RAMP: ROLLOUT-AWARE MODE PRECONDITIONING FOR VIDEO DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-step autoregressive video generators enable efficient inference, but prolonged self-forcing distribution matching distillation (SF-DMD) can lead to oversaturated colors and abnormal contrast. We attribute this failure to SF-DMD favoring narrow, undesirable regions that are easier for an imperfect student to fit than broader high-quality regions, with fake-score errors further amplifying this bias.

These observations motivate reducing undesirable mode before SF-DMD. We therefore propose RAMP, a rollout-aware mode-preconditioning framework that gradually shifts consistency distillation from teacher-forced states to states visited by the student's own autoregressive rollouts before switching to pure SF-DMD. Autoregressive rollout can move states out of the basin of narrow undesirable modes and into transition regions where the broader high-quality mode dominates the teacher posterior. Consistency supervision can then correct these rollout-induced deviations, while the resulting improved prefixes provide better contexts for subsequent autoregressive blocks, progressively reducing undesirable visitation before SF-DMD. Compared with existing autoregressive distillation methods, RAMP achieves the best overall VBench score and lowest HSV saturation, while maintaining strong CFM-7B performance, effectively reducing oversaturation without compromising generation quality.

1 INTRODUCTION

Video diffusion models have made remarkable progress in visual fidelity, motion realism, and text controllability, and now underpin world models, embodied agents, and interactive content creation (Team et al., 2026b;a; HY-World et al., 2026; Wang et al., 2026; Zhang et al., 2026; Gao et al., 2026). However, most state-of-the-art models rely on bidirectional temporal modeling (Wan et al., 2025; Kong et al., 2024), which is poorly suited to real-time interactive generation. Autoregressive models instead generate video incrementally, making them a natural choice for interactive world modeling (Liu et al., 2025; Yang et al., 2025; Zhu et al., 2026a; Gao et al., 2024; ai et al., 2025). However, building a high-quality few-step autoregressive model from scratch remains challenging and computationally expensive. A practical alternative is therefore to distill a pretrained bidirectional teacher into a few-step autoregressive student, preserving its generative capability while enabling causal, low-latency generation.

Recent methods commonly rely on self-forcing distribution matching distillation (SF-DMD), which trains on the student's own rollout trajectories and reduces the train–test mismatch of teacher forcing (Huang et al., 2025; Zhao et al., 2026a; Zhu et al., 2026b; Zheng et al., 2026; Wu et al., 2026). However, subsequent distribution matching can lead to severe oversaturation across diverse prompts (Li et al., 2026; Chen et al., 2026; You et al., 2026), as shown in Fig. 1.We interpret this behavior as SF-DMD shifting probability from broad, high-quality *good modes* toward narrow, color-degraded *bad modes* retained by the student initialization. High-quality samples span diverse appearances and greater variability, whereas the observed bad modes occupy a much narrower region of the output space, making the good modes harder to cover and the bad modes easier to fit. Reverse-KL optimization can amplify this mode-dependent fitting imbalance, while fake-score errors may further bias the update toward undesirable modes. **This makes the pre-SF-DMD student distribution particularly important: reducing bad-mode probability before distribution matching can mitigate the tendency toward collapse.**

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: **Bad-mode collapse under SF-DMD.** Representative generations show that SF-DMD from conventional initializations can amplify oversaturated visual modes across different prompts.

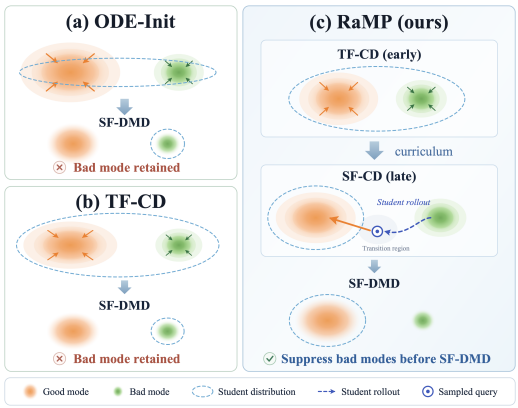

Figure 2: Conceptual comparison of preconditioning strategies before SF-DMD.

The consistency objective provides a useful clue for how such preconditioning can be achieved. Unlike reverse-KL optimization, its target is determined by the teacher distribution around the queried state. In the transition region between a high-probability good mode and a low-probability bad mode, the larger probability of the good mode gives it a broader region of dominance, allowing consistency supervision to guide more states toward the good-mode trajectory. However, this mode-correction behavior is difficult to exploit with conventional initialization strategies. ODE-based distillation relies on a finite set of teacher-generated trajectories, while teacher-forcing consistency distillation (TF-CD) queries states sampled from the teacher distribution (Zhao et al., 2026a; Zheng et al., 2026). Their state coverage therefore remains largely confined to teacher-defined distributions, which tends to preserve the bad-mode branch and leaves much of the transition region unexplored, as illustrated in Fig. 2(a,b).

Self-forcing rollout changes this picture. Autoregressive errors move the student beyond these teacher-defined states, allowing some samples near the narrow bad mode to leave its basin and enter the transition region where consistency supervision favors the good mode. Applying CD to these rollout-induced states redirects them toward good-mode trajectories, producing improved prefixes that further influence subsequent autoregressive blocks and progressively reduce bad-mode visitation.

Based on this observation, we propose RAMP, which injects self-forcing into consistency distillation to precondition the student before SF-DMD. By applying consistency supervision directly to student-visited states, RAMP reduces bad-mode visitation and provides a more favorable initialization for subsequent pure SF-DMD (See Fig. 2(c)).

Our main contributions are summarized as follows.

- We interpret oversaturation and related visual degradations in SF-DMD as manifestations of bad-mode amplification, and analyze two mechanisms that can drive this amplification: reverse-KL preference under imperfect initialization and online fake-score estimation errors.

- We propose RAMP, which combines self-forcing with consistency distillation to suppress bad modes in the student distribution before SF-DMD begins.
- Compared with representative autoregressive video distillation methods, RAMP achieves the best overall VBench score and lowest HSV saturation, while maintaining strong CFM-7B performance and causal, real-time inference.

2 Related Work

2.1 Video Generation and Diffusion Models

Diffusion models generate high-quality samples by progressively reversing a noise-perturbation process and have become the dominant framework for image and video generation (Ho et al., 2020; Rombach et al., 2022). Early video diffusion models extended image diffusion by jointly modeling spatial and temporal dimensions (Ho et al., 2022; Blattmann et al., 2023). More recent advances in Transformer-based diffusion and large-scale training have further improved video generation quality and text–video alignment (Peebles & Xie, 2023; Wan et al., 2025). However, these models typically rely on bidirectional spatiotemporal attention and require dozens of iterative denoising steps, making them difficult to deploy in interactive applications that demand low latency and continuous generation.

2.2 Autoregressive Diffusion Distillation

Existing diffusion distillation methods use several complementary objectives, including direct regression to teacher-generated trajectories, consistency training across noise levels, and distribution matching (Salimans & Ho, 2022; Song et al., 2023; Yin et al., 2024). Recent autoregressive video distillation methods often combine these objectives, using regression or consistency distillation for initialization and SF-DMD for further refinement (Yin et al., 2025; Huang et al., 2025; Zhu et al., 2026b; Zheng et al., 2026). Several recent works further address degradation during few-step or autoregressive distillation. Data-Forcing and DistillAlign mitigate mode collapse by introducing mode-covering supervision (Chen et al., 2026; Li et al., 2026), while Mask Forcing modifies self-rollout coverage to reduce oversaturation and over-smoothing (Zhao et al., 2026b). Related methods also target diversity collapse and long-horizon error accumulation (Duan et al., 2026; Liu et al., 2026). These methods primarily modify DMD itself, leaving bad-mode probability in the pre-DMD initialization unaddressed. In contrast, RAMP does not modify the SF-DMD objective itself. Instead, it preconditions the student's free-running distribution on student-visited states before switching to standard SF-DMD, reducing the initial visitation of undesirable modes.

3 Method

3.1 Motivation

We study the progressive oversaturation that emerges when autoregressive video generators are optimized with SF-DMD, as low-probability undesirable color modes become increasingly prevalent in the student's free-running distribution. We call this phenomenon *bad-mode collapse* and analyze two contributing mechanisms: imperfect initialization under mode-seeking optimization and inaccurate online fake-score estimation. These mechanisms motivate suppressing bad modes before SF-DMD begins. We further show that consistency supervision can favor the good mode in the transition region, while student rollout exposes states to this corrective signal.

Bad-mode collapse from imperfect initialization. We first analyze the effect of student initialization. Consider a teacher and a student distribution consisting of a good mode and a bad mode:

$$p(x) = \pi_{\mathrm{G}} p_{\mathrm{G}}(x) + \pi_{\mathrm{B}} p_{\mathrm{B}}(x), \qquad q(x) = w_{\mathrm{G}} q_{\mathrm{G}}(x) + w_{\mathrm{B}} q_{\mathrm{B}}(x). \tag{1}$$

When the two modes are sufficiently separated, the reverse KL objective can be approximated as

$$D_{\mathrm{KL}}(q\|p) \approx w_{\mathrm{G}} \left(D_{\mathrm{G}} + \log \frac{w_{\mathrm{G}}}{\pi_{\mathrm{G}}} \right) + w_{\mathrm{B}} \left(D_{\mathrm{B}} + \log \frac{w_{\mathrm{B}}}{\pi_{\mathrm{B}}} \right), \tag{2}$$

where $D_k = D_{\mathrm{KL}}(q_k \| p_k)$ measures the fitting error within mode k. Optimizing only the mixture weights gives

$$\frac{w_{\mathrm{B}}^*}{w_{\mathrm{G}}^*} = \frac{\pi_{\mathrm{B}}}{\pi_{\mathrm{G}}} \exp\left(D_{\mathrm{G}} - D_{\mathrm{B}}\right). \tag{3}$$

This shows that even a low-probability bad mode can gain weight when it is easier to fit than the good mode. In practical video generation, high-quality samples span diverse appearances, motions, and temporal dynamics, while the observed bad modes are concentrated around a narrower set of patterns. The good mode is therefore harder to cover, whereas the bad mode is easier to fit, making $D_{\mathrm{G}} > D_{\mathrm{B}}$ more likely. This gap can grow further in high dimensions as fitting errors accumulate across dimensions. We evaluate this prediction in Sec. 4.3.

Bad-mode collapse from inaccurate score estimation. The analysis above assumes access to the exact student score. In practice, SF-DMD relies on an online fake-score model that must continuously track the evolving student distribution. Limited model capacity and optimization budget can make this tracking inaccurate, thereby distorting the distribution-matching update. Moreover, the estimation error can differ across modes, leading to imbalanced update strengths between the good and bad regions. As a result, bad-mode probability may increase even when the student is initially well matched to the teacher. We examine this effect in a controlled two-mode experiment in Sec. 4.3.

Good-mode correction under consistency supervision. These two failure mechanisms make the initial student distribution particularly consequential for SF-DMD. Once optimization begins, mode-seeking updates may amplify bad-mode probability already present in the initialization, while inaccurate score estimates can further bias the student toward the same region. *A natural strategy is therefore to suppress bad modes before SF-DMD begins, rather than relying on SF-DMD to remove them.*

Unlike reverse-KL DMD, consistency distillation directly supervises each queried state with a local teacher target. To see when consistency supervision favors the good mode over the bad mode, we examine the local teacher field between the two modes. For a noised state x_t between the two mode centers, the teacher target depends on the relative posterior responsibility of the two components. At diffusion timestep t, the ratio between the posterior responsibilities of the good and bad modes is

$$\frac{\gamma_{\mathrm{G}}(x_t)}{\gamma_{\mathrm{B}}(x_t)} = \frac{\pi_{\mathrm{G}} \mathcal{N}(x_t; \mu_{\mathrm{G},t}, V_{\mathrm{G},t})}{\pi_{\mathrm{B}} \mathcal{N}(x_t; \mu_{\mathrm{B},t}, V_{\mathrm{B},t})}. \tag{4}$$

Here, π_k is the mixture weight of mode $k \in \{\mathrm{G}, \mathrm{B}\}$, while $\mu_{k,t}$ and $V_{k,t}$ denote its mean and covariance after noising. Moreover, $\gamma_k(x_t) = p(k \mid x_t)$ denotes the posterior responsibility of mode k for the observed state x_t.

At sufficiently noisy timesteps, diffusion noise dominates the intrinsic variance difference between the two components, such that $V_{\mathrm{G},t} \approx V_{\mathrm{B},t} \approx V_t$. Near the center of the transition region, the distances to the two mode centers are also comparable, so their projected teacher displacements are locally similar. Under this local approximation, we analyze the resulting boundary in one dimension the general condition and multivariate extension to Appendix A.2. The boundary $\gamma_{\mathrm{G}}(x_t^\dagger) = \gamma_{\mathrm{B}}(x_t^\dagger)$ is then

$$x_t^\dagger = \frac{\mu_{\mathrm{G},t} + \mu_{\mathrm{B},t}}{2} + \frac{V_t}{\mu_{\mathrm{B},t} - \mu_{\mathrm{G},t}} \log \frac{\pi_{\mathrm{G}}}{\pi_{\mathrm{B}}}. \tag{5}$$

Since $\pi_{\mathrm{G}} > \pi_{\mathrm{B}}$, this boundary shifts toward the bad-mode mean. The good mode consequently dominates a larger portion of the transition region, allowing consistency supervision to redirect states after only a limited displacement from the bad-mode trajectory.

Exposing corrective states through student rollout. The correction described above is useful only when the preconditioning stage queries states in the good-dominated transition region. TF-CD constructs its queries around teacher trajectories, where the local teacher target generally preserves the originating mode. States sampled near the bad mode are therefore supervised toward the same bad-mode branch, which may remain in the initialization presented to SF-DMD.

Self-forcing consistency distillation (SF-CD) addresses this limitation by changing which states receive consistency supervision. Its free-running rollout both exposes states displaced from the bad-mode trajectory and applies consistency supervision on the student's own causal states. Once such a

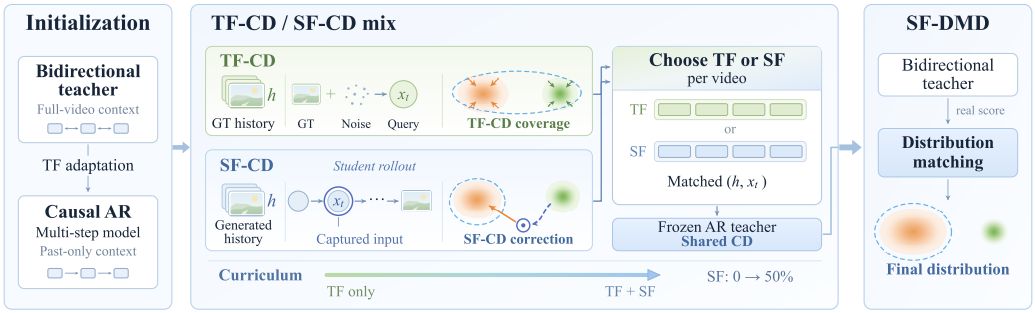

Figure 3: Overview of RAMP.

state enters the good-dominated region, the teacher target pulls it toward the good-mode trajectory. The resulting correction also changes the context seen by subsequent blocks, progressively reducing bad-mode visitation before SF-DMD begins.

3.2 RAMP

We introduce RAMP, short for *Rollout-Aware Mode Preconditioning*, a training strategy that reshapes the student's distribution through rollout-aware causal distillation before transitioning to SF-DMD. As illustrated in Fig. 3, RAMP uses states visited by the autoregressive student during rollout, instead of ground-truth samples. This allows SF-CD to correct states that drift away from narrow bad modes and guide them toward the dominant good mode.

Teacher-forced and rollout-aware consistency preconditioning. We divide a video into causal blocks and generate them autoregressively. For block i, let h_i^{TF} and h_i^{SF} denote the causal KV caches constructed from the ground-truth prefix $x_{<i}$ and the generated prefix $\hat{x}_{<i}$, respectively. We similarly denote the corresponding query states by $x_t^{i,\mathrm{TF}}$ and $x_t^{i,\mathrm{SF}}$.

Both TF-CD and SF-CD use the same consistency target. Given a causal context h and a query state x_t, the frozen causal teacher advances x_t from time t to an earlier time $s < t$:

$$\tilde{x}_s = \mathrm{sg}(x_t) + (s-t)\, v_{\mathrm{T}}(\mathrm{sg}(x_t), t \mid h, c)\,, \tag{6}$$

where c is the text condition, $\mathrm{sg}(\cdot)$ denotes stop-gradient, and v_{T} is the frozen teacher velocity field. Let $G_\theta(x, t \mid h, c)$ denote the clean endpoint predicted by the online student and $G_{\bar{\theta}}$ its exponential-moving-average copy. The consistency loss for a generic pair (h, x_t) is

$$\ell_{\mathrm{CD}}(h, x_t) = \|G_\theta(\mathrm{sg}(x_t), t \mid h, c) - \mathrm{sg}[G_{\bar{\theta}}(\tilde{x}_s, s \mid h, c)]\|_2^2\,. \tag{7}$$

TF-CD constructs both the causal context and query state from the teacher-forced trajectory:

$$\mathcal{L}_{\text{TF-CD}} = \mathbb{E}_{i,t,s}\left[\ell_{\mathrm{CD}}\left(h_i^{\mathrm{TF}}, x_t^{i,\mathrm{TF}}\right)\right]. \tag{8}$$

In contrast, SF-CD uses the student's own autoregressive rollout. Conditioned on h_i^{SF}, the student follows a few-step denoising trajectory $\{x_{t_k}^{i,\mathrm{SF}}\}_{k=0}^{K}$ with clean endpoint $x_0^{i,\mathrm{SF}} = \hat{x}_i$, from which we sample the query state $x_t^{i,\mathrm{SF}}$. Its objective is

$$\mathcal{L}_{\text{SF-CD}} = \mathbb{E}_{i,t,s}\left[\ell_{\mathrm{CD}}\left(h_i^{\mathrm{SF}}, x_t^{i,\mathrm{SF}}\right)\right]. \tag{9}$$

Because h_i^{SF} is constructed from the student's previously generated blocks and $x_t^{i,\mathrm{SF}}$ is sampled from its current denoising trajectory, SF-CD applies consistency supervision directly to states encountered under the student's free-running distribution.

Coherent teacher-to-self-forcing curriculum. Directly applying self-forcing training to an insufficiently initialized student can make optimization unstable. We therefore gradually shift training from TF-CD to SF-CD:

$$\mathcal{L}_{\mathrm{pre}} = (1-\rho)\mathcal{L}_{\text{TF-CD}} + \rho\mathcal{L}_{\text{SF-CD}}, \tag{10}$$

where ρ is gradually increased from zero to a target value. The two branches are sampled at the video level. TF-CD uses the coherent teacher-forced pair $(h_i^{\mathrm{TF}}, x_t^{i,\mathrm{TF}})$, whereas SF-CD uses the coherent rollout pair $(h_i^{\mathrm{SF}}, x_t^{i,\mathrm{SF}})$. We do not construct crossed pairs between generated histories and unrelated ground-truth current states, since such pairs do not correspond to coherent causal trajectories. This curriculum preserves stable teacher-forced supervision early in training while progressively exposing the student to its own rollout distribution, thereby reducing persistent visitation of narrow bad modes before SF-DMD.

Transition to distribution matching. After preconditioning, we disable consistency distillation and use the resulting checkpoint to initialize standard SF-DMD. The subsequent stage is optimized solely with the SF-DMD objective using student-generated rollouts and the online fake-score model. Therefore, RAMP does not modify the SF-DMD update itself. Instead, it provides SF-DMD with an initialization whose free-running distribution assigns less probability to unstable bad modes, reducing their subsequent amplification during distribution matching.

4 Results and Discussion

4.1 Experimental Setup

Models and Baselines. We conduct all experiments using Wan2.1-T2V-1.3B (Wan et al., 2025), which generates 81-frame videos at a resolution of 832 × 480, as the student backbone. Our model adopts a frame-wise causal architecture and generates video autoregressively one latent at a time. We compare RAMP with five representative autoregressive video distillation methods, including CausVid (Yin et al., 2025), Self-Forcing (Huang et al., 2025), Causal Forcing (Zhu et al., 2026b), Causal Forcing++ (Zhao et al., 2026a) and DistillAlign (Li et al., 2026). We use the official released checkpoints and method-specific sampling schedules for the baselines. All methods are evaluated using the same prompts, video resolution and number of frames whenever applicable.

Metrics and Evaluation Datasets. We evaluate all methods on VBench (Huang et al., 2024) using 30 randomly sampled prompts from each of its 16 dimensions, with the same prompt subset used for all methods. We report the overall Quality, Semantic, and Total scores. To measure oversaturation, we compute the mean HSV saturation over all pixels and frames. We additionally evaluate perceptual color fidelity with CFM-7B (Fang et al., 2026), which is designed to reflect human judgments of color realism. For each video, the CFM-7B score is averaged over 16 uniformly sampled frames.

Implementation. In the main experiments, we build upon the three-stage pipeline of Causal Forcing++. We start from the official multi-step autoregressive initialization released by Causal Forcing++ and perform 5,500 steps of RAMP preconditioning using a curriculum that gradually shifts from TF-CD to SF-CD. We then disable the consistency objective and train the resulting checkpoint using pure SF-DMD for an additional 1000 steps. All training experiments are conducted on 8 NVIDIA H200 GPUs.

4.2 Main Results

Table 1 compares RAMP with the bidirectional model and representative autoregressive video-distillation methods. RAMP achieves the best VBench Quality, Semantic, and Total scores of 0.8596, 0.7184, and 0.8314, respectively. RAMP also obtains the lowest mean HSV saturation of 0.4220 while maintaining a competitive CFM-7B score of 11.02. Together, these results indicate that RAMP reduces oversaturation without sacrificing perceptual color fidelity, while improving overall video quality and semantic alignment. Importantly, RAMP uses a comparable number of training steps to prior distillation methods and is trained on only 8 GPUs, while several existing methods report substantially larger training setups. Despite this relatively modest training configuration, RAMP maintains strong quantitative performance with substantially reduced oversaturation, highlighting the stability of rollout-aware preconditioning during subsequent SF-DMD optimization. The qualitative comparisons in Fig. 4 further support this trend, showing that RAMP maintains a more natural color balance and avoids the pronounced oversaturation observed in existing methods.

Table 1: Main quantitative comparison of generation quality, color fidelity, and first-frame latency.

Method	VBench			Color metrics		Latency (s) ↓
	Quality ↑	Semantic ↑	Total ↑	CFM-7B ↑	HSV ↓	
Wan2.1-T2V-1.3B	0.8438	0.6911	0.8132	9.89 ± 3.41	0.4397 ± 0.1773	–
CausVid	0.8415	0.7054	0.8143	5.74 ± 2.96	0.5132 ± 0.1699	0.35
Self-Forcing	0.8546	0.7077	0.8252	**11.50** ± 1.33	0.4553 ± 0.1515	0.35
Causal Forcing	0.8554	0.7109	0.8265	8.95 ± 2.48	0.4697 ± 0.1558	0.35
Causal Forcing++ (2-step)	0.8431	0.6606	0.8066	10.43 ± 1.87	0.4706 ± 0.1534	0.18
DistillAlign	0.8511	0.6765	0.8162	11.01 ± 1.49	0.4258 ± 0.1396	0.35
RAMP (2-step)	**0.8596**	**0.7184**	**0.8314**	11.02 ± 1.55	**0.4220** ± 0.1485	0.18



Figure 4: **Qualitative comparison of color fidelity.** Each row shows videos generated by a different method for two representative prompts, with three frames sampled from each video. Existing methods often amplify red tones and warm lighting, whereas RAMP preserves a more natural color balance across frames.

Beyond these qualitative comparisons, Fig. 5 examines how pre-DMD initialization affects subsequent SF-DMD optimization. Using the official released checkpoints, Fig. 5(a) shows that RAMP provides a more favorable high-CFM initialization. After SF-DMD, both methods shift toward lower-CFM regions, consistent with the mode-seeking tendency of DMD to favor narrower, easier-to-fit regions. Fig. 5(b) further compares optimization stability under a matched setup, where we retrain Causal-Forcing++ using the same number of GPUs and training steps as RAMP. Causal-Forcing++ exhibits a continued CFM-7B decline, whereas RAMP largely stabilizes after an initial drop, consistent with our mechanism that rollout-aware preconditioning reduces subsequent bad-mode amplification.

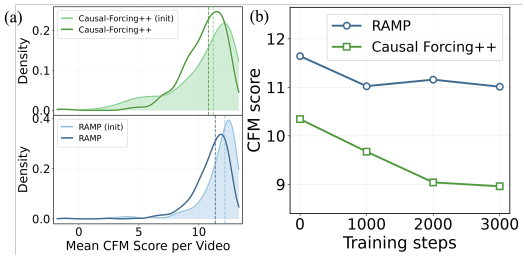

Figure 5: **Effect of pre-DMD initialization on SF-DMD optimization.** (a) CFM-7B distributions before and after SF-DMD, evaluated using the official released checkpoints. (b) CFM-7B dynamics under a matched SF-DMD setup.

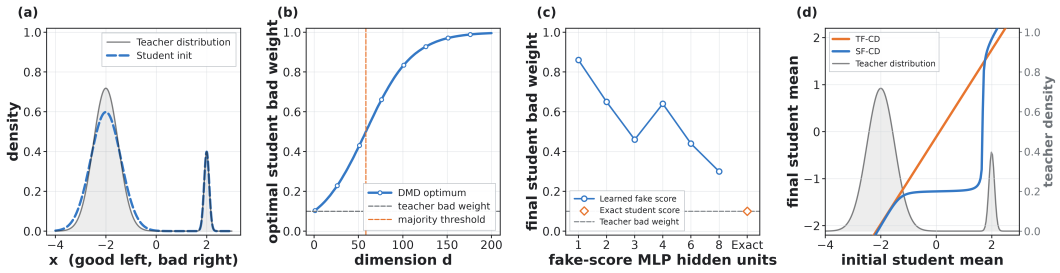

Figure 6: **Controlled Gaussian-mixture experiments.** The teacher contains a 90% broad good mode and a 10% narrow bad mode. (a) The student has a mild good-mode width mismatch. (b) The resulting bad-mode preference strengthens with dimension. (c) Learned fake scores increase the bad-mode weight from a matched initialization, whereas the exact score preserves 0.1. (d) SF-CD enlarges the good-mode attraction region across initial means, while TF-CD largely preserves the initial mean.

4.3 Controlled Mechanism Validation

We use a two-component Gaussian mixture to study the two mechanisms analyzed above and the corrective effect of rollout-aware consistency. All four panels in Fig. 6 use the same teacher distribution,

$$p_d(x) = 0.9\,\mathcal{N}(-2\mathbf{1}_d, 0.5^2 I_d) + 0.1\,\mathcal{N}(2\mathbf{1}_d, 0.1^2 I_d), \tag{11}$$

where the good mode has greater probability and broader support, while the bad mode is rare and narrow. We use $d = 1$ in panels (a), (c), and (d), and vary d in panel (b).

Imperfect student initialization. Fig. 6(a) shows the one-dimensional setup. The student shares the teacher's mode centers and mixture weights, with only a mild mismatch in the good-mode width (0.6 versus 0.5), while the bad mode is matched exactly. Despite this seemingly small discrepancy, its effect grows rapidly with dimension. As shown in Fig. 6(b), the reverse-KL optimum progressively shifts away from the teacher bad-mode weight of 0.1. The bad mode becomes dominant at around $d = 58$, and its optimal weight approaches 1 in high dimensions. The full-mixture Monte Carlo results closely follow the separated-mode prediction in Eq. 3. This demonstrates that even a mild imperfection in student initialization can be amplified by reverse-KL optimization into collapse toward the narrow bad mode.

Inaccurate fake-score estimation. Fig. 6(c) removes the initialization mismatch by starting from a student matched to the teacher. We approximate the student score using online MLP fake-score models with different capacities, each trained by denoising score matching on samples from the noised student distribution. The fake-score model is pretrained for $5{,}000$ steps and receives five refresh steps after each student update. During SF-DMD, the component means and variances remain fixed, and only the mixture weight is optimized. With the exact student score, the matched initialization remains at the teacher bad-mode weight of 0.1. In contrast, learned fake scores shift the student toward the bad mode, with the effect generally decreasing as fake-score capacity increases. This result suggests that fake-score estimation error can itself introduce a bias toward the bad mode, even from a matched student initialization.

Rollout-aware consistency correction. Fig. 6(d) evaluates the CD objectives across a sweep of initial student means. The left axis reports the final student mean after TF-CD or SF-CD, while the gray curve on the right axis shows the teacher density. TF-CD leaves the final student mean close to its initial position and shows no systematic shift of bad-side initializations toward the good basin. SF-CD instead sends most transition states toward the broad, high-probability good mode; convergence to the bad mode occurs only for initializations close to its narrow peak. The toy result shows that once a state moves outside the narrow bad-mode basin, SF-CD tends to pull it toward the broader good mode. In the autoregressive setting, such states arise naturally from rollout errors, so the correction is applied to trajectories that the student actually visits.

4.4 Ablation Studies

Initialization–LoRA swapping. We further examine how the pre-DMD initialization affects the subsequent SF-DMD update. As illustrated in Fig. 7, Init_1 and Init_2 are obtained from TF-CD and RAMP, respectively, and SF-DMD trained from them produces LoRA_1 and LoRA_2. We then swap the two LoRA updates across the initializations and evaluate all four combinations. This ablation uses an independent LoRA-based SF-DMD run, so its absolute CFM values are not directly comparable to those in Table 1.

The swapping results in Table 2 show that the benefit of RAMP is not simply inherited from a better pre-DMD checkpoint. On Init_1, replacing LoRA_1 with LoRA_2 improves CFM from 10.36 to 10.85, indicating that SF-DMD trained from the RAMP initialization learns a more favorable update. Conversely, replacing LoRA_2 with LoRA_1 on Init_2 lowers CFM from 11.36 to 10.21, showing that the better initialization alone does not explain the final gain. In other words, the benefit of RAMP is not simply that SF-DMD starts from a better checkpoint and therefore degrades more slowly. Rather, the improved initialization leads SF-DMD to learn a more favorable update.

Table 2: Initialization–update ablation.

Preconditioning	Stage	CFM↑
TF-CD	LoRA_1	10.36
TF-CD	LoRA_2	10.85
RAMP	LoRA_1	10.21
RAMP	LoRA_2	**11.36**

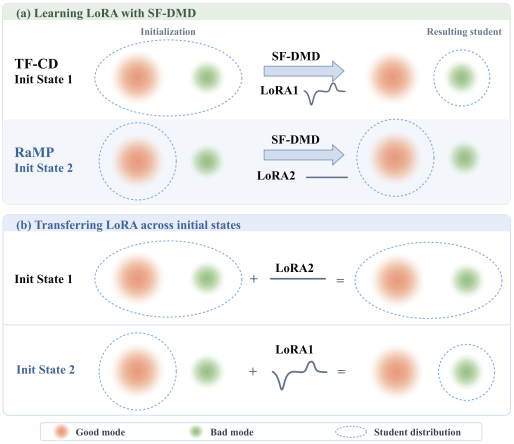

Figure 7: Disentangling initialization and SF-DMD updates.

Ablation of rollout-aware preconditioning. We compare the complete RAMP preconditioning stage with two alternatives while keeping the subsequent SF-DMD stage unchanged. Pure SF-CD trains directly on student trajectories but removes the gradual transition from stable teacher-forced supervision. Diffusion-Forcing consistency distillation (DF-CD) perturbs ground-truth trajectories and can therefore generate states outside the original mode centers, but these synthetic states need not match the structured errors or causal histories encountered during autoregressive rollout.

Table 3: Preconditioning ablation.

Preconditioning	Stage	CFM↑
Pure SF-CD	SF-DMD	8.96
DF-CD	SF-DMD	9.87
RAMP	SF-DMD	**11.02**

As shown in Table 3, pure SF-CD and DF-CD obtain CFM scores of 8.96 and 9.87, respectively, both below the 11.02 achieved by the complete RAMP pipeline. The lower score of pure SF-CD highlights the importance of the gradual TF-to-SF curriculum, while the gap between DF-CD and RAMP shows that generic trajectory perturbations cannot replace states encountered during actual student rollouts. Together, these results show that both the TF-to-SF curriculum and rollout-based training are important for effective preconditioning.

5 Conclusion & Limitations

We investigate bad-mode collapse in few-step autoregressive video generators trained with SF-DMD and identify two contributing mechanisms: imperfect initialization can bias the mode-seeking objective toward rare, easier-to-fit modes, while inaccurate fake-score estimation can further distort optimization. We introduce RAMP, which uses rollout-aware consistency distillation to reduce bad-mode probability before SF-DMD. Controlled experiments validate both mechanisms and show that the pre-DMD distribution strongly affects the final optimization basin. On autoregressive video generation, RAMP improves video quality and color statistics using only eight GPUs. Nevertheless, our analysis uses controlled abstractions of complex video distributions, and the relative contributions of these mechanisms in full-scale video models remain difficult to isolate. Validating them across larger architectures and broader settings remains future work.

6 GENERATIVE AI USAGE STATEMENT

In this work, we did not use generative AI tools for tasks requiring disclosure under the ICLR 2027 policy. We used generative AI tools to aid and polish writing, including improving the clarity, conciseness, grammar, and presentation of text written by the authors. Additionally, we used generative AI tools to assist with the preparation and refinement of scientific figures, including their layout and visual presentation. All AI-assisted text and figures were reviewed and verified by the authors, and the underlying scientific content, experimental results, and conclusions were produced and evaluated by the authors. We take full responsibility for the final content of this work, including its text, figures, claims, and results.

REFERENCES

Sand. ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong Pan, Jie Wang, Jiexin Sheng, Min Cui, Min Hu, Ming Yan, Shucheng Yin, Siran Zhang, Tingting Liu, Xianping Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yuqiao Li. Magi-1: Autoregressive video generation at scale, 2025. URL https://arxiv.org/abs/2505.13211.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575. IEEE, 2023.

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. *arXiv preprint arXiv:2606.18478*, 2026.

Zixuan Duan, Xunzhi Xiang, Yabo Chen, Xin Zhang, Changhan Liu, Haibin Huang, Chi Zhang, Qi Fan, and Xuelong Li. Uncertainty dmd: Restoring diversity in few-step autoregressive video distillation, 2026. URL https://arxiv.org/abs/2609.11265.

Zhengyao Fang, Zexi Jia, Yijia Zhong, Pengcheng Luo, Jinchao Zhang, Guangming Lu, Jun Yu, and Wenjie Pei. Too vivid to be real? benchmarking and calibrating generative color fidelity, 2026. URL https://arxiv.org/abs/2603.10990.

Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing. *arXiv preprint arXiv:2411.16375*, 2024.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, et al. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2024. doi: 10.1109/CVPR52733.2024.02060.

Team HY-World, Chenjie Cao, Xuhui Zuo, Zhenwei Wang, Yisu Zhang, Junta Wu, Zhenyang Liu, Yuning Gong, Yang Liu, Bo Yuan, et al. Hy-world 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds. *arXiv preprint arXiv:2604.14268*, 2026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Jiaxing Li, Kai Zou, Cindy Zhou, Kaichen Huang, Junyao Gao, Zile Wang, Yang Liu, Bin Liu, Bo An, and Yangguang Li. Distillalign: Coordinating mode covering and mode seeking in autoregressive video distillation, 2026. URL https://arxiv.org/abs/2607.26811.

Jinxiu Liu, Xuanming Liu, Kangfu Mei, Yandong Wen, Ming-Hsuan Yang, and Weiyang Liu. Streaming autoregressive video generation via diagonal distillation, 2026. URL https://arxiv.org/abs/2603.09488.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

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

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Ruicong Liu, Xiaojie Xu, Xuangeng Chu, Zhen Li, Zhengyuan Lin, Zhixiang Wang, Zian Meng, and Zihui Gao. Alayaworld: Long-horizon and playable video world generation, 2026a. URL https://arxiv.org/abs/2607.06291.

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Mingliang Zhai, Ruicong Liu, Xiaojie Xu, Xuangeng Chu, Zhen Li, Zhengyuan Lin, Zhixiang Wang, Zian Meng, and Zihui Gao. Alayaworld: Interactive long-horizon world modeling – full technical report, 2026b. URL https://arxiv.org/abs/2607.18367.

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.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

Bin Wu, Mengqi Huang, Shaojin Wu, Weinan Jia, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Stream-r1: Reliability-perplexity aware reward distillation for streaming video generation, 2026. URL https://arxiv.org/abs/2605.03849.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024.

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. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Yuyang You, Yongzhi Li, Jiahui Li, Yadong Mu, Quan Chen, and Peng Jiang. Adaptive video distillation: Mitigating oversaturation and temporal collapse in few-step generation, 2026. URL https://arxiv.org/abs/2603.21864.

Jie Zhang, Xiaoyue Chen, Anzhe Chen, Dayiheng Liu, Deqing Li, Gengze Zhou, Hale Yin, Haoqi Yuan, Haoyang Li, Jiahao Li, et al. Qwen-robotworld technical report: Unifying embodied world modeling through language-conditioned video generation. *arXiv preprint arXiv:2606.17030*, 2026.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation, 2026a. URL https://arxiv.org/abs/2605.15141.

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. *arXiv preprint arXiv:2609.09123*, 2026b.

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. *arXiv preprint arXiv:2606.25473*, 2026.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. SANA-WM: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026a.

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. *arXiv preprint arXiv:2602.02214*, 2026b.

A Detailed Theoretical Analysis

This appendix provides the derivations underlying the two collapse mechanisms and the rollout-aware correction discussed in Sec. 3.1. Throughout the analysis, we use the two-component Gaussian-mixture model from our controlled experiments, with $k \in \{\mathrm{G}, \mathrm{B}\}$ denoting the good and bad modes. The teacher and student components may have different means and covariances, and the modes are assumed to be well separated. This setting matches the controlled experiments and provides closed-form expressions for the within-mode KL terms and the noised responsibility boundaries used below.

A.1 Bad-Mode Preference under SF-DMD

Mode-wise decomposition of reverse KL. Let the teacher and student distributions be

$$p(x) = \sum_k \pi_k p_k(x), \qquad q(x) = \sum_k w_k q_k(x), \qquad \sum_k \pi_k = \sum_k w_k = 1. \tag{12}$$

Assume first that the two modes have disjoint supports $\mathcal{S}_\mathrm{G}$ and $\mathcal{S}_\mathrm{B}$. On $x \in \mathcal{S}_k$, we have $p(x) = \pi_k p_k(x)$ and $q(x) = w_k q_k(x)$. The reverse KL therefore decomposes exactly as

$$\begin{aligned} D_{\mathrm{KL}}(q\|p) &= \sum_k \int_{\mathcal{S}_k} w_k q_k(x) \log \frac{w_k q_k(x)}{\pi_k p_k(x)} \, \mathrm{d}x \\ &= \sum_k w_k \left[D_{\mathrm{KL}}(q_k\|p_k) + \log \frac{w_k}{\pi_k} \right]. \end{aligned} \tag{13}$$

For separated but slightly overlapping modes, the same expression holds up to terms controlled by the overlap probability. Denote the within-mode error by $D_k = D_{\mathrm{KL}}(q_k\|p_k)$. Let

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\Delta} D_{\mathrm{KL}}(q\|p), \qquad \Delta = \left\{ \mathbf{w} : w_k \geq 0, \ \sum_k w_k = 1 \right\}, \tag{14}$$

denote the optimal student mixture weights. Introducing a Lagrange multiplier λ for the normalization constraint $\sum_k w_k = 1$, the stationarity condition for each component is

$$D_k + \log \frac{w_k}{\pi_k} + 1 + \lambda = 0. \tag{15}$$

Solving Eq. 15 gives $w_k = C\pi_k e^{-D_k}$, where C is a normalization constant shared by all components. Enforcing $\sum_k w_k = 1$ gives

$$w_k^* = \frac{\pi_k e^{-D_k}}{\sum_j \pi_j e^{-D_j}}. \tag{16}$$

Consequently,

$$\frac{w_\mathrm{B}^*}{w_\mathrm{G}^*} = \frac{\pi_\mathrm{B}}{\pi_\mathrm{G}} \exp(D_\mathrm{G} - D_\mathrm{B}). \tag{17}$$

Eq. 17 makes the competition explicit. In this two-mode setting, we define *bad-mode dominance* by $w_\mathrm{B}^* > w_\mathrm{G}^*$, or equivalently $w_\mathrm{B}^* > 1/2$. We use this as an operational criterion for bad-mode collapse. The smaller teacher probability of the bad mode contributes only the fixed ratio $\pi_\mathrm{B}/\pi_\mathrm{G}$, whereas its fitting advantage enters through the exponential factor $\exp(D_\mathrm{G} - D_\mathrm{B})$. Consequently, even a moderate gap in within-mode fitting errors can rapidly offset a substantial prior-mass disadvantage. When the student fits the bad mode substantially better than the good mode, its weight can therefore grow despite $\pi_\mathrm{B} \ll \pi_\mathrm{G}$, making collapse toward the bad mode likely.

Why dimensionality can reverse the mass advantage. For diagonal Gaussian components $p_k = \mathcal{N}(\mu_{p,k}, \mathrm{diag}(\sigma_{p,k}^2))$ and $q_k = \mathcal{N}(\mu_{q,k}, \mathrm{diag}(\sigma_{q,k}^2))$, the within-mode error is

$$D_k = \frac{1}{2} \sum_{j=1}^{d} \left[\frac{\sigma_{q,k,j}^2}{\sigma_{p,k,j}^2} + \frac{(\mu_{q,k,j} - \mu_{p,k,j})^2}{\sigma_{p,k,j}^2} - 1 + \log \frac{\sigma_{p,k,j}^2}{\sigma_{q,k,j}^2} \right]. \tag{18}$$

Both good- and bad-mode errors accumulate over the same ambient dimension. What matters is not a difference in dimensionality, but a difference in their per-coordinate fitting errors. A concentrated bad mode may be nearly matched, whereas small mismatches in a diverse good mode can accumulate across many independent visual or temporal directions.

As a simple isotropic example, suppose the means are matched and the bad component is fitted exactly, so that $D_{\mathrm{B}} = 0$. We define the standard-deviation ratio

$$a = \frac{\sigma_{q,\mathrm{G}}}{\sigma_{p,\mathrm{G}}} > 0, \qquad a \neq 1, \tag{19}$$

where $a < 1$ represents under-coverage and $a > 1$ represents over-dispersion of the good component. In either case,

$$D_{\mathrm{G}} = \frac{d}{2}\left(a^2 - 1 - \log a^2\right). \tag{20}$$

The term $a^2 - 1 - \log a^2$ is nonnegative for every $a > 0$ and equals zero only at $a = 1$. Thus, no absolute value is required: both underestimating and overestimating the good-mode width produce a positive discrepancy. Combining Eq. 17 with Eq. 20, and assuming $\pi_{\mathrm{G}} > \pi_{\mathrm{B}}$, the optimal bad-mode weight exceeds the good-mode weight when

$$d > \frac{2\log(\pi_{\mathrm{G}}/\pi_{\mathrm{B}})}{a^2 - 1 - \log a^2}. \tag{21}$$

This threshold explains Fig. 6(a)–(b). Even a modest per-coordinate width mismatch in the good component can outweigh its teacher mass advantage once the discrepancy accumulates over sufficiently many dimensions.

A.2 Why Rollout-Aware Consistency Reduces Bad-Mode Probability

Direct analysis of the SF-CD update. For a rollout-generated history h_i and query state x_t^i, the frozen teacher produces

$$x_s^i = x_t^i + (s-t)v_{\mathrm{T}}(x_t^i, t \mid h_i, c), \tag{22}$$

and the stop-gradient consistency target is

$$y_T^i = \mathrm{sg}\left(G_{\bar{\theta}}(x_s^i, s \mid h_i, c)\right). \tag{23}$$

The per-example SF-CD loss in Eq. 9 is therefore

$$\ell_i = \left\| G_\theta(x_t^i, t \mid h_i, c) - y_T^i \right\|_2^2. \tag{24}$$

Its negative gradient with respect to the student output is

$$-\nabla_{G_\theta}\ell_i = 2\left(y_T^i - G_\theta(x_t^i, t \mid h_i, c)\right), \tag{25}$$

The key question is which teacher mode governs y_T^i at a sampled rollout state. Suppressing (h_i, c) for notation, write the noised teacher distribution as

$$p_t(x_t) = \sum_{k\in\{\mathrm{G},\mathrm{B}\}} \pi_k p_{k,t}(x_t). \tag{26}$$

For an optimal conditional flow-matching teacher, the velocity is a posterior mixture of the component velocities,

$$v_{\mathrm{T}}(x_t, t) = \sum_k \gamma_k(x_t) v_k(x_t, t), \qquad \gamma_k(x_t) = \frac{\pi_k p_{k,t}(x_t)}{\sum_j \pi_j p_{j,t}(x_t)}. \tag{27}$$

The teacher first advances the query state according to $x_s - x_t = (s-t)v_{\mathrm{T}}(x_t, t)$. Substituting Eq. 27 into this step and defining the component-wise displacement $d_k(x_t, t) = (s-t)v_k(x_t, t)$ gives

$$x_s - x_t = \gamma_{\mathrm{G}}(x_t) d_{\mathrm{G}}(x_t, t) + \gamma_{\mathrm{B}}(x_t) d_{\mathrm{B}}(x_t, t). \tag{28}$$

Eq. 28 connects the teacher mixture to the actual CD target: it determines which component moves the teacher-advanced state x_s more strongly before the EMA endpoint is evaluated.

Let u be the unit vector from the good-mode mean to the bad-mode mean. In the transition region, the two component-wise displacements point in opposite directions along this axis. Define their positive projected magnitudes by

$$\rho_{\mathrm{G}} = -u^{\top} d_{\mathrm{G}}(x_t, t), \qquad \rho_{\mathrm{B}} = u^{\top} d_{\mathrm{B}}(x_t, t). \tag{29}$$

Projecting Eq. 28 onto u gives

$$u^{\top}(x_s - x_t) = -\gamma_{\mathrm{G}}(x_t)\rho_{\mathrm{G}} + \gamma_{\mathrm{B}}(x_t)\rho_{\mathrm{B}}. \tag{30}$$

Thus, without assuming equal displacement magnitudes, the teacher step points toward the good mode when $\gamma_{\mathrm{G}}\rho_{\mathrm{G}} > \gamma_{\mathrm{B}}\rho_{\mathrm{B}}$, and the exact directional switch satisfies

$$\gamma_{\mathrm{G}}(x_t)\rho_{\mathrm{G}} = \gamma_{\mathrm{B}}(x_t)\rho_{\mathrm{B}}. \tag{31}$$

For the equal-noised-covariance Gaussian setting considered below, the two component velocities share the same time-dependent scale and point from x_t toward their respective means. Their projected magnitudes are therefore comparable near the central transition region. Under the local approximation $\rho_{\mathrm{G}} \approx \rho_{\mathrm{B}}$, the directional switch reduces to

$$\gamma_{\mathrm{G}}(x_t) = \gamma_{\mathrm{B}}(x_t). \tag{32}$$

To express the approximate responsibility boundary in Eq. 32 explicitly in state space, consider the time-t densities of the two teacher components. For $p_{k,0} = \mathcal{N}(\mu_k, \Sigma_k)$, the forward process $x_t = a_t x_0 + b_t \epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$, gives

$$p_{k,t} = \mathcal{N}(\mu_{k,t}, V_{k,t}), \qquad \mu_{k,t} = a_t \mu_k, \qquad V_{k,t} = a_t^2 \Sigma_k + b_t^2 I, \tag{33}$$

where a_t and b_t are the diffusion-schedule coefficients. We consider timesteps at which the shared noise covariance $b_t^2 I$ dominates the difference between the clean component covariances, yielding $V_{\mathrm{G},t} \approx V_{\mathrm{B},t} \approx V_t$. Substituting these Gaussian densities into the posterior responsibilities and imposing $\gamma_{\mathrm{G}}(x_t) = \gamma_{\mathrm{B}}(x_t)$ yields the following hyperplane in the noised state space

$$\mathcal{H}_t = \left\{ x_t \mid (\mu_{\mathrm{B},t} - \mu_{\mathrm{G},t})^{\top} V_t^{-1} x_t = \frac{1}{2}\left(\mu_{\mathrm{B},t}^{\top} V_t^{-1} \mu_{\mathrm{B},t} - \mu_{\mathrm{G},t}^{\top} V_t^{-1} \mu_{\mathrm{G},t}\right) + \log \frac{\pi_{\mathrm{G}}}{\pi_{\mathrm{B}}} \right\}. \tag{34}$$

Eq. 34 is the full equal-responsibility boundary in the d-dimensional noised state space. The scalar boundary used in the main text describes where this hyperplane intersects the line joining the two mode centers. For the isotropic case $V_t = \sigma_t^2 I$, let $u_t = (\mu_{\mathrm{B},t} - \mu_{\mathrm{G},t})/\|\mu_{\mathrm{B},t} - \mu_{\mathrm{G},t}\|$ denote the direction of this line and use $z_t = u_t^{\top} x_t$ as its scalar coordinate. Substituting this parameterization into Eq. 34 gives the intersection

$$z_t^{\dagger} = u_t^{\top} \frac{\mu_{\mathrm{G},t} + \mu_{\mathrm{B},t}}{2} + \frac{\sigma_t^2}{\|\mu_{\mathrm{B},t} - \mu_{\mathrm{G},t}\|} \log \frac{\pi_{\mathrm{G}}}{\pi_{\mathrm{B}}}.$$

When $\pi_{\mathrm{G}} > \pi_{\mathrm{B}}$, the boundary shifts toward the bad-mode mean, so the good component governs a larger portion of the transition region. Under the comparable-displacement approximation, the teacher step in this region points toward the good mode. If the EMA target preserves this local direction, SF-CD corrects sampled rollout states toward the good mode and can reduce subsequent bad-mode visitation.

Why the self-rollout sampling measure matters. The analysis above determines the correction direction at a given state. The key difference between TF-CD and SF-CD is therefore not the local consistency target itself, but which states are actually sampled. Let $\mathcal{R}_t$ denote rollout-induced transition states for which the teacher target is good-directed, and let ν_t^{TF} and ν_t^{SF} denote the joint distributions over histories and query states used by TF-CD and SF-CD, respectively. TF-CD samples from teacher-forced trajectories and therefore rarely visits the off-trajectory states created by accumulated student rollout errors. These states arise naturally under self-rollout, so in the rollout-mismatch regime,

$$\nu_t^{\mathrm{SF}}(\mathcal{R}_t) \gg \nu_t^{\mathrm{TF}}(\mathcal{R}_t). \tag{35}$$

Thus, although TF-CD and SF-CD would apply similar good-directed corrections if evaluated at the same recovery state, SF-CD places substantially more training weight on the states actually induced by the student's own autoregressive errors. Diffusion forcing can also broaden the training-state distribution, but its perturbations are prescribed rather than generated by the student's free-running dynamics and therefore need not match these structured, history-dependent rollout errors. By repeatedly correcting the student on such rollout-specific transition states, SF-CD strengthens its ability to recover from deviations toward the bad mode, making persistent bad-mode trajectories less likely and providing a more favorable free-running distribution for the subsequent SF-DMD stage.

B More Experimental Details

Training setup. All stages use the frame-wise causal student and are trained on eight NVIDIA H200 GPUs with a total batch size of eight. Training clips contain 21 latent frames, with one latent frame in each causal block. We use mixed-precision training, gradient checkpointing, and hybrid-sharded FSDP. Unless otherwise specified, the optimizer is AdamW with $(\beta_1, \beta_2) = (0, 0.999)$ and weight decay 0.01. An exponential-moving-average student with decay 0.99 is maintained for evaluation.

RAMP preconditioning. We initialize both the online student and the frozen causal teacher from the official autoregressive checkpoint released by Causal Forcing++ and train the preconditioning stage for $5{,}500$ steps. The student follows the four-step schedule $\{1000, 750, 500, 250\}$ with a timestep shift of 5.0 and guidance scale 3.0. Each training example is drawn entirely from either the TF-CD or SF-CD branch. The probability of selecting SF-CD increases linearly from 0 to 0.5 during the first $2{,}000$ steps and remains at 0.5 thereafter. A TF-CD example uses a ground-truth history and its corresponding current state. An SF-CD example instead uses a detached student-generated prefix together with an intermediate state visited by the same four-step rollout. We uniformly sample one of the four visited denoising states and use the same timestep index for all causal blocks in that example. The teacher, online student, and EMA student are evaluated under the same coherent history.

Pure SF-DMD refinement. After preconditioning, we disable CD and initialize a separate pure SF-DMD stage from the $5{,}500$-step checkpoint. The final student uses the two-step schedule $\{1000, 500\}$, while the first causal block retains the four-step schedule $\{1000, 750, 500, 250\}$. The real-score model is Wan2.1-T2V-14B and the online fake-score model is Wan2.1-T2V-1.3B. The generator and fake-score model use learning rates of 2×10^{-6} and 4×10^{-7}, respectively. The fake-score model is updated every training iteration, and the generator is updated once every five iterations. We train this stage for $1{,}000$ iterations, with no consistency objective active.

C Additional Visualization of RAMP

Fig. 8 presents additional qualitative results across diverse prompts, including scene construction, human motion, and moving-camera generation. The sampled frames preserve coherent subjects and backgrounds while maintaining natural color saturation throughout the autoregressive rollout. These examples complement the quantitative color and video-quality metrics reported in the main paper and show that the improvement is consistent across different visual domains.

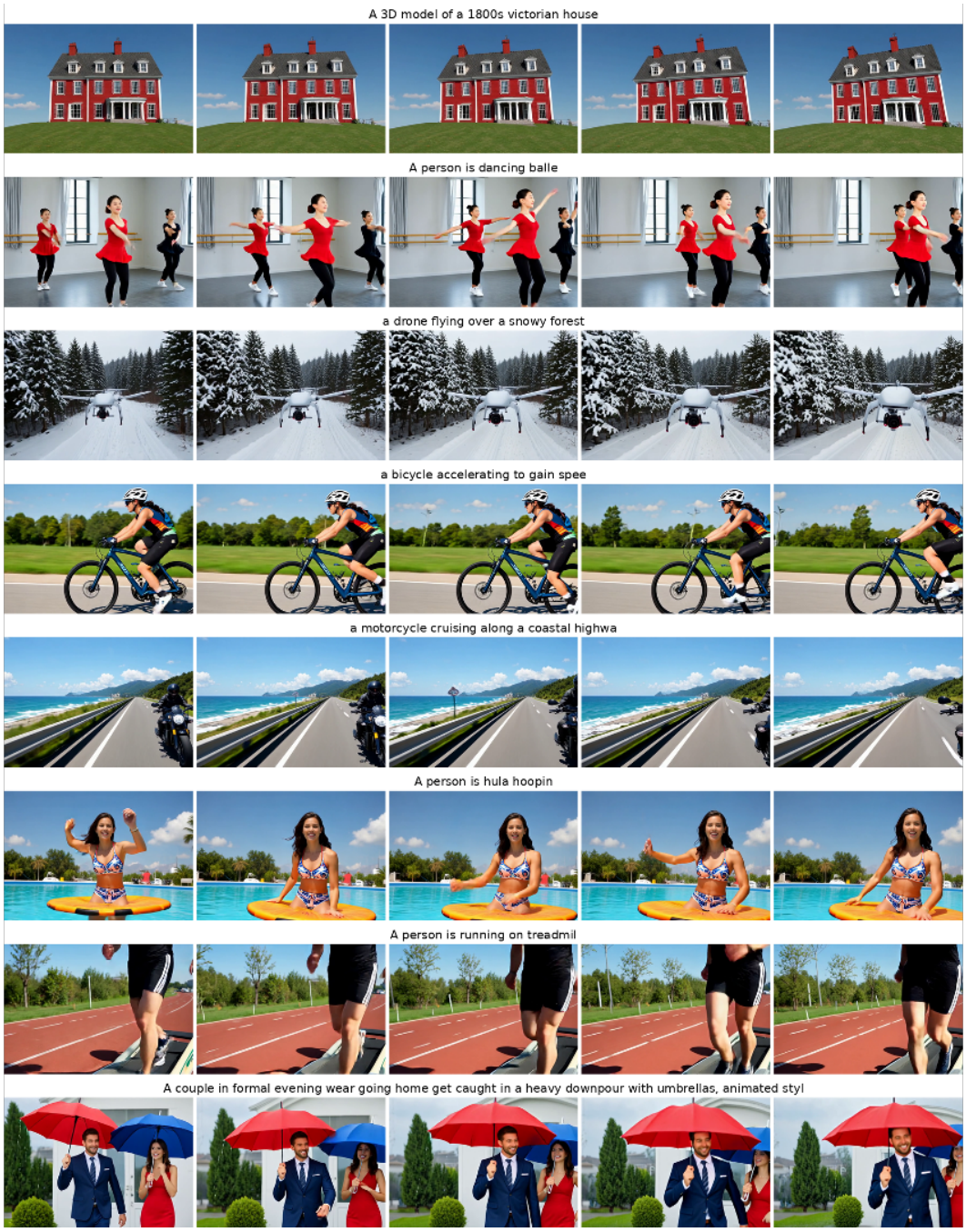

Figure 8: Additional qualitative results of RAMP.