yin et al., 2024b) optimize a generator through score-based distribution matching. DMD uses the difference between a target score and a score fitted to the student's evolving distribution. DMD2 (Yin et al., 2024a) removes the regression loss and strengthens online critic training, while SenseFlow (Ge et al., 2025) improves stability for large flow-based models. SiD (Zhou et al., 2024b) and SGMD (Wu et al., 2026b) derive alternative objectives from score identities and Fisher-based matching. DP-DMD (Wu et al., 2026a) and PhasedDMD (Fan et al., 2026) distribute supervision across generation steps or noise intervals. rCM (Zheng et al., 2026) combines score regularization with consistency training, AnyFlow (Gu et al., 2026) learns flow-map transitions for any-step video generation, and AMD (Bai et al., 2026) adapts distribution matching using reward proxies. These methods still rely on score estimation. FSF-DMD (Kim et al., 2026) removes the auxiliary fake-score network but derives a fake-score surrogate from the generator's own flow map. Our method instead estimates no score and learns the required log-density ratio directly through classification.

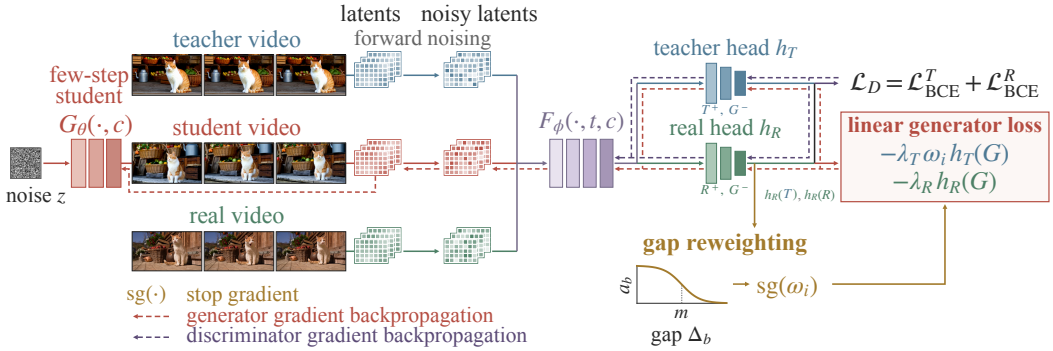


Figure 3: **Overview of DMAD (drawn for video distillation).** T, R, and G denote noisy teacher, real, and student samples, with time and text conditions omitted. A shared backbone F_ϕ feeds a teacher head and a real head, and a linear loss on each logit trains the generator. The real-data head also provides the gap for teacher reweighting. Video frames are decoded only for illustration.

Adversarial distillation. Adversarial distillation typically trains discriminators against real images or videos (Sauer et al., 2023; Lin et al., 2025a;b), or uses intermediate teacher predictions as adversarial targets (Sauer et al., 2024; Lin et al., 2024). D2O-F (Zheng & Yang, 2025) finetunes a pretrained diffusion model with a projected discriminator. DMD2, SiDA (Zhou et al., 2024a), SenseFlow, and ADM (Lu et al., 2025) add adversarial supervision to a score-based distribution-matching objective. DDO (Zheng et al., 2025) connects likelihood ratios to an implicit discriminator, while f-distill (Xu et al., 2025) uses density ratios to reweight score differences. Unlike these methods, we do not use adversarial training to supplement a score-based objective. The discriminator logits themselves provide the distribution-matching gradient, eliminating student-score estimation. Separate heads match teacher samples and real data, with their balance adapted across noise levels.

3 Method

We reformulate DMD's score-estimation pipeline as joint classification-based matching to teacher samples and real data. Linear losses on two discriminator logits train the few-step student. The same real-data head also scores teacher samples, and the gap it leaves between them and real data adapts teacher supervision across noise levels. Figure 3 illustrates the pipeline for video distillation. Image distillation is the same, with only the discriminator heads adapted to the backbone (Section B).

3.1 Reformulating Distribution Matching Distillation

We start from the idealized distribution-matching update underlying DMD (Yin et al., 2024b). Let G_θ be a student with parameters θ and the final rollout output $\boldsymbol{x}_0 = G_\theta(\boldsymbol{z}, \boldsymbol{c})$ for initial noise $\boldsymbol{z}$ and condition $\boldsymbol{c}$, inducing a distribution p_θ. At noise level t, we form $\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ is independent Gaussian noise. The forward schedule specifies signal scale α_t and noise standard deviation σ_t. Let p_t be the noisy marginal of a clean distribution p and $\boldsymbol{s}_p = \nabla_{\boldsymbol{x}_t} \log p_t$ its score. All densities are conditional on $\boldsymbol{c}$, omitted below. As in DMD, the analysis assumes the same fixed distribution over $t, \boldsymbol{c}$ for target and student samples, and all expectations below use it.

For a fixed target distribution p, the update above is motivated by the reverse-KL objective

$$\mathcal{L}_{\text{DM}}(\theta; p) = \mathbb{E}_{t,\boldsymbol{c}} \left[D_{\text{KL}}(p_{\theta,t} \,\|\, p_t) \right]. \tag{1}$$

Its generator gradient depends on the score difference, with $\boldsymbol{s}_\theta = \nabla_{\boldsymbol{x}_t} \log p_{\theta,t}$ the student score:

$$\nabla_\theta \mathcal{L}_{\text{DM}} = -\mathbb{E}_{\boldsymbol{z},\boldsymbol{\epsilon},t,\boldsymbol{c}} \left[\left(\frac{\partial \boldsymbol{x}_t}{\partial \theta} \right)^{\mathsf{T}} \left(\boldsymbol{s}_p(\boldsymbol{x}_t, t, \boldsymbol{c}) - \boldsymbol{s}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}) \right) \right]. \tag{2}$$

DMD estimates these two scores separately using a frozen teacher and a critic trained on student samples. DMAD instead recovers the score difference directly from a discriminator logit.

To estimate the required score difference directly, we train a single binary classifier to tell target samples from student samples at each noise level. Let h_ϕ be the scalar logit of a discriminator D

with parameters ϕ, and $D_\phi = S(h_\phi)$ its target-class probability, where $S(u) = (1 + e^{-u})^{-1}$ is the logistic sigmoid function. We minimize the balanced binary cross-entropy (BCE) $\mathcal{B}(h_\phi; p)$, which weights target and student samples equally, so its optimum is the log-density ratio alone (Section A).

Proposition 1 (Classification-based distribution-matching update)**.** *At fixed $t, \boldsymbol{c}$, assume positive differentiable densities on common support and the regularity needed to exchange differentiation and integration. The unrestricted, unregularized BCE optimum is $h^\star = \log p_t - \log p_{\theta,t}$, so*

$$\nabla_{\boldsymbol{x}_t} h^\star = \boldsymbol{s}_p - \boldsymbol{s}_\theta. \tag{3}$$

At the current parameters, the generator gradients of $-\mathbb{E}[h^\star]$ and $\mathcal{L}_{\mathrm{DM}}$ are equal with $h^\star$ fixed.

The proof follows from the optimal discriminator identity (Goodfellow et al., 2014) and the chain rule, and is given in Section A. We therefore update the generator with the linear loss

$$\mathcal{L}_G^{\mathrm{lin}}(\theta; \phi) = -\mathbb{E}_{\boldsymbol{z},\boldsymbol{\epsilon},t,\boldsymbol{c}}[h_\phi(\boldsymbol{x}_t, t, \boldsymbol{c})]. \tag{4}$$

With ϕ fixed, gradients through the discriminator's input train the generator without auxiliary score fitting, and the learned discriminator approximates the ideal logit gradient (Figure 2). Unlike the linear loss, the non-saturating loss $-\log D_\phi$ multiplies each sample's logit-gradient contribution by $1 - D_\phi$. We analyze this weighting in Section A and compare the losses in Section 4.3. One classifier therefore replaces both score estimates, and the linear loss carries its gradient unweighted.

3.2 Distillation with Teacher and Real Targets

To provide supervision beyond the teacher's sampled distribution, we train on both teacher-generated samples and real data, with distributions p_T and p_R. A shared backbone F_ϕ feeds two heads with scalar logits h_T and h_R, each trained to distinguish its own target samples from student samples. The combined discriminator objective is the sum of the two balanced classification losses:

$$\mathcal{L}_D = \mathcal{L}_{\mathrm{BCE}}^T + \mathcal{L}_{\mathrm{BCE}}^R, \tag{5}$$

where $\mathcal{L}_{\mathrm{BCE}}^T = \mathcal{B}(h_T; p_T)$ and $\mathcal{L}_{\mathrm{BCE}}^R = \mathcal{B}(h_R; p_R)$. With fixed nonnegative coefficients λ_T and λ_R setting the relative strength of the two targets, the generator loss is the linear combination:

$$\mathcal{L}_G = -\mathbb{E}\left[\lambda_T h_T(\boldsymbol{x}_t, t, \boldsymbol{c}) + \lambda_R h_R(\boldsymbol{x}_t, t, \boldsymbol{c})\right]. \tag{6}$$

At their optima, each head supplies its own target's score difference (Equation (3)), so both targets contribute distribution-matching gradients within one objective. Our ablations in Table 4 show that real supervision improves image FID, while teacher supervision preserves detail and text alignment.

3.3 Gap-Based Reweighting

In Equation (6), λ_T weights the teacher equally at every noise level. The real-data head's mean-logit gap between real and teacher samples varies across noise levels and over training (Figure 7(a)). It depends on the current student and discriminator. We use this empirical signal to assign larger teacher weights to noise bands with lower gaps.

To measure the gap at a usable resolution, we group noise levels into bands and write $b(t)$ for the band containing t. For each band b, we maintain an exponential moving average (EMA):

$$\Delta_b \leftarrow \mathrm{EMA}_\beta\left[\overline{h_R(\boldsymbol{x}_t^R, t, \boldsymbol{c})}_b - \overline{h_R(\boldsymbol{x}_t^T, t, \boldsymbol{c})}_b\right]. \tag{7}$$

The bars average the same real head's logits over noisy real samples $\boldsymbol{x}_t^R$ and teacher samples $\boldsymbol{x}_t^T$ within a band. We use EMA decay $\beta = 0.99$ and update only bands observed in both sources.

The gap then sets a raw teacher weight a_b, centered at the median gap over the measured bands $\mathcal{R}$:

$$m = \mathrm{median}_{b \in \mathcal{R}}\, \Delta_b, \qquad a_b = S\left(\frac{m - \Delta_b}{\tau}\right). \tag{8}$$

A band with a lower mean-logit gap receives a larger weight. The temperature $\tau > 0$ sets how sharply the weight varies around the median, and centering at the median gap keeps the weights relative as the gaps drift over the course of training.

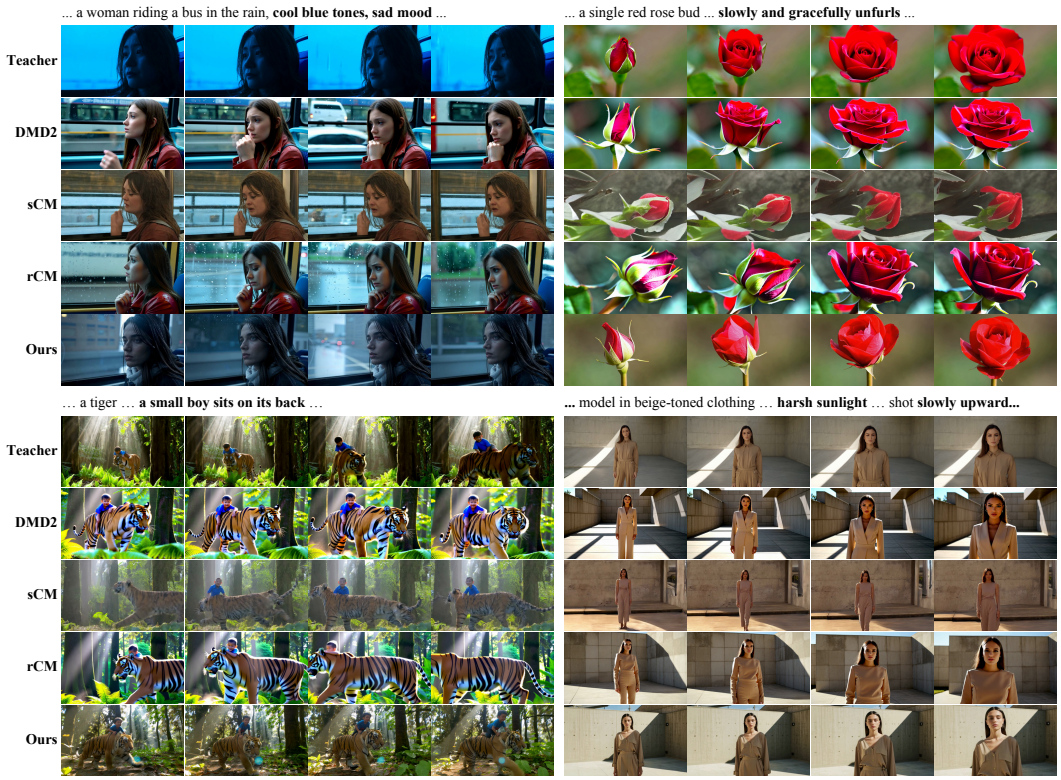


Figure 4: **Qualitative comparisons of our DMAD and baseline models on Wan2.1-T2V-1.3B.** Rows show the teacher, DMD2, sCM, rCM, and ours. Within each example, columns show frames 1, 28, 55, and 81 in order. All compared videos are generated using identical initial noise and text prompts. Prompts are abbreviated for presentation.

We normalize these weights to control their overall scale. For sample i in a minibatch of size B, let $\omega_i = a_{b(t_i)}/C$. For ImageNet and SDXL, $C = \max(B^{-1}\sum_j a_{b(t_j)}, B^{-1})$; for Wan, $C = \max(|\mathcal{R}|^{-1}\sum_{b\in\mathcal{R}} a_b, 10^{-3})$. MiniMax-H3 uses the same band-based normalization with a floor of 10^{-4}. Collecting these terms, the final generator loss is

$$\widehat{\mathcal{L}}_G = -\frac{1}{B}\sum_{i=1}^{B}\left[\lambda_T\omega_i h_T(\boldsymbol{x}_{t_i,i}, t_i, \boldsymbol{c}_i) + \lambda_R h_R(\boldsymbol{x}_{t_i,i}, t_i, \boldsymbol{c}_i)\right]. \tag{9}$$

where $\boldsymbol{x}_{t_i,i}$ is the noisy student output for sample i. We hold the gap statistics and weights fixed during generator backpropagation, so each ω_i scales its term without adding a gradient.

4 Experiments

4.1 Experimental Setup

We evaluate DMAD on class-conditional image generation, text-to-image synthesis, and text-to-video synthesis. Following DMD2 (Yin et al., 2024a), we distill EDM (Karras et al., 2022) for ImageNet-64×64 (Deng et al., 2009) and SDXL (Podell et al., 2023) for zero-shot generation on 10K COCO 2014 prompts (Lin et al., 2014). We use Fréchet Inception Distance (FID) (Heusel et al., 2017) to measure image quality and diversity, and CLIP score (Radford et al., 2021) to evaluate text–image alignment. For SDXL, we additionally report patch FID (Lin et al., 2024; Chai et al., 2022) to assess high-resolution details, following DMD2. For video, we distill Wan2.1-T2V 1.3B and 14B (Wan et al., 2025) to four steps and evaluate them with VBench (Huang et al., 2024) on the same augmented prompt set as rCM (Zheng et al., 2026). We further distill MiniMax-H3 (MiniMax, 2026) to four steps and evaluate them in a user study and AVGen-Bench (Zhou et al., 2026). Training and evaluation details are provided in Section B. More results are shown in our **supplementary demo video**.

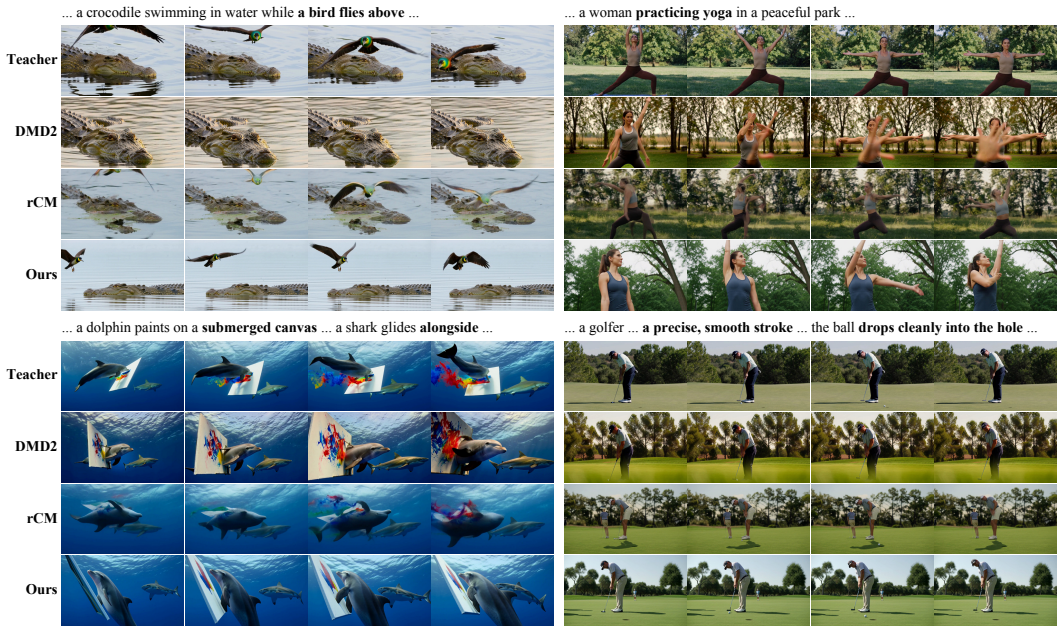


Figure 5: **Qualitative comparisons of our DMAD and baseline models on MiniMax-H3.** Rows show the teacher, DMD2, rCM, and ours, in that order. Within each example, columns show frames 1, 42, 83, and 124 using the same indices for all methods. All compared videos are generated using identical initial noise and text prompts. Prompts are abbreviated for presentation.

Table 1: **Comparisons with the EDM (Karras et al., 2022) on ImageNet-64×64.**

Method	# Fwd Pass (↓)	FID (↓)
BigGAN-deep (Brock et al., 2019)	1	4.06
RIN (Jabri et al., 2023)	1000	1.23
StyleGAN-XL (Sauer et al., 2022)	1	1.52
Consistency Model (Song et al., 2023)	1	6.20
iCT-deep (Song & Dhariwal, 2024)	1	3.25
CTM (Kim et al., 2024)	1	1.92
DMD (Yin et al., 2024b)	1	2.62
D2O-F (Zheng & Yang, 2025)	1	1.16
DMD2 (Yin et al., 2024a)	1	1.51
DMD2+longer training (Yin et al., 2024a)	1	1.28
SiD (Zhou et al., 2024b)	1	1.52
SiDA (Zhou et al., 2024a)	1	1.35
SiD^{2}A (Zhou et al., 2024a)	1	1.11
DMAD (Ours)	1	1.24
+ projected discriminator (Ours)	1	**1.04**
EDM (Teacher, ODE) (Karras et al., 2022)	511	2.32
EDM (Teacher, SDE) (Karras et al., 2022)	511	1.36

Table 2: **Comparisons with the SDXL (Podell et al., 2023) on 10K prompts from COCO 2014.**

Method	# Fwd Pass (↓)	FID (↓)	Patch FID (↓)	CLIP (↑)
LCM-SDXL (Luo et al., 2023a)	1	81.62	154.40	0.275
	4	22.16	33.92	0.317
SDXL-Turbo (Sauer et al., 2023)	1	24.57	23.94	0.337
	4	23.19	23.27	0.334
SDXL Lightning (Lin et al., 2024)	1	23.92	31.65	0.316
	4	24.46	24.56	0.323
DMD2 (Yin et al., 2024a)	1	19.01	26.98	0.336
	4	19.32	20.86	0.332
SenseFlow (Ge et al., 2025)	4	22.83	26.05	0.335
DMAD (Ours)	1	16.11	32.42	**0.338**
	4	**14.47**	**19.88**	0.328
SDXL Teacher, cfg=6	100	19.36	21.38	0.332
SDXL Teacher, cfg=8	100	20.39	23.21	0.335

4.2 Main Results

Class-conditional image generation. Table 1 compares DMAD with recent baselines. Our one-step model achieves an FID of 1.24, outperforming DMD2, including its longer-training variant, and SiDA trained directly from the teacher. It also outperforms the EDM teacher with both ODE and SDE samplers. With a projected discriminator following D2O-F (Zheng & Yang, 2025), DMAD achieves an FID of **1.04**, surpassing D2O-F and SiD^{2}A, which is initialized from a SiD-distilled generator. The advantage comes from our formulation: the discriminator itself provides the distillation gradient, so discriminative representations directly guide generator training. DMD2, in contrast, requires a critic that must be refitted continually to the student's evolving distribution, five times per student update.

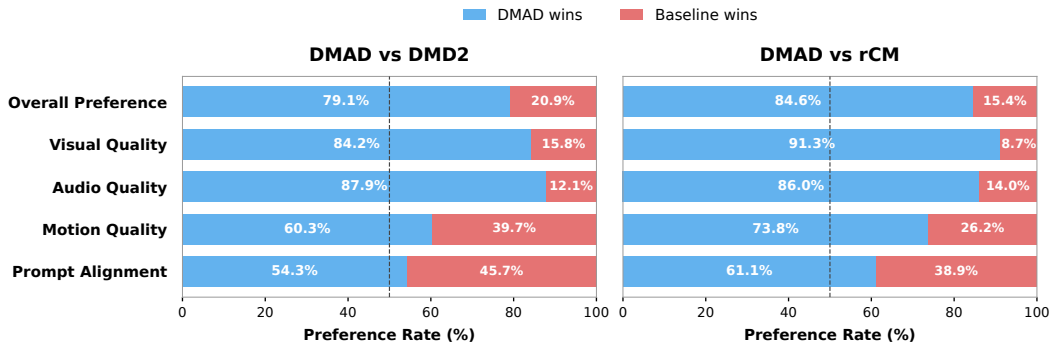


Figure 6: **User study comparing DMAD with DMD2 and rCM on MiniMax-H3.** All models use 4 sampling steps. Preference rates use prompt-level majority votes, excluding ties. Dashed lines indicate 50%.

Table 3: **VBench results for Wan2.1 T2V (480p).** †Retested with Diffusers and augmented prompts.

Model	Resolution	NFE	Total (↑)	Quality (↑)	Semantic (↑)
Wan2.1 T2V - 1.3B					
Wan2.1 T2V Teacher (Wan et al., 2025)†	832 × 480 × 81	50 × 2	83.02	83.95	79.26
DMD2 (Yin et al., 2024a)	832 × 480 × 81	4	84.56	85.58	80.50
sCM (Lu & Song, 2025)	832 × 480 × 81	4	82.06	84.00	74.30
rCM (Zheng et al., 2026)	832 × 480 × 81	4	84.43	85.38	80.63
DMAD (Ours)	832 × 480 × 81	4	**84.70**	85.98	79.55
Wan2.1 T2V - 14B					
Wan2.1 T2V Teacher (Wan et al., 2025)†	832 × 480 × 81	50 × 2	83.59	84.26	80.92
rCM (Zheng et al., 2026)	832 × 480 × 81	4	84.92	85.43	82.88
DMAD (Ours)	832 × 480 × 81	4	**85.15**	85.96	81.90

Text-to-image synthesis. We follow DMD2 (Yin et al., 2024a) for both training data and evaluation. Our generators are trained by distilling SDXL (Podell et al., 2023) using 500K real images from LAION-Aesthetics (Schuhmann et al., 2022) and corresponding teacher-generated samples, and evaluate on zero-shot COCO 2014 (Lin et al., 2014). Our four-step generator achieves an FID of 14.47 and a patch FID of 19.88 (Table 2), outperforming DMD2, SenseFlow (Ge et al., 2025), and the SDXL teacher on both metrics. With a single forward pass, DMAD achieves an FID of 16.11 and a CLIP score of 0.338, the best one-step FID and highest CLIP score among the compared models. Qualitative comparisons appear in Figure 8.

Text-to-video synthesis. We distill Wan2.1-T2V 1.3B, 14B (Wan et al., 2025) and MiniMax-H3 (MiniMax, 2026) into four-step generators using real videos from UltraVideo (Xue et al., 2025) and corresponding videos generated by the teacher. For VBench (Huang et al., 2024) evaluation, we generate videos using the same augmented prompt set as rCM (Zheng et al., 2026). Our 1.3B model achieves a VBench total score of 84.70, outperforming DMD2, sCM (Lu & Song, 2025), rCM, and the 50-step teacher (Table 3). At 14B, our model reaches 85.15, also surpassing rCM and the teacher. Both models achieve the highest quality scores among the compared methods. Additional samples from our four-step Wan2.1-T2V-14B student appear in Figure 9. Qualitative comparisons in Figure 4 show more natural colors and better preservation of prompted details, including petal texture and tiger anatomy. Our models achieve the highest VBench total and quality scores among the compared methods, while rCM scores higher on semantics.

For MiniMax-H3, Figure 5 shows improved visual quality and preservation of prompted elements, such as the bird in the crocodile scene. A user study on VideoGenEval (Yang et al., 2025) prompts evaluates five dimensions (Figure 6). DMAD receives overall preference rates of 79.1% against DMD2 and 84.6% against rCM, excluding ties, with the strongest preferences in visual and audio quality. Additional quantitative audio-video results are provided in Section C.

Table 4: **Ablation experiments of our DMAD.** One-step image generation on ImageNet, four-step image generation on COCO-10K (SDXL), and four-step video generation on VBench (Wan2.1-1.3B).

Setting	ImageNet	COCO-10K (SDXL)			VBench (Wan2.1-1.3B)		
	FID (↓)	FID (↓)	Patch FID (↓)	CLIP (↑)	Total (↑)	Quality (↑)	Semantic (↑)
w/o Teacher	1.30	**13.85**	26.03	0.312	82.20	83.91	75.36
w/o Real	1.67	21.37	23.70	0.325	84.38	85.57	79.59
w/o Gap-reweight	1.33	16.62	20.57	0.327	83.56	84.53	79.69
w/o Linear loss	1.31	14.78	21.04	0.326	84.12	85.19	79.86
DMAD (Ours)	**1.24**	14.47	**19.88**	**0.328**	**84.70**	85.98	79.55

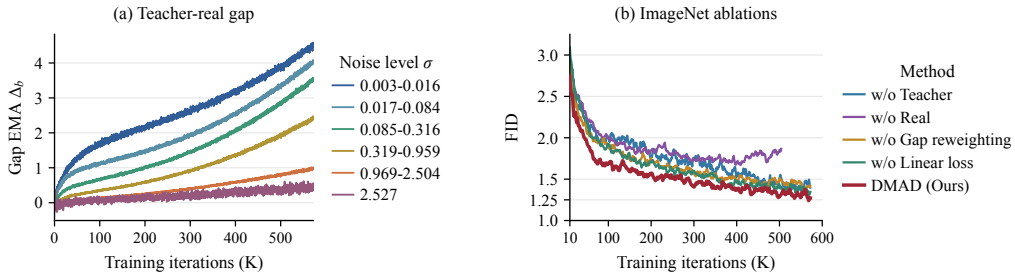


Figure 7: **ImageNet training dynamics.** (a) EMA of the real-head gap between real and teacher samples, the signal that drives gap-based reweighting (Section 3.3). (b) FID of DMAD and four ablations without smoothing on a 1K-iteration grid. The no-real curve ends at 504K iterations.

4.3 Ablation Studies

We study four ablations in Table 4: removing the teacher head, removing the real head, disabling gap reweighting with fixed loss weights, and replacing the linear generator loss with the non-saturating GAN loss. Removing a head also removes its discriminator and generator losses. For removing the real head, we also disable gap-based reweighting, which depends on the real-data head.

Removing the real head causes the largest FID increase on ImageNet and SDXL in Table 4. The teacher-only curve in Figure 7(b) also stays above the full model and rises again after about 400K iterations. Removing the teacher instead slows the decrease in ImageNet FID. Although it gives the lowest SDXL FID, it also gives the worst patch FID and CLIP, and reduces the VBench semantic score by 4.19 points. Our combined adversarial objective improves ImageNet convergence and SDXL detail and text alignment over either target alone. Replacing the linear generator loss with the non-saturating loss slows the reduction in ImageNet FID. As shown in Figure 7(b), our model first records FID below 1.5 at 209K iterations, compared with 354K for the non-saturating variant. The linear loss also gives better image FID and VBench total score in Table 4. Our linear objective uses the discriminator's logit gradient directly, without the additional sample weighting introduced by the non-saturating loss. Disabling gap reweighting worsens ImageNet FID and all SDXL metrics in Table 4, and lowers VBench total by 1.14 points. Adapting the teacher weight across noise levels lowers FID at matched iterations in Figure 7(b), while the real-data coefficient stays fixed.

5 Conclusion

We presented DMAD, a reformulation of distribution matching distillation that jointly learns from teacher-generated samples and real data without auxiliary score fitting. Starting from DMD's score-difference update, DMAD trains the student through linear losses on two discriminator logits, which at the discriminator optimum recover the same distribution-matching gradient. Training requires no online teacher evaluation and uses one discriminator update per student update. We further use the real-data head's gap between real and teacher samples to adapt teacher supervision across noise levels. Experiments show improvements over the compared few-step methods and multi-step teachers in image FID and VBench total score. On MiniMax-H3, users prefer our four-step student over DMD2 and rCM, particularly in visual and audio quality. These results support discriminative learning as an effective approach to distribution matching distillation.

AI USE STATEMENT

We used large language models to assist with polishing the writing, retrieving and discovering related literature, and drafting portions of the manuscript. The authors take responsibility for the final content of this work, including its claims, citations, and any text produced with the aid of these tools.

ETHICS STATEMENT

This work develops efficient distillation methods for image and video generation. While faster generation can benefit creative applications and research, it can also lower the cost of producing misleading or harmful synthetic media. Distilled models may inherit biases and inappropriate generation behaviors from their teachers and training data. We encourage responsible use of these methods, with appropriate safeguards against impersonation, misinformation, and non-consensual content.

REPRODUCIBILITY STATEMENT

Section 3 describes our method, and Appendix A provides the theoretical derivation. The experimental setup and implementation details, including model architectures, training hyperparameters, and evaluation protocols, are provided in Section 4 and Appendix B. We will publicly release our training and evaluation code, dataset preparation and preprocessing scripts, and experiment configurations to facilitate independent reproduction and further research.

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. *arXiv preprint arXiv:2602.07345*, 2026.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019.

Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *ECCV*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41667–41676, 2026.

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. *arXiv preprint arXiv:2506.00523*, 2025.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2672–2680, 2014.

Yuchao Gu, Guian Fang, Yuxin Jiang, Weijia Mao, Song Han, Han Cai, and Mike Zheng Shou. Anyflow: Any-step video diffusion model with on-policy flow map distillation. *arXiv preprint arXiv:2605.13724*, 2026.

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. In *NeurIPS*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

Allan Jabri, David Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. In *ICML*, 2023.

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

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24174–24184, 2024.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *ICLR*, 2024.

Youngjoong Kim, Deokyeong Lee, and Jaesik Park. Distribution matching distillation without fake score network. *arXiv preprint arXiv:2605.19256*, 2026.

Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024.

Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*, 2025a.

Shanchuan Lin, Ceyuan Yang, Hao He, Jianwen Jiang, Yuxi Ren, Xin Xia, Yang Zhao, Xuefeng Xiao, and Lu Jiang. Autoregressive adversarial post-training for real-time interactive video generation. *arXiv preprint arXiv:2506.09350*, 2025b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, volume 2025, pp. 50611–50649, 2025.

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. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16818–16829. IEEE, 2025.

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

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. *CoRR*, abs/2311.05556, 2023a. doi: 10.48550/ARXIV.2311.05556. URL https://doi.org/10.48550/arXiv.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. In *NeurIPS*, 2023b.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14297–14306, 2023.

MiniMax. Minimax h3: An open model breaking the boundaries between tasks and modalities, July 2026. URL https://www.minimax.io/blog/minimax-h3.

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. *arXiv preprint arXiv:2307.01952*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *SIGGRAPH*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

Neta Shaul, Chao Liu, Arash Vahdat, and Julius Berner. Parallel decoding distillation for fast image and video generation. *arXiv preprint arXiv:2607.26004*, 2026.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *ICLR*, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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. *Advances in neural information processing systems*, 36:8406–8441, 2023.

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. *arXiv preprint arXiv:2602.03139*, 2026a.

Zhuguanyu Wu, Ruihao Gong, Yang Yong, Yushi Huang, Xiangyu Fan, Lei Yang, Dahua Lin, and Xianglong Liu. Sgmd: Score gradient matching distillation for few-step video diffusion distillation. *arXiv preprint arXiv:2605.30116*, 2026b.

Yilun Xu, Weili Nie, and Arash Vahdat. One-step diffusion models with f-divergence distribution matching. *arXiv preprint arXiv:2502.15681*, 2025.

Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. *arXiv preprint arXiv:2506.13691*, 2025.

Yuhang Yang, Ke Fan, Shangkun Sun, Hongxiang Li, Ailing Zeng, FeiLin Han, Wei Zhai, Wei Liu, Yang Cao, and Zheng-Jun Zha. Videogen-eval: Agent-based system for video generation evaluation. *arXiv preprint arXiv:2503.23452*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024b.

Bowen Zheng and Tianming Yang. Revisiting diffusion models: From generative pre-training to one-step generation. *arXiv preprint arXiv:2506.09376*, 2025.

Kaiwen Zheng, Yongxin Chen, Huayu Chen, Guande He, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Direct discriminative optimization: Your likelihood-based visual generative model is secretly a gan discriminator. In *International Conference on Machine Learning*, pp. 78067–78094. PMLR, 2025.

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. In *International Conference on Learning Representations*, volume 2026, pp. 2582–2603, 2026.

Mingyuan Zhou, Huangjie Zheng, Yi Gu, Zhendong Wang, and Hai Huang. Adversarial score identity distillation: Rapidly surpassing the teacher in one step. *arXiv preprint arXiv:2410.14919*, 2024a.

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. In *Forty-first International Conference on Machine Learning*, 2024b.

Ziwei Zhou, Zeyuan Lai, Rui Wang, Yifan Yang, Zhen Xing, Yuqing Yang, Qi Dai, Lili Qiu, and Chong Luo. AVGen-Bench: A task-driven benchmark for multi-granular evaluation of text-to-audio-video generation, 2026. URL `https://arxiv.org/abs/2604.08540`.

A Proof of Proposition 1

Optimal discriminator. Fix t and $\boldsymbol{c}$ and write $P = p_t$, $Q = p_{\theta,t}$ for the target and student noisy marginals. As in the main text, densities are conditional on $\boldsymbol{c}$, which we omit from their arguments. Assume positive differentiable densities on common support and an unrestricted, unregularized discriminator with equal positive and negative class weights. The balanced BCE objective is

$$\mathcal{B}(h_\phi; p) = -\mathbb{E}_{t,\boldsymbol{c}}\left[\mathbb{E}_{\boldsymbol{x}_t\sim p_t} \log D_\phi(\boldsymbol{x}_t, t, \boldsymbol{c}) + \mathbb{E}_{\boldsymbol{x}_t\sim p_{\theta,t}} \log\big(1 - D_\phi(\boldsymbol{x}_t, t, \boldsymbol{c})\big)\right], \tag{10}$$

where $D_\phi = S(h_\phi)$ and $S(u) = (1 + e^{-u})^{-1}$. Pointwise, this is

$$-P(\boldsymbol{x}) \log S(h(\boldsymbol{x})) - Q(\boldsymbol{x}) \log\big(1 - S(h(\boldsymbol{x}))\big).$$

Using $\frac{\partial}{\partial h} \log S(h) = 1 - S(h)$ and $\frac{\partial}{\partial h} \log(1 - S(h)) = -S(h)$, the stationarity condition is $-P(1 - S(h)) + QS(h) = 0$, *i.e.* $S(h^\star) = P/(P + Q)$, which is the familiar optimal discriminator of Goodfellow et al. (2014). Hence,

$$h^\star(\boldsymbol{x}) = \log \frac{S(h^\star)}{1 - S(h^\star)} = \log \frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})} = \log p_t(\boldsymbol{x}) - \log p_{\theta,t}(\boldsymbol{x}).$$

Since the integrand is strictly convex in h for $P, Q > 0$, this stationary point is the unique minimizer. Differentiating in $\boldsymbol{x}$ yields Equation (3).

If positives and negatives instead use different fixed sampling priors $\pi_+(t, \boldsymbol{c})$ and $\pi_-(t, \boldsymbol{c})$, the optimal joint-classification logit additionally contains $\log[\pi_+(t, \boldsymbol{c})/\pi_-(t, \boldsymbol{c})]$. On their common support, this term is constant in $\boldsymbol{x}$, so its input gradient vanishes. The conditional log-ratio expression above assumes matching priors, not samplewise pairing.

Gradient of the distribution-matching objective. Write $\boldsymbol{x}_t = \alpha_t G_\theta(\boldsymbol{z}, \boldsymbol{c}) + \sigma_t \boldsymbol{\epsilon}$ and let $p_{\theta,t}$ be its density. The target p does not depend on θ. Assume the densities and reparameterization are sufficiently regular to interchange differentiation and integration, with support independent of θ. Then

$$\begin{aligned}\nabla_\theta D_{\mathrm{KL}}(p_{\theta,t} \| p_t) &= \nabla_\theta \, \mathbb{E}_{\boldsymbol{z},\boldsymbol{\epsilon}}\Big[\log p_{\theta,t}(\boldsymbol{x}_t) - \log p_t(\boldsymbol{x}_t) \Big] \\ &= \mathbb{E}_{\boldsymbol{z},\boldsymbol{\epsilon}}\left[\left(\frac{\partial \boldsymbol{x}_t}{\partial \theta} \right)^{\mathsf{T}} \big(\boldsymbol{s}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}) - \boldsymbol{s}_p(\boldsymbol{x}_t, t, \boldsymbol{c}) \big) \right] \\ &\quad + \mathbb{E}_{\boldsymbol{x}_t \sim p_{\theta,t}}\left[\nabla_\theta \log p_{\theta,t}(\boldsymbol{x})|_{\boldsymbol{x}=\boldsymbol{x}_t} \right],\end{aligned}$$

where the second term collects the dependence of the density on θ at fixed argument. That term vanishes, since $\mathbb{E}_{p_{\theta,t}}[\nabla_\theta \log p_{\theta,t}] = \int \nabla_\theta p_{\theta,t} = \nabla_\theta \int p_{\theta,t} = \mathbf{0}$. Averaging over t and $\boldsymbol{c}$ gives Equation (2). Their sampling distributions are held fixed as θ varies.

Equivalence. Replace h_ϕ in Equation (4) by the optimal logit at the current student distribution. Hold this function fixed during the generator update, while retaining its derivative with respect to $\boldsymbol{x}_t$. The chain rule gives

$$\begin{aligned}\nabla_\theta \mathcal{L}_G^{\mathrm{lin}} &= -\,\mathbb{E}\left[\left(\frac{\partial \boldsymbol{x}_t}{\partial \theta} \right)^{\mathsf{T}} \nabla_{\boldsymbol{x}_t} h^\star(\boldsymbol{x}_t, t, \boldsymbol{c}) \right] \\ &= -\,\mathbb{E}\left[\left(\frac{\partial \boldsymbol{x}_t}{\partial \theta} \right)^{\mathsf{T}} \big(\boldsymbol{s}_p - \boldsymbol{s}_\theta \big) \right] = \nabla_\theta \mathcal{L}_{\mathrm{DM}}(\theta; p).\end{aligned}$$

This is a local equality of generator gradients under an optimal, fixed critic. It does not differentiate through the dependence of the optimal critic on θ, or assert equality of the two loss values.

Extension to weighted objectives. The same equivalence holds when each noise level is assigned a nonnegative weight w_t that is independent of θ. In particular,

$$\mathbb{E}_{t,\boldsymbol{c}}\left[w_t D_{\mathrm{KL}}(p_{\theta,t} \| p_t) \right]$$

has the same generator gradient as

$$-\mathbb{E}\left[w_t h^\star(\boldsymbol{x}_t, t, \boldsymbol{c}) \right],$$

with $h^\star$ held fixed. This extension does not cover DMD's sample-dependent score normalization. It also treats our gap-based weights as fixed during each generator update rather than differentiating through their evolution over training.

Choice of generator loss. For any differentiable ℓ, a fixed critic gives

$$\nabla_\theta \, \mathbb{E}\big[w_t \ell(h_\phi) \big] = \mathbb{E}\left[w_t \ell'(h_\phi) \left(\frac{\partial \boldsymbol{x}_t}{\partial \theta} \right)^{\mathsf{T}} \nabla_{\boldsymbol{x}_t} h_\phi \right]. \tag{11}$$

Relative to the linear loss, the per-sample multiplier is $-\ell'(h_\phi)$, which is nonnegative for a non-increasing loss. The linear loss $\ell(h) = -h$ has multiplier one. The non-saturating loss $\ell(h) = -\log S(h)$ instead gives $1 - D_\phi$, reducing the contribution of samples the discriminator already accepts. At an optimal critic, this reweights the ideal distribution-matching gradient for each sample and can change the direction of the aggregate update. We compare the two losses in Section 4.3.

Two fixed targets. Applying the result separately to teacher samples and real data gives

$$h_T^\star = \log \frac{p_{T,t}}{p_{\theta,t}}, \qquad h_R^\star = \log \frac{p_{R,t}}{p_{\theta,t}}. \tag{12}$$

These are the unrestricted optima of the two balanced classification tasks. Assume both logits are optimal and the coefficients λ_T, λ_R are constant and nonnegative. With gap reweighting disabled, the generator update is the gradient of

$$\mathbb{E}_{t,\boldsymbol{c}}[\lambda_T D_{\mathrm{KL}}(p_{\theta,t} \| p_{T,t}) + \lambda_R D_{\mathrm{KL}}(p_{\theta,t} \| p_{R,t})].$$

Let $\Lambda = \lambda_T + \lambda_R > 0$ and $\rho = \lambda_T/\Lambda$. At each fixed $t, \boldsymbol{c}$, define

$$M_t(\boldsymbol{x}) = Z_t^{-1} p_{T,t}(\boldsymbol{x})^{\rho} p_{R,t}(\boldsymbol{x})^{1-\rho}, \qquad Z_t = \int p_{T,t}(\boldsymbol{x})^{\rho} p_{R,t}(\boldsymbol{x})^{1-\rho} \, \mathrm{d}\boldsymbol{x}.$$

For $0 < Z_t < \infty$, the weighted KL sum equals $\Lambda[D_{\mathrm{KL}}(p_{\theta,t} \| M_t) - \log Z_t]$. Since Z_t is independent of θ, the corresponding gradient is that of $\Lambda \mathbb{E}_{t,\boldsymbol{c}}[D_{\mathrm{KL}}(p_{\theta,t} \| M_t)]$. The geometric mixture is defined separately at each noise level. Forward noising does not generally commute with geometric mixing, and the marginals $p_{\theta,t}$ all arise from a single clean student distribution. This identity does not identify a clean geometric mixture as the global optimum, or apply as a fixed objective when gap reweighting changes the coefficients.

B Implementation Details

Training and evaluation configuration. We use AdamW with $(\beta_1, \beta_2) = (0, 0.99)$ and one discriminator update per generator update. All reported models use generator EMA weights. ImageNet and SDXL use an EMA half-life of 10^6 generated images with ramp-up factor 0.05. Wan and MiniMax-H3 uses power-function EMA (Karras et al., 2024) with $\sigma_{\text{rel}} = 0.1$. For ImageNet-$64 \times 64$, we train on 8 H100 GPUs with global batch size 384, learning rates $\eta_G = \eta_D = 2 \times 10^{-6}$, $\lambda_T = \lambda_R = 0.003$, and 570K iterations. For SDXL, we train on 8 H100 GPUs with global batch size 64, $\eta_G = \eta_D = 5 \times 10^{-7}$, $\lambda_T = \lambda_R = 0.005$, and 17K iterations. For Wan2.1-T2V-1.3B, we train on 8 H100 GPUs with global batch size 8, $\eta_G = \eta_D = 2 \times 10^{-6}$, $\lambda_T = \lambda_R = 1$, and 17K iterations. For Wan2.1-T2V-14B, we train on 64 H100 GPUs with global batch size 8, $\eta_G = \eta_D = 10^{-6}$, $\lambda_T = \lambda_R = 1$, and 19K iterations, using FSDP and context parallelism with groups of 8 GPUs. For evaluation, we follow the protocols of DMD2 (Yin et al., 2024a) for ImageNet and SDXL, and rCM (Zheng et al., 2026) for Wan2.1-T2V. For MiniMax-H3, we train on 8 H200 GPUs with global batch size 8 at a resolution of 1344×768 for 800 iterations, using $\eta_G = \eta_D = 4 \times 10^{-5}$ and $\lambda_T = \lambda_R = 1$. Both the generator and the discriminator backbone use LoRA with rank and alpha set to 128. For evaluation, we follow the protocols of DMD2 (Yin et al., 2024a) for ImageNet and SDXL, rCM (Zheng et al., 2026) for Wan2.1-T2V and AVGen-Bench (Zhou et al., 2026) for MiniMax-H3. For user study in MiniMax-H3, we conduct a pairwise user study using 387 prompts from VideoGenEval (Yang et al., 2025). For each prompt, we compare the audio-video output of DMAD with those of DMD2 and rCM, yielding 774 pairs. All models use four sampling steps. Pairs are presented as A/B comparisons with the left/right placement of DMAD balanced within each baseline. Each pair is evaluated by three annotators on five dimensions: prompt alignment, visual quality, motion quality, audio quality, and overall preference. For each dimension, annotators select A, B, or a tie. This yields 2,322 pairwise evaluations. For each prompt and dimension, we aggregate the three judgments by majority vote, recording a tie if all three judgments differ. Each baseline therefore has 387 prompt-level outcomes per dimension. We report the preference rate for DMAD as $N_{\text{win}}/(N_{\text{win}} + N_{\text{loss}})$ over these outcomes, excluding ties. For overall preference, the win/tie/loss counts are 204/129/54 against DMD2 and 247/95/45 against rCM, with tie rates of 33.3% and 24.5%, respectively.

Discriminator architecture. The real and teacher heads share a feature backbone. For ImageNet and SDXL, the diffusion discriminator uses a pretrained U-Net encoder and middle block with two convolutional heads. Each head downsamples the features using 4×4 convolutions, group normalization, and SiLU, then predicts a scalar logit. For the projected-discriminator variant on ImageNet, we follow D2O-F (Zheng & Yang, 2025) and use frozen ImageNet-pretrained VGG16-BN and EfficientNet-Lite0 backbones with two trainable sets of multi-scale discriminator heads. For Wan, two MLP heads with a SiLU activation operate on final-block tokens, and their logits are averaged over the video. For MiniMax-H3, final-block features undergo non-affine RMS normalization before entering two MLP heads, each consisting of two linear layers with a SiLU activation between them. For noise-band reweighting, ImageNet uses six bands over Karras indices 20–500, grouped by 100 indices, with the last containing only index 500 ($\sigma \approx 2.527$), SDXL uses 100-timestep bands, Wan uses 10 equal-probability bands under its discriminator-time sampling distribution, and MiniMax-H3 uses ten equal-width bands in the base time variable over $[0, 1]$. We use $\tau = 0.5$ for ImageNet and SDXL training and $\tau = 2$ for Wan and MiniMax-H3.

B.1 Training Speed and Memory

Table 5 compares the online training cost of DMD2 and DMAD on NVIDIA H200 GPUs, with matched batch sizes and input shapes within each setting. ImageNet and Wan-14B use full-parameter training, while MiniMax-H3 uses rank-128 LoRA. We measure CUDA-synchronized update times after warm-up and peak allocated GPU memory. Time includes one generator update and its associated critic updates. Following the official DMD2 configuration, we use a 5:1 critic-to-generator update ratio for DMD2, compared with 1:1 for DMAD. Evaluation and checkpointing are excluded.

Table 5: Online training time per generator update, including its critic updates, and peak allocated memory per GPU. Lower is better.

		Time (s)		Memory (GiB/GPU)	
Setting	H200 GPUs	DMD2	Ours	DMD2	Ours
ImageNet-64 × 64	1	2.321	**0.496**	40.34	**33.99**
Wan2.1-T2V-14B	4	162.9	**38.50**	131.04	**124.56**
MiniMax-H3-33B	4	818.1	**233.3**	127.23	**114.72**

B.2 Toy Gradient Estimation

Distributions. We compare two estimates of the same correction field using two-dimensional Gaussian mixtures with known parameters. The target p represents real data, and q is a fixed student distribution. For $k = 0, \dots, 7$, define

$$\boldsymbol{\mu}_k = 2\big(\cos(2\pi k/8), \sin(2\pi k/8)\big)^{\mathsf{T}}, \qquad \boldsymbol{\delta} = (0.08, -0.04)^{\mathsf{T}}.$$

The clean distributions are

$$p(\boldsymbol{x}) = \sum_{k=0}^{7} \tfrac{1}{8}\, \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, 0.35^2 \boldsymbol{I}),$$

$$q(\boldsymbol{x}) = \sum_{k=0}^{7} \pi_k\, \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k + \boldsymbol{\delta}, 0.35^2 \boldsymbol{I}),$$

where $(\pi_0, \dots, \pi_7) = (0.16, 0.14, 0.12, 0.10, 0.10, 0.12, 0.13, 0.13)$. Adding independent noise $\sigma\boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, gives marginals p_σ and q_σ. Their component means and weights are unchanged, and their component covariance is $v_\sigma \boldsymbol{I}$, where $v_\sigma = 0.35^2 + \sigma^2$.

Analytic reference. For a mixture f_σ with weights a_k and means $\boldsymbol{m}_k$, its posterior component weights and score are

$$r_k(\boldsymbol{x}, \sigma) = \frac{a_k \exp\big(-\|\boldsymbol{x} - \boldsymbol{m}_k\|^2/(2v_\sigma)\big)}{\sum_j a_j \exp\big(-\|\boldsymbol{x} - \boldsymbol{m}_j\|^2/(2v_\sigma)\big)},$$

$$\boldsymbol{s}(\boldsymbol{x}, \sigma) = \nabla_{\boldsymbol{x}} \log f_\sigma(\boldsymbol{x}) = \frac{\sum_k r_k(\boldsymbol{x}, \sigma)\boldsymbol{m}_k - \boldsymbol{x}}{v_\sigma}.$$

Applying this expression separately to the target and student gives the ground-truth field

$$\boldsymbol{g}^\star = \boldsymbol{s}_p - \boldsymbol{s}_q = \nabla_{\boldsymbol{x}} \log \frac{p_\sigma}{q_\sigma}.$$

Balanced binary classification has the ideal logit $h^\star = \log(p_\sigma/q_\sigma)$, motivating the comparison with $\nabla_{\boldsymbol{x}} h_\phi$. Analytic scores and density ratios are used only for evaluation. Neither estimator receives them as training supervision.

Training. We first train a teacher noise predictor on samples drawn directly from p for 20,000 updates. It has two width-64 tanh hidden layers and a linear two-dimensional output. Its input is $(x_1/3, x_2/3, \log(\sigma/0.5))$, and its denoising objective is $\frac{1}{2}\mathbb{E}\|\epsilon_\omega(\boldsymbol{x}_0 + \sigma\boldsymbol{\epsilon}, \sigma) - \boldsymbol{\epsilon}\|^2$. Noise levels are sampled uniformly from $\{0.2, 0.5, 1.0\}$ for each example during pretraining and adaptation.

Both estimators inherit this pretrained teacher. The fake noise predictor copies all its parameters. The discriminator copies both hidden layers and uses a new scalar linear head, whose weights have independent Gaussian initialization with standard deviation 0.02 and whose bias is zero. The teacher then stays frozen. All parameters of the fake predictor and discriminator, including their hidden layers, are adapted for 1,000 updates. The fake predictor uses denoising on q. The discriminator uses balanced binary cross-entropy on noisy target positives and student negatives.

Pretraining and adaptation use Adam with learning rate 0.001, $(\beta_1, \beta_2) = (0.9, 0.999)$, and numerical constant 10^{-8}. Each adaptation optimizer starts with fresh state. The denoising batch contains 512 student examples. The classification batch contains 256 target and 256 student examples, sharing those student inputs with the first half of the denoising batch. Pretraining also uses batch size 512. Both branches therefore receive 512 training inputs per update. Computation uses float64.

Evaluation. The experiment in Figure 2 uses seed 0 and $\sigma = 0.5$. The labels *DMD gradient* and *Ours gradient* refer respectively to

$$\widehat{\boldsymbol{g}}_{\mathrm{DMD}} = -\frac{\epsilon_{\mathrm{teacher}}}{\sigma} + \frac{\epsilon_{\mathrm{fake}}}{\sigma}, \qquad \widehat{\boldsymbol{g}}_{\mathrm{ours}} = \nabla_{\boldsymbol{x}} h_\phi.$$

Both are evaluated on the same $N = 2{,}048$ independent held-out samples $\boldsymbol{x}_i \sim q_{0.5}$ using

$$\mathrm{NMSE}(\widehat{\boldsymbol{g}}) = \frac{N^{-1}\sum_i \|\widehat{\boldsymbol{g}}(\boldsymbol{x}_i) - \boldsymbol{g}^\star(\boldsymbol{x}_i)\|^2}{N^{-1}\sum_i \|\boldsymbol{g}^\star(\boldsymbol{x}_i)\|^2}.$$

Their NMSE values are 1.539462 and 0.144375. This experiment isolates single-target gradient estimation with a fixed student distribution. Target samples for both teacher pretraining and discriminator training are drawn directly from p.

C Additional Audio-Video Evaluation

We follow the evaluation metrics and score aggregation of AVGen-Bench (Zhou et al., 2026), using Qwen3-Omni-Thinking as the VLM judge. The same H3-enhanced prompts are used for generation and judging across all methods. Table 6 reports all eleven metrics for MiniMax-H3. DMAD achieves a total score of 65.4, compared with 62.6 for DMD2 and 59.8 for rCM, while the teacher scores 67.2.

Table 6: **AVGen-Bench results on MiniMax-H3.** All scores are normalized to 0–100 (↑). Bold marks the best total among distilled models.

		Quality		Synchronization		Fine-grained evaluation						
Method	Total	Visual	Audio	AV	Lip	Face	Lo-Phy	Hi-Phy	Holistic	Text	Speech	Music
DMD2	62.6	90.1	67.4	56.7	22.7	46.1	80.2	95.0	87.1	64.6	81.0	0.0
rCM	59.8	84.0	65.8	48.7	15.9	49.0	79.7	95.3	87.3	48.3	81.3	6.2
DMAD (Ours)	**65.4**	93.3	70.4	49.1	30.6	53.1	82.3	93.2	87.4	70.9	80.6	11.8
Teacher	67.2	92.6	70.6	58.3	39.0	59.2	79.7	94.5	87.0	74.6	84.0	1.6

D Additional Qualitative Results

D.1 SDXL Comparisons

Qualitative comparisons in Figure 8 show that DMAD produces more natural colors and textures and follows the prompts more closely, while some DMD2 and SenseFlow samples appear oversaturated. The tennis example highlights natural skin texture, while the bird example shows distinct wing structure and fine feather detail. In the bench example, only our model satisfies the instruction to exclude cats. It also generates exactly four people in the family portrait, whereas the teacher, DMD2, and SenseFlow include extra people.

D.2 MiniMax-H3 Teaser Prompts

We list the prompts for the six samples in Figure 1, ordered from left to right and top to bottom.

1. A medium shot opens on a seasoned, grey-bearded man in sunglasses and a paisley shirt, Beside him, a younger man in a tank top. In the background, a vibrant mural splashes across a wall, hinting at an urban setting. Faint city murmurs and distant chatter drift in, accompanied by a mellow, soulful hip-hop beat that adds a contemplative yet grounded atmosphere. The older man speaks while the younger man turns toward him, laughs softly, shakes his head in amusement, and then looks back at the camera. The camera slowly pushes in. "The city always got a story," the older man murmurs with a slight nod. The younger man chuckles and replies, "Yeah, you always say that."
2. A breathtaking, painterly 2D animated continuous visual narrative, rendered with the lush, vibrant, and slightly surreal, almost dreamlike, infused with the intricate, delicate detail of traditional Japanese woodblock prints (Ukiyo-e), follows a young, adventurous, and kind-hearted girl (perhaps with bright, curious eyes and wearing simple, practical, beautifully patterned traditional



Figure 8: **Quality comparison of DMAD with the SDXL teacher, DMD2 (Yin et al., 2024a), and SenseFlow (Ge et al., 2025).** All distilled models use 4 sampling steps, while the teacher uses 50 sampling steps. Images in each row share the same text prompt and initial Gaussian noise. Our model produces more natural colors and realistic textures with better text alignment. (Zoom in for details.)

Japanese farm attire) as she befriends a colossal, gentle, ancient Forest Spirit. The Spirit is a magnificent, awe-inspiring creature, its form a harmonious blend of animal and plant – perhaps with moss-covered, antler-like branches, fur like shimmering leaves that change color with its mood, and eyes like deep, tranquil forest pools. They meet in a sun-dappled, sacred grove deep within an ancient, primeval forest, where impossibly tall, gnarled trees form a living cathedral and tiny, glowing, friendly forest sprites (Kodama-like) peek from behind mossy rocks and giant, fantastical mushrooms. The girl, initially awestruck, offers the massive Spirit a small, carefully cultivated offering – perhaps a perfectly ripe persimmon or a handful of wild berries – her gesture one of pure, innocent respect and affection. The Forest Spirit responds with a slow, gentle inclination of its massive head, its leafy fur rustling like a thousand whispers, and perhaps causes a shower of magical, luminous flower petals to drift down from the canopy, or a tiny, new sapling to sprout at the girl's feet. The animation captures the incredible, detailed textures of the forest, the Spirit's majestic yet gentle presence, and the profound, unspoken emotional connection forming between the child and this ancient guardian of nature. The color palette is a rich symphony of deep forest greens, earthy browns, vibrant floral hues, and the soft, magical

glow of the sprites and the Spirit's own subtle luminescence. This continuous, sweeping visual journey is a celebration of the profound, often mystical, bond between humanity and nature, the innocence and courage of childhood, and the power of kindness and respect to bridge even the most fantastical of divides, an affectionate, visually intoxicating ode to ecological harmony and interspecies understanding. The only implied sounds are the gentle rustling of leaves, the distant calls of unseen forest birds, the girl's soft, respectful breathing, the Spirit's deep, resonant, almost inaudible hum, and a soaring, emotionally resonant, orchestral score.

3. The video opens with a medium, eye-level shot of a beautiful woman with dark hair and warm brown eyes. She wears a magnificent, high-fashion flamingo dress with layers of pink and fuchsia feathers, complemented by whimsical pink, heart-shaped sunglasses. She walks with serene confidence through the crystal-clear, shallow turquoise water of a sun-drenched lagoon. The camera slowly pulls back to a medium-wide shot, revealing the breathtaking scene as the dress's long train glides and floats gracefully on the water's surface behind her. The cinematic, dreamlike atmosphere is enhanced by the vibrant colors of the dress against the serene, minimalist landscape, capturing a moment of pure elegance and high-fashion fantasy.
4. An adorable happy otter confidently stands on a surfboard wearing a yellow lifejacket, riding along turquoise tropical waters near lush tropical islands, 3D digital render art style.
5. Static overhead shot of a printed photo of a tree trunk lying on a wooden table.Action Sequence:1, A real human hand enters, places a single pinecone on the paper next to the tree hole, and exits immediately.2, A realistic 3D squirrel emerges from the 2D hole in the photo. The squirrel comes out empty-handed.3, The squirrel sniffs the pinecone sitting on the paper, looks curious, blinks, and tilts its head.4, The squirrel reaches out, grabs that specific pinecone from the table.Key details: Seamless interaction between real world and photo, surreal VFX, squirrel paws are empty initially, heavy weight perception on the pinecone.
6. An origami dragon unfolds from paper in stop-motion style, then flies over folded mountains. The camera pans with paper dust fluttering midair. Every flap and fold is sharply geometric.

D.3 WAN2.1-T2V-14B SAMPLES

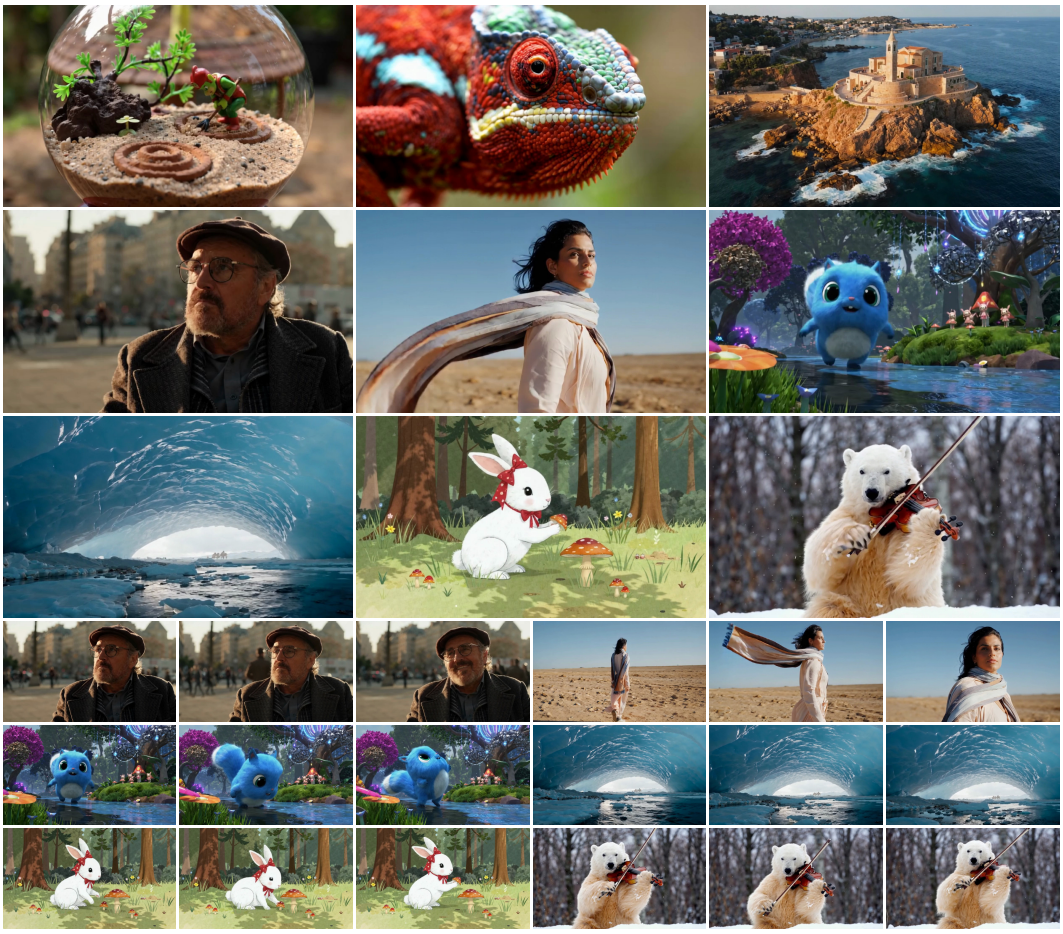


Figure 9: Video samples from our four-step Wan2.1-T2V-14B student. Top: representative frames from nine generated videos. Bottom: three frames in chronological order from each of six videos. Full prompts are provided in Section D.3.

We list the prompts for the nine samples in the upper grid of Figure 9, ordered from left to right and top to bottom.

1. A close up view of a glass sphere that has a zen garden within it. There is a small dwarf in the sphere who is raking the zen garden and creating patterns in the sand.
2. This close-up shot of a chameleon showcases its striking color changing capabilities. The background is blurred, drawing attention to the animal's striking appearance.
3. A drone camera circles around a beautiful historic church built on a rocky outcropping along the Amalfi Coast, the view showcases historic and magnificent architectural details and tiered pathways and patios, waves are seen crashing against the rocks below as the view overlooks the horizon of the coastal waters and hilly landscapes of the Amalfi Coast Italy, several distant people are seen walking and enjoying vistas on patios of the dramatic ocean views, the warm glow of the afternoon sun creates a magical and romantic feeling to the scene, the view is stunning captured with beautiful photography.
4. An extreme close-up of an gray-haired man with a beard in his 60s, he is deep in thought pondering the history of the universe as he sits at a cafe in Paris, his eyes focus on people offscreen as they walk as he sits mostly motionless, he is dressed in a wool coat suit coat with a button-down shirt , he wears a brown beret and glasses and has a very professorial appearance, and the end he offers a subtle closed-mouth smile as if he found the answer to the mystery of life, the lighting is very cinematic with the golden light and the Parisian streets and city in the background, depth of field, cinematic 35mm film.

5. A female traveler traverses a desert alone. The camera first shows a distant view of her slightly unsteady footsteps and flowing clothes, then switches to a medium side shot of her scarf billowing in the wind, and finally focuses on her determined gaze.
6. 3D animation of a small, round, fluffy creature with big, expressive eyes explores a vibrant, enchanted forest. The creature, a whimsical blend of a rabbit and a squirrel, has soft blue fur and a bushy, striped tail. It hops along a sparkling stream, its eyes wide with wonder. The forest is alive with magical elements: flowers that glow and change colors, trees with leaves in shades of purple and silver, and small floating lights that resemble fireflies. The creature stops to interact playfully with a group of tiny, fairy-like beings dancing around a mushroom ring. The creature looks up in awe at a large, glowing tree that seems to be the heart of the forest.
7. The camera slowly pushes forward into a breathtaking ice cave, its jagged walls sculpted by nature into intricate patterns of blues and whites, reflecting the ethereal light from an opening ahead. The crunch of ice underfoot and the drip-drip of melting water create a serene, echoing soundscape. As the camera moves closer, a gentle, ambient melody begins, swelling with the light from the cave's exit. The camera emerges from the narrow opening into a vast, sun-drenched valley, revealing a group of polar bears playfully sliding down an ice slope, their roars echoing with joy.
8. In the style of watercolor illustration, a little white rabbit wearing a bow is picking mushrooms in the forest.
9. A polar bear is playing the violin in the snow.

D.4 IMAGENET-64×64 SAMPLES

Figures 10 and 11 show one-step samples from our FID of 1.24 and FID of 1.04 results on ImageNet. Each figure shows a 32×32 grid of generated images at their native 64×64 resolution.



Figure 10: One-step samples from DMAD on ImageNet-64×64 (FID 1.24). Zoom in for details.

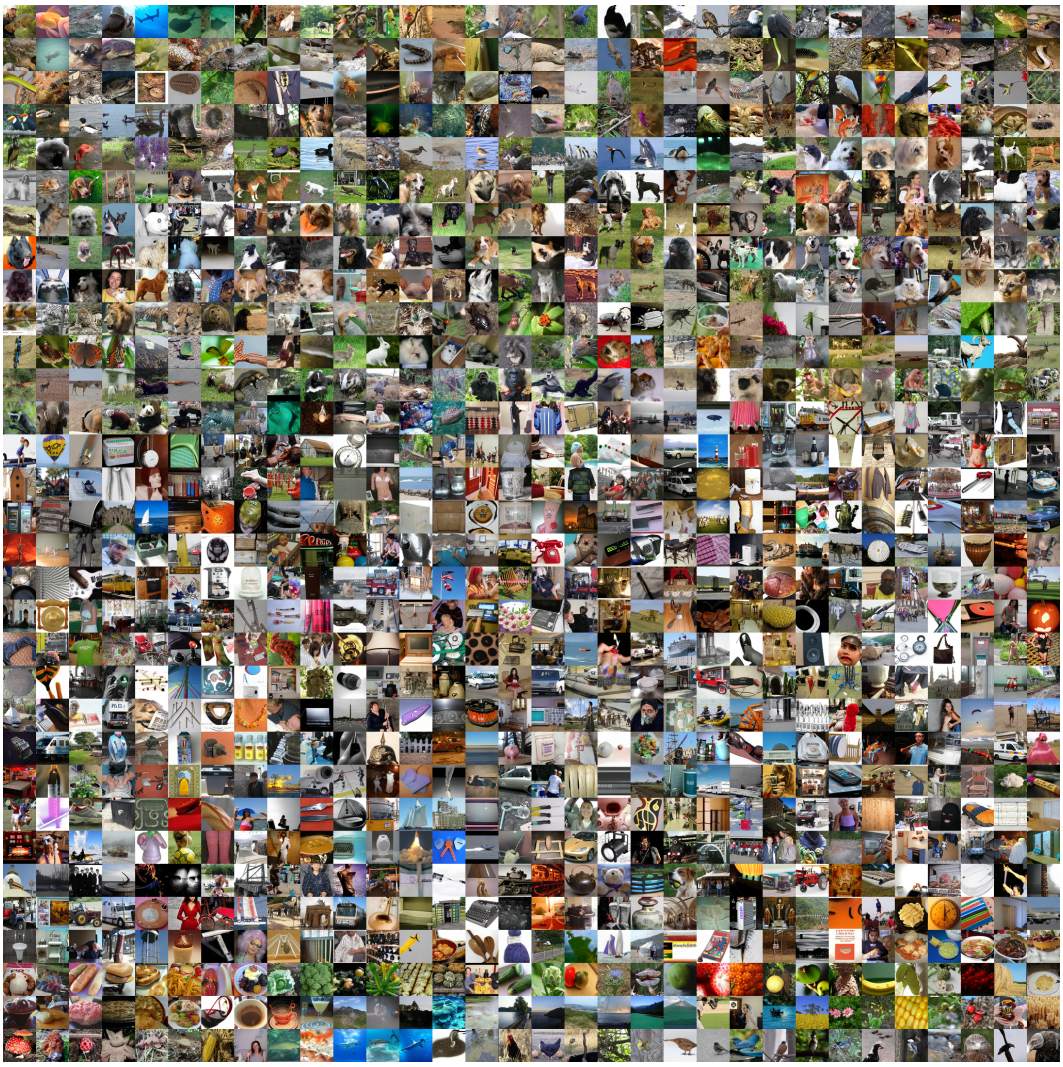
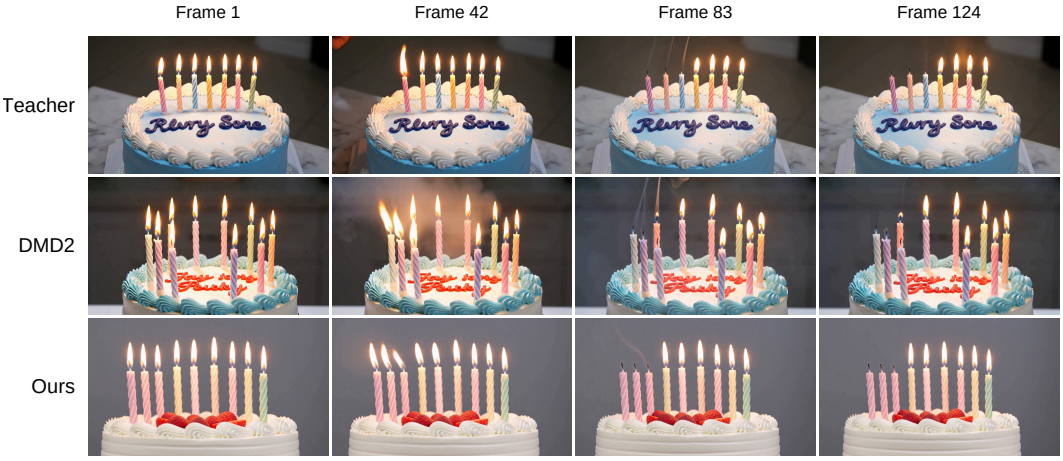


Figure 11: One-step samples from DMAD on ImageNet-64×64 (FID 1.04). Zoom in for details.

D.5 More results on MiniMax-H3

Prompt: Birthday cake with **exactly nine lit candles**; a child blows out **three candles on the left**, **leaving six burning**. Smoke rises only from the extinguished three.



Prompt: On the rooftop of a Korean high school, distant city lights glimmer in the background with a soft wind rustling, and stars twinkle in the night sky. The girl leans against the railing, lost in thought. The boy walks over with **two cans of cola**, **hands one to her**, and she takes it and pops the tab open. Boy (casual tone, Korean): "숙제 다 했어? 왜 여기 있어?" Girl (sighing, Korean): "시험이 너무 무서워". Boy (gentle tone, Korean): "걱정 마, 넌 잘할 거야."



Figure 12: **Qualitative comparisons of our DMAD and baseline models on MiniMax-H3.** Within each case, rows show Teacher, DMD2, and Ours, from top to bottom. Columns show frames 1, 42, 83, and 124, using the same indices for all methods. All compared videos are generated using identical initial noise and text prompts.

Prompt: An empty white-walled room containing only **a single wooden chair in the far corner**: **no windows**, no pictures, **no lamps**, no other furniture. The camera slowly dollies forward toward the chair.



Prompt: A beautiful woman wearing red clothes. A round of **glances and a laugh**.

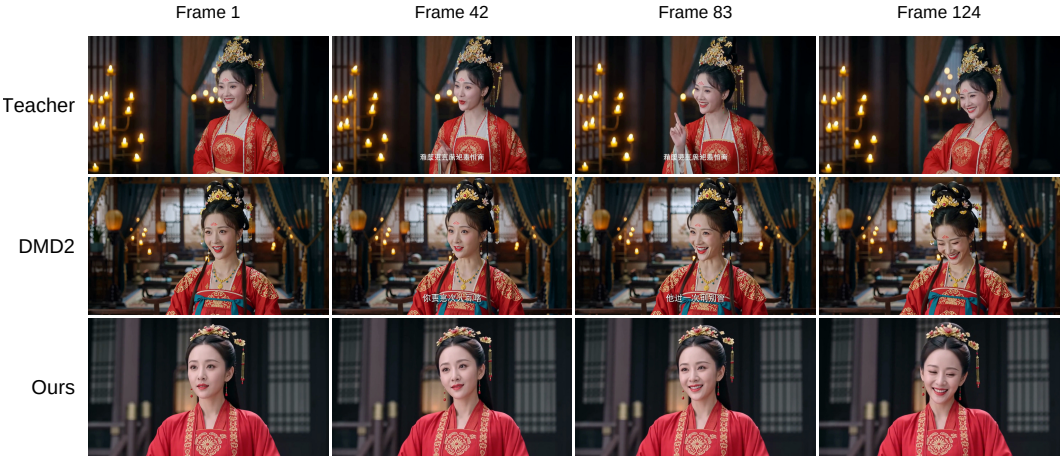
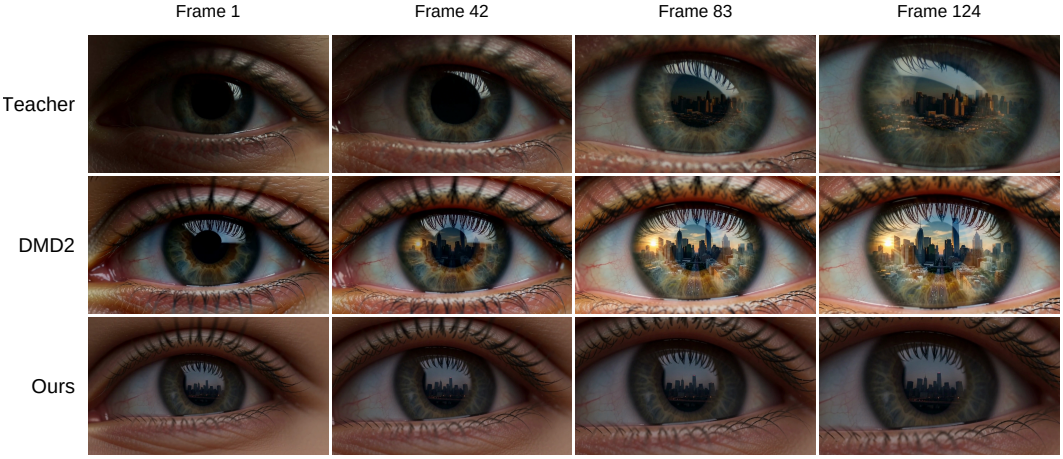


Figure 13: **Qualitative comparisons of our DMAD and baseline models on MiniMax-H3.** Within each case, rows show Teacher, DMD2, and Ours, from top to bottom. Columns show frames 1, 42, 83, and 124, using the same indices for all methods. All compared videos are generated using identical initial noise and text prompts.

Prompt: Extreme close-up of a an eye with **city reflected in it**.

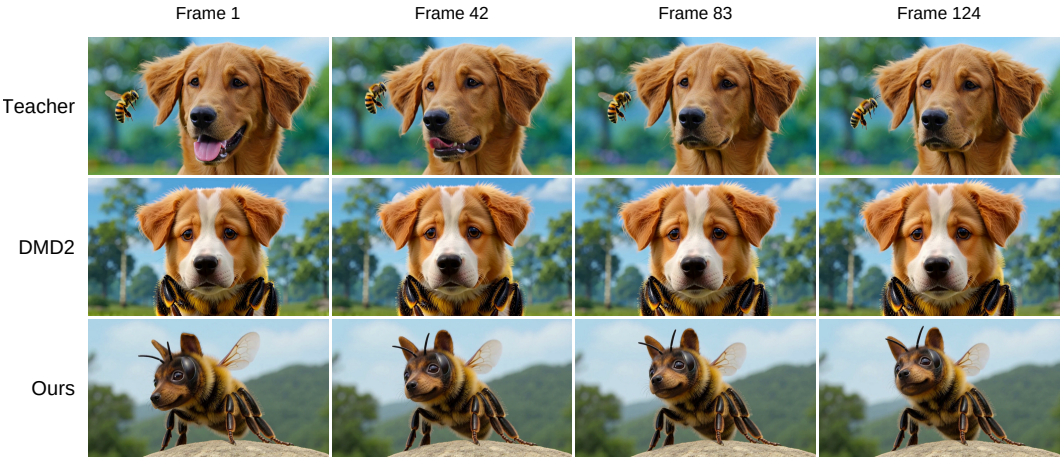


Prompt: Talking to the barista **from across the counter**, a woman in a coffee shop places an order for a cup of coffee with cream and sugar, and a chocolate croissant. The barista listens to the order, responds sure :) and then turns to the commercial espresso machine that is behind him. The woman **patiently waits across the counter** as her order is being prepared by the barista.



Figure 14: **Qualitative comparisons of our DMAD and baseline models on MiniMax-H3.** Within each case, rows show Teacher, DMD2, and Ours, from top to bottom. Columns show frames 1, 42, 83, and 124, using the same indices for all methods. All compared videos are generated using identical initial noise and text prompts.

Prompt: A **bee with a puppy's head**



Prompt: **First-person perspective**, high-speed flight, symmetrical composition, rotation, countless lightning bolts amidst dark clouds, motion blur.

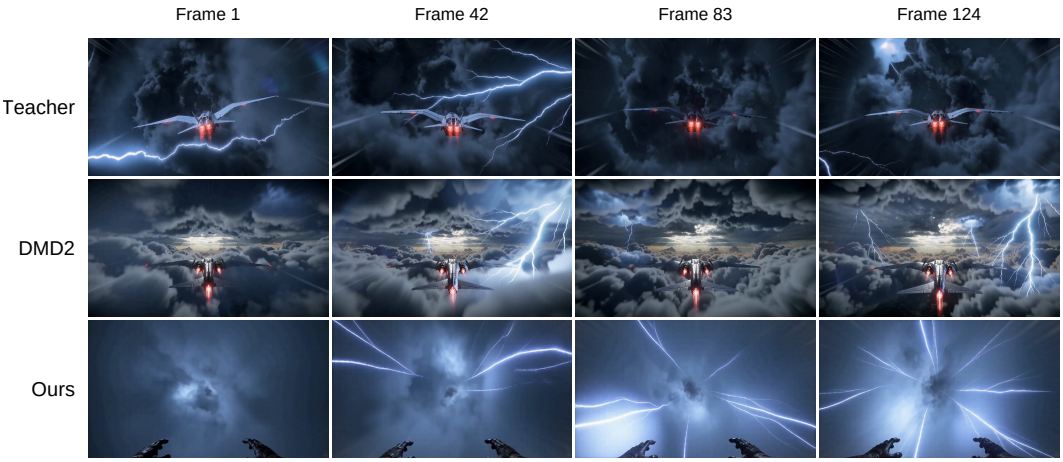


Figure 15: **Qualitative comparisons of our DMAD and baseline models on MiniMax-H3.** Within each case, rows show Teacher, DMD2, and Ours, from top to bottom. Columns show frames 1, 42, 83, and 124, using the same indices for all methods. All compared videos are generated using identical initial noise and text prompts.