DISTILLALIGN: COORDINATING MODE COVERING AND MODE SEEKING IN AUTOREGRESSIVE VIDEO DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing autoregressive video distillation methods commonly adopt a Distribution Matching Distillation (DMD)-based multi-stage pipeline. However, they typically decouple the initialization and DMD stages—which then pursue different target distributions—and judge the intermediate student mainly by visual scores such as VBench. In this paper, we revisit this design from a distributional perspective. Given the mode-seeking nature of the distribution matching loss, a good initialization should match the mode coverage of the target DMD teacher, rather than merely pursuing high quality. To analyze this, we introduce a distributional evaluation protocol that measures precision and coverage between student and teacher distributions in a shared latent space. It exposes differences hidden by visual scores: some initializations reach high precision but low coverage, leading to suboptimal refinement, while mode-covering ones preserve broader support. Furthermore, even when the target distributions are aligned, DMD's reverse-KL objective can still drive the student toward high-probability teacher regions in late training, reducing coverage and diversity. To address this, we propose joint distillation, which combines DMD's mode-seeking objective with a Consistency Distillation-based mode-covering constraint. Experiments show that our method improves generation quality, coverage, and diversity; notably, even with a Wan-1.3B DMD teacher, it outperforms baselines refined with Wan-14B, underscoring the importance of distributional alignment in autoregressive video distillation. Code, data, and models will be publicly released.

1 INTRODUCTION

Driven by the emerging vision of world models (Qian et al., 2026; He et al., 2025; Wang et al., 2026; Team et al., 2026) and interactive content creation (PixVerse Research, 2026), the demand for high-quality, real-time autoregressive video generation is rapidly increasing. To reduce the sampling cost of video diffusion models, recent methods commonly distill bidirectional video diffusion teachers into few-step causal generators (Huang et al., 2025; Yin et al., 2025a; Zhu et al., 2026; Yang et al., 2025).

Existing methods typically follow a multi-stage pipeline: the student is first initialized by ODE distillation (Huang et al., 2025), or by causal training followed by causal ODE or Consistency Distillation (CD) (Zhu et al., 2026), and is then refined with Distribution Matching Distillation (DMD). In practice, the two stages are treated separately and may pursue different target distributions: the initialization is trained on base-model trajectories (Huang et al., 2025; Zhu et al., 2026) or external video data (Zhao et al., 2026), whereas DMD may use a stronger teacher. The intermediate student is then judged by visual scores such as VBench (Huang et al., 2024), implicitly assuming that a better-scoring initialization yields a better final generator.

We revisit this assumption from a distributional perspective. Trajectory- or consistency-based initialization mainly provides mode coverage, whereas DMD performs mode-seeking refinement around the current student distribution. Therefore, the effectiveness of DMD depends on whether its target modes have already been covered by the initialization distribution. When the target distributions of the two stages are misaligned, DMD may provide unsupported gradients, leading to coverage collapse and suboptimal refinement, as illustrated in Fig. 1.

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.

To verify this hypothesis, we conduct a controlled target-distribution swap on the Causal Forcing pipeline. Tab. 1 shows that final performance is not determined solely by teacher capacity or data quality, but also by whether the initialization and DMD target distributions are aligned: Even weaker initialization data or a weaker DMD teacher can lead to better results when their distributions are matched. This suggests initialization quality should not be judged solely by visual quality, but also by whether it provides matched mode coverage for subsequent DMD refinement.



Figure 1: **Illustration of Hypothesis I.** Mismatched target distributions across stages can lead to suboptimal refinement.

Table 1: **Controlled target-distribution swap.** Built upon the Causal Forcing pipeline, we use samples from the initialization target data for causal model training and consistency distillation, and apply DMD with different teachers.

Init. Target Data	DMD Teacher	Matched	VBench ↑
Wan-14B	Wan-14B	✓	84.74
Wan-1.3B	Wan-1.3B	✓	84.50
Wan-14B	Wan-1.3B	✗	84.16
Wan-1.3B	Wan-14B	✗	83.89

*The last row corresponds to the original Causal Forcing setting.

Based on this finding, we propose a distributional evaluation protocol to assess distribution matching in distillation pipelines. We project teacher and student samples into the same latent feature space, and measure how well the student distribution matches the teacher distribution using precision and coverage. This evaluation reveals differences that are difficult to capture with VBench scores alone: some initializations achieve high visual quality and precision but low coverage, suggesting that they have collapsed to a small set of high-quality modes. In contrast, mode-covering initializations may appear blurrier, but preserve broader distributional support for subsequent DMD refinement.

Furthermore, even when the initialization distribution is aligned with the DMD target distribution, pure DMD can still suffer from late-stage distribution drift (Fig. 2). As supported by the toy experiment in Tab. 3, diversity score keeps decreasing, while the visual score first increases and then drops. This suggests that drifting toward high-probability teacher regions does not necessarily improve visual quality. To address this, we propose joint distillation, which combines DMD's mode-seeking objective with CD's mode-covering constraint to sharpen high-quality modes while preserving broader teacher coverage, thereby suppressing late-stage distribution drift.

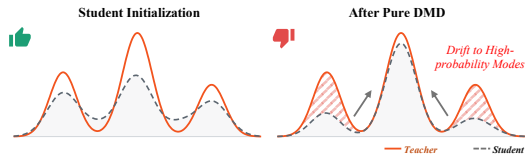


Figure 2: **Illustration of Hypothesis II.** Even with matched initialization, long-term reverse-KL refinement can push the student toward high-probability teacher regions.

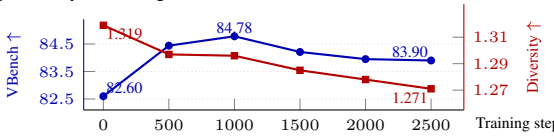


Figure 3: **Late-stage drift in pure DMD.** Using the first setting in Tab. 1, VBench score first increases and then decreases, while diversity consistently drops. Diversity is reported as the raw Vendi score under the protocol in Sec. 4.1.

Extensive experiments show that by aligning target distributions across stages and coordinating mode-covering and mode-seeking constraints, our method can improve final visual quality and the coverage of the student distribution over the teacher distribution. Notably, even with Wan-1.3B as the DMD teacher, our method already outperforms all baselines with Wan-14B teacher, further highlighting the importance of distributional alignment in autoregressive video distillation.

2 Background

2.1 Flow Matching

Flow matching (Lipman et al., 2022; Liu et al., 2022) formulates generation as learning a velocity field between a simple noise distribution and the data distribution. Given a data sample $\mathbf{x}_0 \sim p_{\text{data}}$ and a noise sample $\mathbf{x}_1 \sim p_{\text{noise}}$, it defines a linear interpolation path

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \quad t \in [0,1], \tag{1}$$

whose target velocity is $\mathbf{x}_1 - \mathbf{x}_0$. The model is trained by minimizing

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\left[\|\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) - (\mathbf{x}_1 - \mathbf{x}_0)\|_2^2\right], \tag{2}$$

where $\mathbf{c}$ denotes the conditioning signal. During inference, samples are generated by initializing from noise and integrating the learned ODE from $t = 1$ to $t = 0$.

2.2 ODE AND CONSISTENCY DISTILLATION

ODE distillation (Luhman & Luhman, 2021; Salimans & Ho, 2022) learns a student mapping that directly predicts the teacher ODE endpoint. Let $\Psi_{s\to r}$ denote the flow map of the teacher PF-ODE from time s to r. Given $\mathbf{x}_1 \sim p_{\text{noise}}$ and $\mathbf{x}_t = \Psi_{1\to t}(\mathbf{x}_1)$, ODE distillation minimizes

$$\mathcal{L}_{\text{ODE}}(\theta) = \mathbb{E}_{t,\mathbf{x}_1}\left[\|f_\theta(\mathbf{x}_t, t) - \Psi_{1\to 0}(\mathbf{x}_1)\|_2^2\right]. \tag{3}$$

Consistency distillation (CD) (Song et al., 2023) instead enforces local consistency between adjacent time steps. With a discretization $0 = t_0 < \cdots < t_N = 1$, it uses one teacher ODE step to estimate

$$\hat{\mathbf{x}}_{t_n} = \mathbf{x}_{t_{n+1}} + (t_n - t_{n+1})\,\mathbf{v}_\phi(\mathbf{x}_{t_{n+1}}, t_{n+1}), \tag{4}$$

and trains the student by

$$\mathcal{L}_{\text{CD}}(\theta) = \mathbb{E}_{n,\mathbf{x}_0,\mathbf{x}_1}\left[\left\|f_\theta(\mathbf{x}_{t_{n+1}}, t_{n+1}) - f_{\theta^-}(\hat{\mathbf{x}}_{t_n}, t_n)\right\|_2^2\right], \tag{5}$$

where θ^- is an EMA or stop-gradient copy of θ. At convergence, both ODE distillation and CD recover the teacher flow map, $f_\theta(\mathbf{x}_t, t) = \Psi_{t\to 0}(\mathbf{x}_t)$. As regressions onto teacher endpoints, such methods exhibit *mode-covering* behavior, preserving broad support at the cost of sharpness.

2.3 DISTRIBUTION MATCHING DISTILLATION

Distribution matching distillation (DMD) (Yin et al., 2024b;a) distills a multi-step teacher into a few-step generator G_θ that maps noise directly to samples. Let $p_{\theta,t}$ and $p_{\text{teacher},t}$ denote the perturbed marginals of the generator and teacher distributions at noise level t. DMD minimizes a reverse-KL objective over noise levels:

$$\mathcal{L}_{\text{DMD}}(\theta) = \mathbb{E}_t\left[w(t)\,\text{KL}(p_{\theta,t}\,\|\,p_{\text{teacher},t})\right]. \tag{6}$$

The gradient of Eq. equation 6 can be written as a score-difference update:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = \mathbb{E}_{t,\mathbf{x}_1,\boldsymbol{\epsilon}}\left[w(t)\,\left(\mathbf{s}_{\text{fake}}(\mathbf{x}_t, t) - \mathbf{s}_{\text{teacher}}(\mathbf{x}_t, t)\right)\frac{\partial \mathbf{x}_t}{\partial \theta}\right], \tag{7}$$

where $\mathbf{x}_t = (1-t)\,G_\theta(\mathbf{x}_1) + t\,\boldsymbol{\epsilon}$. Here, $\mathbf{s}_{\text{teacher}}$ is provided by the frozen teacher, while $\mathbf{s}_{\text{fake}}$ is estimated by an online critic trained on generator samples. This reverse-KL objective is inherently *mode-seeking*, concentrating the student on high-density teacher modes at the cost of coverage and diversity.

3 METHODOLOGY

3.1 TEACHER-NORMALIZED DISTRIBUTION EVALUATION

Since DMD mainly refines teacher modes that the initializer already covers, evaluating the initializer itself is crucial. What matters here is not the low-level fidelity of raw initializer outputs but the high-level semantic support an initializer reaches, i.e., which scene layouts, object configurations, and motions it makes accessible for later refinement. Measuring this support is non-trivial. Running the full DMD stage for every initializer and comparing the results is computationally expensive, whereas directly evaluating raw outputs is cheaper but unreliable: initializers differ substantially in sharpness and residual denoising artifacts despite similar semantics, and these low-level differences dominate conventional video metrics and obscure distributional coverage.

To address this issue, we evaluate initializers through a teacher-normalized refinement protocol. Given a prompt c and noise z, the initializer first produces $x_\theta = G_\theta(c, z)$. We then re-noise this sample in the latent space as $x_{\theta,\rho} \sim q_\rho^{T_{\text{norm}}}(\cdot \mid x_\theta)$, where $q_\rho^{T_{\text{norm}}}$ follows the forward noising process of a fixed normalization teacher T_{norm} to noise level ρ. Its solver maps the re-noised sample back to a clean video,

$$\tilde{x}_\theta^{(\rho)} = R_{T_{\text{norm}},\rho}(x_{\theta,\rho}; c), \tag{8}$$

where $R_{T_{\text{norm}},\rho}$ denotes the denoising trajectory from noise level ρ to a clean sample. Since all initializers are evaluated with the same re-noising level, normalization teacher, and denoising schedule, low-level fidelity differences are largely normalized, while the result still reflects the semantic

region reached by the initializer, as shown in Fig. 9. Note that T_{norm} is distinct from the reference teacher that defines the target distribution; see Appendix B for details.

We then compare refined initializer samples with teacher reference samples in a frozen video representation space. Let $\phi(\cdot)$ be a video encoder, and define normalized features

$$u_i = \frac{\phi(\tilde{x}_{\theta,i}^{(\rho)})}{\|\phi(\tilde{x}_{\theta,i}^{(\rho)})\|_2}, \qquad v_j = \frac{\phi(x_{T,j})}{\|\phi(x_{T,j})\|_2}. \tag{9}$$

We estimate teacher support using a k-nearest-neighbor radius. Let $r_k(v_j)$ be the distance from teacher feature v_j to its k-th nearest neighbor among teacher features. Precision measures whether refined samples lie inside this support:

$$\text{Precision} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\exists j \;\text{ s.t. }\; \|u_i - v_j\|_2 \le r_k(v_j)\right]. \tag{10}$$

Coverage measures how much teacher support is reached by the initializer:

$$\text{Coverage} = \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}\left[\min_i \|u_i - v_j\|_2 \le r_k(v_j)\right]. \tag{11}$$

3.2 Mode Alignment Matters

3.2.1 Rethinking the Toy Experiment

In Sec. 1, the controlled teacher-source swap shows that matched initialization and DMD targets lead to better final VBench scores. Using our distributional protocol, we revisit these settings and quantify the effect under the same conditions. Specifically, for each row, we compute coverage after initialization and after DMD, using the corresponding DMD teacher as the target distribution. As shown in Tab. 2, matched settings achieve higher coverage than mismatched ones both before and after DMD. This confirms that the advantage of aligned stages does not merely come from stronger teachers or higher-quality data; instead, matched initialization provides better mode coverage over the DMD target distribution, making subsequent DMD refinement better supported. In addition, we observe that DMD generally reduces coverage, which is consistent with its mode-seeking behavior.

Table 2: **Controlled teacher-source swap.** For each setting in Table 1, we recompute the coverage of both the initialization and DMD-refined results against the corresponding teacher distribution. In each of the three numeric columns, darker shading indicates a higher-ranked value.

Init. Data Source	DMD Teacher	Matched	VBench ↑	Coverage ↑	
				Init.	After DMD
Wan-14B	Wan-14B	✓	84.74	0.648	0.527
Wan-1.3B	Wan-1.3B	✓	84.50	0.516	0.461
Wan-14B	Wan-1.3B	✗	84.16	0.503	0.430
Wan-1.3B	Wan-14B	✗	83.89	0.453	0.435

*The last row corresponds to the original Causal Forcing setting.

3.2.2 Rethinking Distillation Pipes from a Distributional Perspective

When the data source and target distribution are already aligned, how do different initialization methods affect coverage and subsequent DMD refinement?

As shown in Fig. 5, we construct five distillation pipelines for comparison, corresponding to different pre-DMD initialization methods. Among them, (D) ODE Init and (E) Causal CD Init correspond to Self Forcing and Causal Forcing, while (A) the bidirectional base model, (B) AR diffusion, and (C) AR diffusion followed by Causal DMD are controlled variants. In particular, (C) and (E) share the same Stage-1 causal teacher and differ only in the distillation objective: Causal DMD is mode-seeking, whereas consistency distillation is mode-covering.

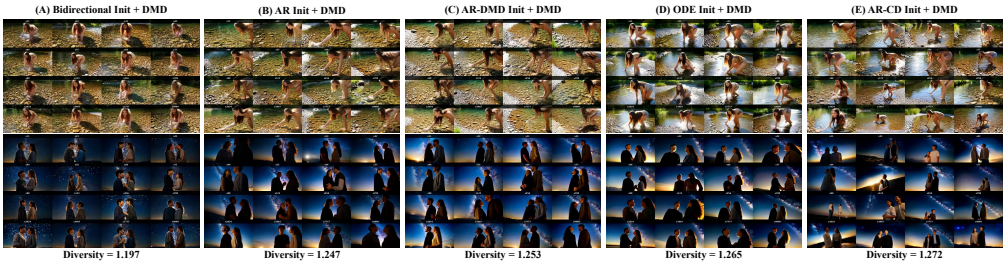


Figure 4: **Visual diversity comparison across different distillation pipelines.** For each pipeline, we sample 16 fixed seeds and visualize the middle frame of each generated 81-frame video as a grid.

The bar charts in Fig. 5 show that, except for direct bidirectional initialization, most AR-based pipelines achieve high VBench scores after DMD, making visual scores alone insufficient to distinguish initialization quality. Causal DMD Init is a representative case: it obtains the highest initialization VBench but does not yield the best final result. Our distribution evaluation explains this gap: AR Init and Causal DMD Init achieve high precision but low coverage, indicating concentration on limited high-quality modes. In contrast, Causal CD Init maintains broader coverage before and after DMD, providing better support for mode-seeking refinement. Moreover, under the same initialization target, coverage differences are also reflected in sample diversity. As shown in Fig. 4, pipelines D and E exhibit higher diversity, consistent with their broader coverage.

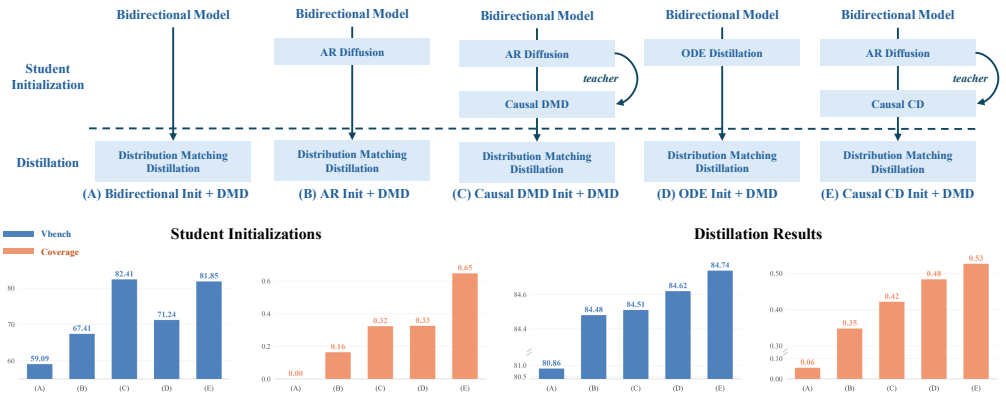


Figure 5: **Comparison across different distillation pipelines.** All pipelines use Wan-14B as the DMD teacher and use its sampled distillation data as the initialization target.

3.3 Balancing Mode Coverage and Mode Seeking

3.3.1 Distribution Evolution

Whereas Sec. 3.2 compares students *across* stages, we now track the dynamics *within* a single matched run, where initialization and distillation share the same target distribution. Fig. 6 projects V-JEPA2 features onto a PCA basis fitted on teacher samples and shows the student covariance contour at successive checkpoints, together with precision and coverage over training.

The left panel shows the evolution of the Causal CD initializer. At early checkpoints, the student distribution is still far from the teacher, but as training proceeds the few-step generator gradually expands its support toward it. This trend indicates the mode-covering behavior of trajectory- or consistency-based initialization: the initializer progressively reaches more teacher modes and improves coverage more than perceptual fidelity, so samples can remain blurry or under-refined, consistent with the averaging effect of mode-covering objectives. The middle panel shows the subsequent DMD stage initialized from this Causal CD model. Unlike the initialization stage, DMD no longer expands coverage; instead, the student distribution gradually drifts toward high-density teacher regions, consistent with the mode-seeking, RKL-style refinement illustrated in Fig. 2. As a result, later DMD checkpoints improve local fidelity and sharpness while reducing coverage and diversity.

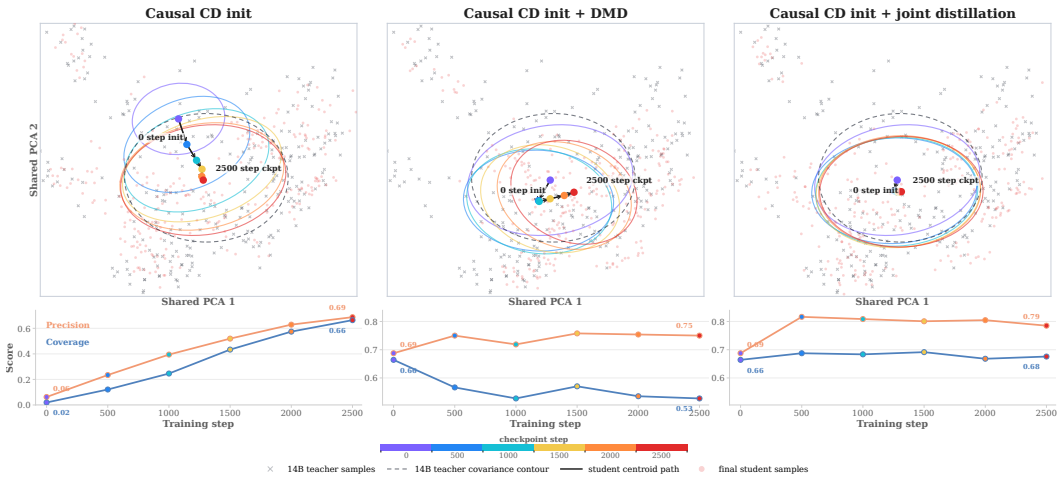


Figure 6: **Distribution evolution under mode-covering and mode-seeking objectives.** We visualize checkpoint distributions in a shared V-JEPA2 PCA space and track precision and coverage over training. Causal CD expands teacher coverage, pure DMD contracts toward high-density teacher regions, and joint distillation preserves coverage while refining samples.

This drift has two implications. On the one hand, concentrating probability mass near high-density teacher regions can improve visual quality by sharpening samples that are already close to plausible teacher modes. On the other hand, high-density regions under the teacher distribution do not necessarily contain all desirable videos. Samples with large motion, rich details, or rare dynamics may lie in relatively lower-density regions. Overly strong mode-seeking refinement can therefore suppress these valuable modes, trading diversity—and sometimes overall quality—for local realism.

3.3.2 Joint Distillation

To mitigate the late-stage distribution drift of DMD, we combine the mode-seeking objective with a mode-covering constraint under the same target distribution:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{DMD}} + \lambda \mathcal{L}_{\text{CD}}.$$

Here, $\mathcal{L}_{\text{CD}}$ is the second-stage consistency distillation loss, which acts as a distributional anchor for the student. While DMD selects and sharpens high-probability modes, the CD term helps maintain coverage over the teacher distribution and prevents excessive mode collapse. As shown in Fig. 6(c), the joint objective enables the student distribution to match teacher high-probability regions while preserving broader coverage, thus balancing mode seeking against mode covering and suppressing the drift of pure DMD.

4 Experiments

4.1 Setup

Implementation details. We adopt Wan2.1-T2V-1.3B (Team, 2025) as the base model for fine-tuning, which generates 81-frame videos at a resolution of 832×480. In the main experiments, we build upon the three-stage pipeline of Causal Forcing. We collect two sets of distillation data from Wan2.1-1.3B and Wan2.1-14B, respectively, each containing 25K samples. The prompt list is from VidProM (Wang & Yang, 2024). In the first stage, we perform autoregressive diffusion training with teacher forcing for 5K steps on each data set. In the second stage, we apply consistency distillation for 2.5K steps to obtain a few-step causal generator. In the final stage, we conduct 1.5K steps of joint DMD-CD training. The DMD teacher is chosen to match the corresponding data source, while the CD teacher is kept the same as the causal teacher used in Stage 2. We set the joint loss weight λ to 0.01.

Evaluation. Perceptual quality is measured with the VBench text-to-video protocol (Huang et al., 2024), reporting quality, semantic, and weighted overall scores under the same prompts and scripts for all methods. Distributional behavior is measured with the teacher-normalized protocol of Sec. 3.1: $N = M = 256$ student and teacher samples on matched prompts and seeds, V-JEPA2 features from 8 uniformly sampled frames within the first 5 seconds, and a k-NN support estimator with $k = 5$. Initialization-stage models are first refined with $\rho = 0.9$, whereas post-DMD models are evaluated directly. Precision and coverage are always computed against the DMD teacher listed in each row and are omitted for non-DMD systems without a target teacher. Diversity is the raw Vendi score (Friedman & Dieng, 2023) on the same features. The first-chunk analysis in Fig. 5 and Tab. 6 instead uses the first 12 consecutive frames; see Appendix B for full details.

4.2 Main Comparison

Table 3 compares our method with both open few-step video generation systems and autoregressive diffusion distillation baselines. The open baselines include LTX (HaCohen et al., 2024), Wan2.1-T2V-1.3B, SkyReels-DF(Chen et al., 2025), and CausVid(Yin et al., 2025b), evaluated with the same VBench protocol. For DMD-based autoregressive methods, we additionally report the DMD teacher, which separates the effect of teacher capacity from the effect of teacher-aligned initialization.

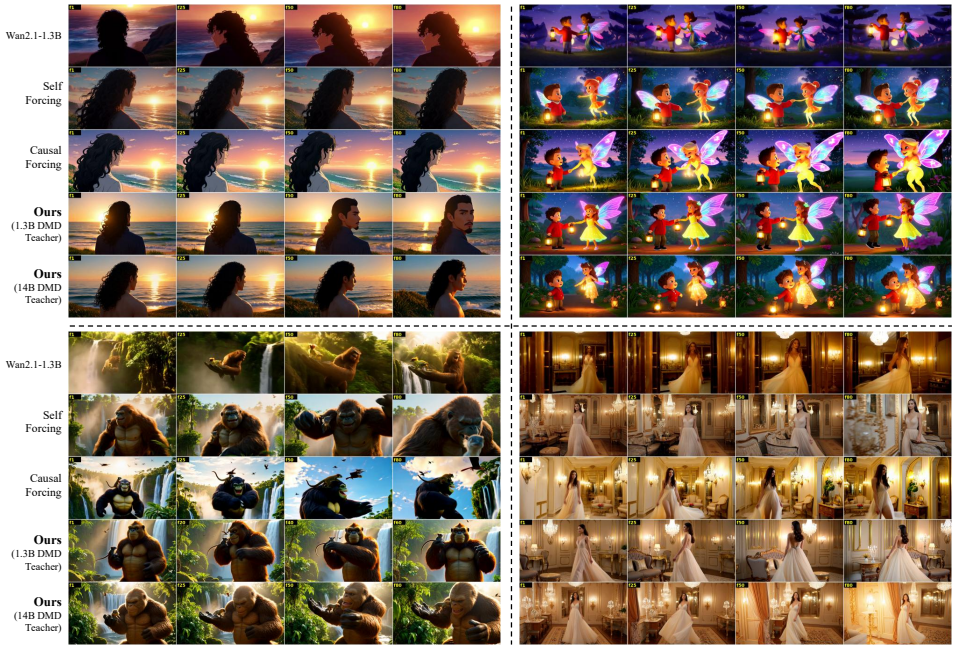


Figure 7: **Qualitative comparison of visual quality and motion.** Compared with representative bidirectional and distilled autoregressive baselines, our method produces sharper subjects, stronger motion, and more temporally coherent details.

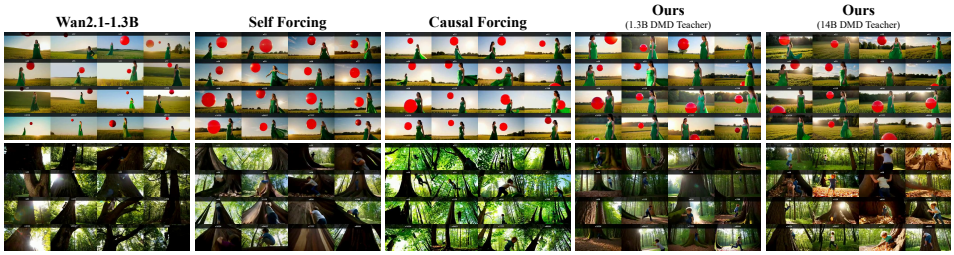


Figure 8: **Qualitative comparison of sample diversity.** Under fixed prompts and different seeds, our method preserves richer appearance and motion variations while maintaining high visual fidelity.

Our method achieves the best overall VBench score among the compared methods and also obtains the strongest coverage among distilled autoregressive models. Importantly, the comparison separates two factors that are often conflated in DMD-based autoregressive distillation: the DMD teacher

Table 3: **Main comparison on text-to-video generation.** VBench is reported with total, quality, and semantic scores. Precision and coverage are teacher-normalized against the listed DMD teacher when applicable; diversity is the raw Vendi score. "DMD Teacher" is only applicable to methods with a DMD refinement stage.

Method	DMD Teacher	VBench Metrics ↑			Distributional Metrics ↑		
		Total	Quality	Semantic	Precision	Coverage	Diversity
Bidirectional models							
LTX Video	–	81.37	83.05	74.66	–	–	1.281
Wan2.1-T2V-1.3B	–	83.56	84.31	80.55	–	–	1.334
Distilled autoregressive models							
CausVid	Wan2.1-14B	83.03	83.99	79.18	0.51	0.21	1.223
Self-Forcing	Wan2.1-14B	84.21	85.04	**80.87**	<u>0.78</u>	0.53	1.269
Causal-Forcing	Wan2.1-14B	84.38	85.47	80.01	0.61	0.29	1.250
Ours (weak)	Wan2.1-1.3B	<u>84.54</u>	<u>85.61</u>	80.28	0.77	<u>0.59</u>	**1.305**
Ours (full)	Wan2.1-14B	**84.83**	**85.92**	<u>80.47</u>	**0.79**	**0.69**	<u>1.304</u>

Table 4: **Ablation on the joint loss weight λ.** A larger λ strengthens the CD-based mode-covering constraint, while a smaller λ gives more weight to DMD refinement, and $\lambda = 0$ recovers pure DMD. For each λ, we report the best checkpoint within a 2500-step training trajectory. Diversity is reported as the raw Vendi score under the same V-JEPA2 protocol as Tab. 3. In each column, darker shading indicates a larger value.

λ	VBench Metrics ↑			Distributional Metrics ↑		
	Total	Quality	Semantic	Precision	Coverage	Diversity
1	84.18	85.17	80.18	0.73	0.71	1.311
0.1	84.67	85.72	80.46	0.76	0.69	1.303
0.01	84.83	85.92	80.47	0.79	0.69	1.304
0.001	84.78	85.85	80.55	0.78	0.64	1.307
0	84.74	85.78	80.54	0.72	0.53	1.292

and the initialization distribution. The weak variant remains competitive even when refined with a smaller Wan2.1-1.3B teacher, while the full variant further improves performance with the Wan2.1-14B teacher. This supports the view developed in Sec. 3: DMD benefits from an initialization that already covers the target teacher distribution, rather than relying on the refinement stage to recover missing modes. The coverage improvement is not merely a diagnostic artifact: it is accompanied by stronger raw diversity and better final VBench under the matched DMD teacher.

Fig. 7 provides qualitative examples consistent with the quantitative trend. Compared with prior distilled autoregressive baselines, our model better preserves object identity and motion details across frames, while avoiding the overly smoothed appearance typical of mode-covering initialization alone.

Fig. 8 further visualizes the diversity difference across seeds. Methods with lower teacher coverage tend to produce more similar appearances or motion patterns, whereas our matched joint distillation preserves broader variations without sacrificing visual quality.

4.3 Ablation on Joint Distillation

We ablate joint distillation from two aspects, complementing the analysis in Sec. 3.3: the joint loss weight λ, which balances the DMD mode-seeking objective against the CD-based mode-covering constraint, and the training dynamics across steps. For each λ, we train for 2500 steps and report the best checkpoint. As shown in Tab. 4, a larger λ yields the highest coverage but lower VBench scores, whereas a smaller λ strengthens mode seeking and improves visual quality and precision, with $\lambda = 0$ reducing to pure DMD. The best trade-off is reached at an intermediate $\lambda = 0.01$, which we adopt as the default.

Table 5: Ablation of joint distillation across training steps. Compared with pure DMD, joint distillation better preserves coverage and diversity while achieving comparable or better VBench scores. Diversity is reported as the raw Vendi score under the same V-JEPA2 protocol as Tab. 3.

Step	Joint Distillation ($\lambda = 0.01$)						Pure DMD ($\lambda = 0$)					
	VBench Metrics ↑			Distributional Metrics ↑			VBench Metrics ↑			Distributional Metrics ↑		
	Total	Quality	Semantic	Precision	Coverage	Diversity	Total	Quality	Semantic	Precision	Coverage	Diversity
0	82.60	84.18	76.27	0.69	0.66	1.319	82.60	84.18	76.27	0.69	0.66	1.319
500	84.60	85.87	79.53	0.82	0.67	1.305	84.44	85.52	80.09	0.75	0.56	1.297
1000	84.78	85.85	80.52	0.80	0.68	1.304	84.74	85.78	80.54	0.72	0.53	1.292
1500	84.83	85.92	80.47	0.79	0.69	1.304	84.40	85.50	80.01	0.76	0.55	1.285
2500	84.27	85.26	80.32	0.79	0.68	1.307	83.91	85.01	79.51	0.75	0.53	1.272

We then track joint distillation across training steps under $\lambda = 0.01$, quantifying the trend visualized in Fig. 6. As shown in Tab. 5, pure DMD improves visual quality early on, but its coverage and diversity decrease continuously as training proceeds. Joint distillation instead maintains higher coverage and diversity throughout training while achieving comparable or better quality, semantic, and total scores, and its advantage widens in the late stage. The CD-based mode-covering constraint thus suppresses DMD-induced distribution drift without sacrificing generation quality.

4.4 Analysis of the Distributional Evaluation Protocol

Table 6 reports the complete metrics for the five pipelines in Fig. 5, at the initialization stage and after DMD, all targeting Wan2.1-14B. Consistent with Sec. 3.2, the mode-covering Causal CD initialization (E) attains the highest coverage both before and after DMD, whereas mode-seeking initializations such as Causal DMD (C) reach high precision but lower coverage. The table also shows why teacher-normalized re-noising is necessary: without it, low-level fidelity confounds coverage. The visibly blurrier ODE initializer (D) obtains only 0.020 raw coverage, far below the sharper Causal DMD (0.324), yet yields *higher* coverage after DMD (0.484 vs. 0.422). After re-noising, ODE reaches 0.326, consistent with the post-DMD ordering, indicating that raw coverage is dominated by the denoising state whereas the re-noised protocol reflects the semantic support available for subsequent refinement (Fig. 9).



Figure 9: **Effect of teacher-normalized re-noising.** Re-noising with a shared teacher preserves the scene semantics of each initializer while removing differences in low-level fidelity.

Table 6: **Complete results across distillation pipelines.** VBench, precision, coverage, and diversity for the five initialization pipelines in Fig. 5, at the student-initialization stage and after DMD refinement. For initialization coverage we report both raw coverage (on the raw outputs) and teacher-normalized re-noised coverage (after the shared refinement with $\rho = 0.9$); all coverage is computed against the Wan2.1-14B reference. In every column, darker shading indicates a larger value.

Pipeline Init.	Student Initialization				Distillation Result			
	VBench ↑	Precision ↑	Raw Coverage ↑	Re-noised Coverage ↑	VBench ↑	Precision ↑	Coverage ↑	Diversity ↑
(A) Bid	59.09	0.000	0.000	0.000	80.86	0.137	0.055	1.197
(B) AR	67.41	0.238	0.008	0.164	84.48	0.691	0.348	1.247
(C) Causal DMD	82.41	0.645	0.324	0.324	84.51	0.672	0.422	1.253
(D) ODE	71.24	0.488	0.020	0.326	84.62	0.867	0.484	1.265
(E) Causal CD	81.85	0.770	0.664	0.648	84.74	0.750	0.527	1.272

5 Conclusion

We revisit autoregressive video distillation from a distributional perspective. Since trajectory- or consistency-based initialization is mode-covering while DMD is mode-seeking, an initialization should be judged by how well it covers the target DMD teacher rather than by visual scores alone. We introduce a teacher-normalized evaluation protocol to measure this coverage. Applying it to the training dynamics within distillation, we further find that pure DMD drifts toward high-density teacher regions even under aligned targets, and counter this drift with joint distillation, which pairs DMD's mode-seeking objective with a CD-based mode-covering constraint. Experiments confirm a better balance among visual quality, coverage, and diversity.

AI USE STATEMENT

In this work, we used generative AI tools for grammar checking and writing refinement. We have not used generative AI tools for any other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. We have not used generative AI tools for any additional tasks with recommended disclosure. We have reviewed all AI-assisted work. All AI-assisted text was carefully reviewed and revised by the authors to ensure accuracy and consistency with the intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies distillation of text-to-video diffusion models and does not involve human subjects, personal data, or newly collected datasets. All prompts are drawn from the public VidProM prompt set, and all teacher models are publicly released open-weight models. Video generation models can be misused to produce misleading or harmful content; our method does not introduce new generation capabilities beyond those of the teacher models it distills, and it inherits their content safeguards and limitations. We will release code, data, and models under a license that requires responsible use.

REPRODUCIBILITY STATEMENT

We have taken several steps to make our results reproducible. The training pipeline, data sources, training steps, and the joint loss weight are specified in Sec. 4.1. The complete distributional evaluation protocol, including the re-noising procedure, video encoder, feature extraction, nearest-neighbor support estimation, and the exact prompt–seed grid, is documented in Appendix B. Full ablation results with the Wan2.1-1.3B teacher are given in Appendix C, and the prompts used for all qualitative figures are listed in Tab. 9. Code, distillation data, and model checkpoints will be publicly released upon acceptance.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Shengqu Cai, Weili Nie, Chao Liu, Julius Berner, Lvmin Zhang, Nanye Ma, Hansheng Chen, Maneesh Agrawala, Leonidas Guibas, Gordon Wetzstein, and Arash Vahdat. Mode seeking meets mean seeking for fast long video generation. In *International Conference on Machine Learning*, 2026.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, 2024.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. In *International Conference on Learning Representations*, 2026.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

Junyao Gao, Jiaxing Li, Wenran Liu, Yanhong Zeng, Fei Shen, Kai Chen, Yanan Sun, and Cairong Zhao. Charactershot: Controllable and consistent 4d character animation. *arXiv preprint arXiv:2508.07409*, 2025.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

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*, 2025.

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.

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 *International Conference on Learning Representations*, 2024.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jiaxing Li, Hongbo Zhao, Yijun Wang, and Jianxin Lin. Towards photorealistic video colorization via gated color-guided image diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10891–10900, 2024.

Jiaxing Li, Jiepeng Wang, Junyao Gao, Yang Liu, Eric Li, Bo An, and Hao-Xiang Guo. Dynamicsboost: Dynamic plausible video generation via annotation-free continuation preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20024–20033, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

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

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, 2020.

PixVerse Research. Pixverse-r1: Next-generation real-time world model. https://pixverse.ai/en/blog/pixverse-r1-next-generation-real-time-world-model, January 2026. Accessed: 2026-06-20.

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, et al. Matrix-game 3.5: Enhancing real-time streaming interactive world models with patch memory. *arXiv preprint arXiv:2608.29910*, 2026.

Shaohao Rui, Xiaofeng Mao, Zhanyu Zhang, Peijia Lin, Yansong Zhu, Yibo Zhang, Haibin Wan, and Weijie Ma. Biwm: Advancing open-source interactive video world models with bidirectional autoregression. *arXiv preprint arXiv:2606.10135*, 2026.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. In *International Conference on Learning Representations*, 2026.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *Advances in Neural Information Processing Systems*, 37:65618–65642, 2024.

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.

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. In *International Conference on Machine Learning*, 2026.

Desai Xie, Zhan Xu, Yicong Hong, Hao Tan, Difan Liu, Feng Liu, Arie Kaufman, and Yang Zhou. Progressive autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2025.

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

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.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T 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, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

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, 2025b.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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. *arXiv preprint arXiv:2605.15141*, 2026.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2023.

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*, 2026.

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 *International Conference on Machine Learning*, 2024.

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. In *International Conference on Machine Learning*, 2026.

Kai Zou, Dian Zheng, Hongbo Liu, Tiankai Hang, Bin Liu, and Nenghai Yu. Hiar: Efficient autoregressive long video generation via hierarchical denoising. *arXiv preprint arXiv:2603.08703*, 2026.

A Related Work

A.1 Autoregressive Video Diffusion Models.

With the rapid progress of video generation (Team, 2025; Yang et al., 2024; Li et al., 2026; Gao et al., 2025; Li et al., 2024), a growing line of work reformulates video generation autoregressively, either by changing the temporal denoising process itself or by converting a pretrained bidirectional video diffusion model into a causal generator. Diffusion Forcing and progressive autoregressive diffusion introduce per-frame or progressively scheduled noise levels to bridge next-token prediction and full-sequence diffusion (Chen et al., 2024; Xie et al., 2025). More closely related to our setting, CausVid distills bidirectional video diffusion models into few-step autoregressive generators with ODE initialization and distribution matching (Yin et al., 2025a); Self-Forcing reduces exposure bias by training on self-generated histories with video-level distribution matching (Huang et al., 2025); Causal Forcing shows that ODE initialization from a bidirectional teacher can be target-misaligned and instead uses an autoregressive teacher before DMD refinement (Zhu et al., 2026); and Causal Forcing++ further scales this idea with causal consistency distillation for frame-wise one- and two-step generation (Zhao et al., 2026). Recent follow-ups push this direction toward longer horizons, stronger controls, and lower latency: Rolling Forcing, Self-Forcing++, Context Forcing, and MotionStream target multi-minute rollout or interactive motion control (Liu et al., 2025; Cui et al., 2026; Chen et al., 2026; Shin et al., 2026); and HiAR and BiWM explore hierarchical or bidirectional autoregressive generation to mitigate accumulated errors (Zou et al., 2026; Rui et al., 2026).

A.2 Forward and Reverse Divergence in Diffusion Models.

A useful lens for diffusion distillation is the divergence implicitly induced by the training distribution. Offline, teacher-supervised objectives—including progressive distillation, consistency distillation, consistency trajectory models, and continuous-time consistency models—train on teacher or data trajectories and thus exhibit forward-divergence or mean-seeking behavior, which tends to cover teacher samples but may average modes or lose fine detail (Salimans & Ho, 2022; Song et al., 2023; Kim et al., 2024; Lu & Song, 2025). In contrast, on-policy score-distillation and distribution-matching objectives, including DMD, DMD2, and SiD, optimize student-generated samples against teacher scores and are commonly associated with reverse-KL-style, mode-seeking behavior: they improve sharpness and fidelity but can collapse onto high-probability modes (Yin et al., 2024b;a; Zhou et al., 2024). Recent work has made this trade-off explicit. (f)-Distill generalizes score-based distribution matching beyond reverse KL and shows that forward KL or Jensen–Shannon alternatives can improve mode coverage (Xu et al., 2025); rCM combines the mode-covering behavior of continuous-time consistency with score regularization to recover reverse-divergence fidelity at large scale (Zheng et al., 2026); ADM/DMDX uses adversarial distribution matching to alleviate the mode collapse caused by reverse-KL DMD (Lu et al., 2025); SGMD analyzes the mode-seeking conservatism of DMD-style video distillation and improves motion dynamics by matching score gradients (Wu et al., 2026); Adaptive Video Distillation targets oversaturation, temporal inconsistency, and temporal collapse in few-step video generation (You et al., 2026); and Mode Seeking meets Mean Seeking decouples a supervised mean-seeking flow-matching head from a reverse-KL distribution-matching head for long-video synthesis (Cai et al., 2026). In autoregressive video generation, HiAR explicitly adds a forward-KL regularizer to preserve motion diversity under self-rollout, and BiWM adds GAN and mass-covering forward-KL objectives to counter DMD-induced mode-seeking degradation (Zou et al., 2026; Rui et al., 2026).

B More Details of Coverage Metrics

This section specifies the complete evaluation pipeline used for the precision, coverage, and diversity results in the paper. We distinguish the *normalization teacher* T_{norm}, which removes nuisance differences among raw initializer outputs, from the *reference teacher* T_{ref}, whose sample distribution defines the target support. The two teachers need not be identical. This distinction is important: the normalization operation is shared across initializers, whereas T_{ref} is selected according to the target DMD teacher reported in each experiment.

Evaluation sample sets. We use a fixed grid of 16 prompts and 16 random seeds, yielding one video for each prompt–seed pair and $N = M = 256$ videos for both the evaluated model and the reference teacher. The seeds are

$$\{11, 22, 33, 42, 44, 55, 66, 77, 88, 123, 456, 789, 2024, 3407, 7777, 9999\}.$$

The same grid is used for every method and teacher. Matching the grid controls the prompt mixture and sampling budget, but the reported precision and coverage remain set-level statistics: they are computed on the pooled 256-sample sets, not by comparing only videos that share a prompt or seed. We require exactly one sample for every key $(\text{prompt}, \text{seed})$ and do not impute failed or missing generations.

Teacher-normalized refinement of initializer samples. For a pre-DMD initializer, we reuse its saved terminal video latent $z_{\theta,0}$; this avoids an additional decode–encode pass. Following the linear path used by the flow-matching teacher (Lipman et al., 2022), we draw $\epsilon \sim \mathcal{N}(0, I)$ and form

$$z_{\theta,\rho} = (1 - \rho) z_{\theta,0} + \rho\epsilon. \tag{12}$$

For a given prompt–seed pair, the noise seed is fixed and shared across initializers, so the normalization itself does not introduce method-dependent randomness. Unless otherwise stated, $\rho = 0.9$. We then apply a fixed Wan2.1-T2V-14B normalization teacher with classifier-free guidance scale 5.0. All methods use the same unconditional negative prompt. The solver is Flow-UniPC (Zhao et al., 2023), configured with 1,000 training timesteps, schedule shift 1, and no dynamic shifting. We use $22 = \mathrm{round}(25\rho)$ model evaluations. Their sigma values are uniformly spaced from 0.9 through

$0.9/22$, after which the scheduler performs the terminal update to zero. The refined latent is decoded at 16 fps to obtain $\tilde{x}_\theta^{(0.9)}$.

The reference set is generated separately with the 25-step sampler of T_{ref}. Thus, for experiments targeting Wan2.1-14B, the shared 1.3B model performs only the normalization step, while the neighborhood radii and all reported precision/coverage values are defined by Wan2.1-14B reference samples. Post-DMD generators are evaluated directly from their final videos without re-noising: their outputs are already fully denoised, and the purpose of normalization is specifically to remove the heterogeneous denoising state of pre-DMD initializers.

Temporal standardization and video representation. Our default distribution protocol uses the frozen `facebook/vjepa2-vith-fpc64-256` V-JEPA2 encoder (Assran et al., 2025). We restrict every video to the first five seconds. In our 16-fps setup this is the 81-frame prefix; for methods that generate longer videos, later frames are discarded before feature extraction. We select eight frames with rounded, uniformly spaced indices over this prefix, including both temporal endpoints. The same frame-selection rule is applied to teacher and student videos. This short-window protocol is applied symmetrically to the student and reference sets, and its absolute values are not mixed with the default 8-frame results in the main comparison.

The selected frames are processed by the model's native video processor and encoded in bfloat16. Let $h_1, \ldots, h_L \in \mathbb{R}^{1280}$ denote the output spatiotemporal tokens. We aggregate them into one video descriptor by concatenating the token-wise mean and sample standard deviation (correction one, matching the extractor implementation),

$$f(x) = \left[\frac{1}{L} \sum_{\ell=1}^{L} h_\ell \; ; \; \mathrm{std}_{\ell=1}^{L}(h_\ell) \right] \in \mathbb{R}^{2560}, \qquad \phi(x) = \frac{f(x)}{\|f(x)\|_2}. \tag{13}$$

All reported results use these raw, ℓ_2-normalized V-JEPA2 features. We do not apply low-pass filtering, sharpness residualization, dimensionality reduction, or PCA before computing the metrics. PCA is used only for the visualizations in Fig. 6 and does not affect any reported number.

Local teacher support. We use a non-parametric k-nearest-neighbor support estimator, following the improved precision construction of Kynkäänniemi et al. (2019) and the coverage statistic of Naeem et al. (2020). Let $V = \{v_j\}_{j=1}^{M}$ be the normalized features of T_{ref} and $U = \{u_i\}_{i=1}^{N}$ those of the evaluated model. Distances are Euclidean in the normalized feature space,

$$d(a, b) = \|a - b\|_2, \qquad d(a, b)^2 = 2 - 2a^\top b, \tag{14}$$

so ranking by this distance is equivalent to ranking by cosine distance. For each teacher feature, we define an adaptive local radius

$$r_k(v_j) = \text{distance from } v_j \text{ to its } k\text{-th nearest element of } V \setminus \{v_j\}. \tag{15}$$

We use $k = 5$ for every experiment; k is fixed globally and is never tuned per method, checkpoint, or teacher.

Precision and coverage. Precision is the fraction of evaluated samples that fall inside at least one adaptive teacher neighborhood:

$$\mathrm{Prec}(U, V) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\exists j : \; d(u_i, v_j) \le r_k(v_j)] \,. \tag{16}$$

Coverage is the fraction of teacher samples whose own neighborhood contains at least one evaluated sample:

$$\mathrm{Cov}(U, V) = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}\Big[\min_i d(v_j, u_i) \le r_k(v_j)\Big] \,. \tag{17}$$

Although reported beside precision, this coverage statistic is not the recall of the improved precision–recall metric. Recall constructs neighborhoods from the generated samples; Eq. 17 deliberately uses only teacher radii, which makes the target support fixed across all compared methods and prevents

isolated student outliers from creating artificially large regions. With $M = 256$, coverage changes in increments of $1/256$ before rounding.

The two statistics measure complementary failure modes. High precision means that most model samples lie near the estimated teacher support, but it does not imply broad support: a strongly mode-seeking generator can have high precision and low coverage. High coverage means that many local teacher neighborhoods are reached, but it is not a standalone perceptual-quality score. Both quantities are conditional on the feature encoder, prompt set, temporal window, and T_{ref}. Consequently, values computed against different reference teachers should be interpreted relative to their respective supports rather than ranked as if they shared one absolute scale.

Diversity statistic. For completeness, all diversity entries in the paper use the raw Vendi score (Friedman & Dieng, 2023) on the same normalized V-JEPA2 descriptors. For a method feature matrix $U \in \mathbb{R}^{N\times 2560}$, we form the shifted cosine kernel

$$K = \frac{UU^\top + \mathbf{1}\mathbf{1}^\top}{2}, \qquad \bar{K} = \frac{K}{N}. \tag{18}$$

If $\{\lambda_q\}$ are the positive eigenvalues of $\bar{K}$, the reported score is

$$\mathrm{Vendi}(U) = \exp\left(-\sum_q \lambda_q \log \lambda_q\right). \tag{19}$$

We report this raw effective-rank value, not a ratio to the teacher. It is a within-method diversity statistic and does not enter the precision or coverage calculation.

C Additional Ablation

C.1 Joint Distillation

To complement the Wan2.1-14B ablations in Sec. 4.3, we repeat the joint distillation ablations on the weak setting, where Wan2.1-1.3B serves as both the initialization data source and the DMD teacher. The conclusions are consistent with the main setting: joint distillation better preserves coverage and diversity than pure DMD, and $\lambda = 0.01$ provides the best overall trade-off.

Loss weight λ. Table 7 ablates λ under the weak setting, reporting the best checkpoint within a 2500-step training trajectory for each λ as in the main setting. As λ decreases, the DMD mode-seeking effect strengthens, improving precision but reducing coverage and diversity; $\lambda = 0$ recovers pure DMD. Consistent with the Wan2.1-14B result in Tab. 4, $\lambda = 0.01$ gives the best overall trade-off among quality, precision, and coverage.

Table 7: **Ablation on the joint loss weight λ under the weak (Wan2.1-1.3B) setting.** A larger λ strengthens the CD-based mode-covering constraint, while a smaller λ gives more weight to DMD refinement, and $\lambda = 0$ recovers pure DMD. Diversity is reported as the raw Vendi score under the same V-JEPA2 protocol as Tab. 3. In each column, darker shading indicates a larger value.

λ	VBench Metrics ↑			Distributional Metrics ↑		
	Total	Quality	Semantic	Precision	Coverage	Diversity
1	84.18	85.17	80.18	0.65	0.53	1.295
0.1	84.34	85.41	80.06	0.67	0.52	1.298
0.01	84.54	85.61	80.28	0.77	0.59	1.305
0.001	84.52	85.63	80.09	0.71	0.50	1.299
0	84.50	85.58	80.19	0.72	0.44	1.275

Across training steps. Table 8 tracks joint distillation and pure DMD across training steps under the weak setting. Pure DMD steadily loses coverage and diversity as training proceeds, whereas joint distillation maintains higher coverage and diversity while reaching comparable or better total scores in the later stage, mirroring the Wan2.1-14B trend in Tab. 5.

Table 8: **Ablation of joint distillation across training steps under the weak (Wan2.1-1.3B) setting.** Compared with pure DMD, joint distillation better preserves coverage and diversity while achieving comparable or better VBench scores. Diversity is reported as the raw Vendi score under the same V-JEPA2 protocol as Tab. 3. In each column, darker shading indicates a larger value.

Step	Joint Distillation ($\lambda = 0.01$)						Pure DMD ($\lambda = 0$)					
	VBench Metrics ↑			Distributional Metrics ↑			VBench Metrics ↑			Distributional Metrics ↑		
	Total	Quality	Semantic	Precision	Coverage	Diversity	Total	Quality	Semantic	Precision	Coverage	Diversity
0	82.16	83.38	77.29	0.61	0.52	1.301	82.16	83.38	77.29	0.61	0.52	1.301
500	83.84	85.07	78.92	0.66	0.52	1.295	84.44	85.52	80.09	0.68	0.48	1.293
1000	84.17	85.38	79.29	0.69	0.51	1.292	84.48	85.52	80.35	0.72	0.46	1.284
1500	84.38	85.33	80.58	0.72	0.53	1.297	84.50	85.58	80.19	0.72	0.44	1.275
2500	84.54	85.61	80.28	0.77	0.59	1.305	84.09	85.22	79.59	0.73	0.43	1.269

D PROMPTS

For reproducibility, we list the prompts used for the qualitative examples and visual comparisons in this paper. For fair comparison, all methods in the same figure use the same prompt, and when applicable, the same set of fixed random seeds. The reference column indicates the figures in which each prompt is used.

Table 9: Prompts used in this paper.

Prompt	**Reference**
Four teenage girls, with windswept hair and determined expressions, ride powerful waves in frigid, crystal-clear coldwater. Their vibrant surfboards cut through frothy whitecaps as they lean into turns with athletic grace. Dynamic camera angles show drone sweeps, close-up shoulder shots, and overhead tracking capture their synchronized movements. Each girl wears snug wetsuits, laughing and shouting as they paddle, carve, and ride the swell. Motion is fluid and energetic, emphasizing their youthful spirit against the vast, chilly ocean backdrop.	Fig. 9
Zara, a young woman with flowing hair and curious expression, stumbles joyfully upon a small, crystal-clear brook, its surface shimmering with sunlight as it bubbles gently over smooth pebbles visible beneath. She bends slightly, reaching out to touch the cool water, her reflection dancing in the ripples. The camera glides softly around her, capturing her delighted surprise from low angles and overhead shots, emphasizing the serene, natural beauty and the soothing motion of water swirling around her feet.	Fig. 4
A young couple stands side by side under a vast, starry night sky, gazing upward in wonder as distant stars flicker with soft, ethereal light. Their faces glow with awe, hands lightly touching, shoulders relaxed. The camera slowly pans around them, tilting upward to capture the infinite cosmos, then gently zooms in on their eyes reflecting the stars. Shot in cinematic 4K HD, with deep blues and warm golden hues, the scene pulses with quiet romance and cosmic wonder as gentle wind rustles their hair.	Fig. 4
A vibrant red ball soars dynamically through the air, arcing gracefully toward a graceful lady in a flowing green dress, standing poised in a sunlit field. Her posture is serene, eyes tracking the ball's approach as it nears her midsection. The camera tracks smoothly from a low angle, rising slightly as the ball descends. Soft wind ripples her dress. The scene glows with cinematic realism, rich colors, and natural motion — the ball's flight and her subtle lean create anticipation.	Fig. 8
Sammy, a curious young boy with tousled brown hair and a determined expression, sprints through the dappled forest toward a towering ancient oak. He leaps with joy, sliding headfirst into its hollow trunk as leaves rustle around him. The camera follows him in a dynamic tracking shot, then swoops low to capture his playful disappearance. Sunlight filters through the canopy, casting dancing shadows. His laughter echoes as he vanishes, leaving the forest alive with mystery and motion.	Fig. 8

Prompt	Reference
A man with long, flowing curly black hair stands alone on a windswept beach cliff, gazing at the sun setting over the endless ocean captured from behind in a stunning anime-style digital illustration by Makoto Shinkai. His hollow cheeks and contemplative posture evoke melancholy beauty, as if frozen in his last moment of solitude. The scene glows with warm, cinematic hues, enhanced by soft atmospheric lighting and gentle waves crashing below. The camera slowly pans right, then tilts up to frame the sky, blending romanticism with aestheticism. Motion: slow, reflective, emotionally resonant.	Fig. 7
The story ends with little boy Teddy, wearing a cozy red sweater and holding a glowing lantern, joyfully dancing beside fairy Sparkle, whose wings shimmer with aurora-like colors. They float side-by-side through a dreamy forest at dusk, laughing and twirling as camera circles them in slow motion. 3D animation, 4K resolution, 16:9, cinematic lighting, soft depth of field, whimsical fantasy style. Their bond glows with warmth as they leap into the next adventure, hand in hand.	Fig. 7
Cinematic wide shot of a tiny tarsier perched dynamically on King Kong's giant palm, surrounded by lush forest near cascading waterfalls, with fluttering birds soaring in golden sunlight that highlights the misty background. King Kong leaps joyfully with playful companions, capturing natural motion through fluid camera pans and dolly movements. The tarsier's curious expression and agile movements contrast with the colossal scale of its surroundings, emphasizing whimsy and wonder in a vibrant, high-detail fantasy scene.	Fig. 7
A young, radiant beauty gracefully walks through an opulent luxury room, mid-shot, her flowing gown swaying with each step. Soft golden light bathes her as she glides past gilded furniture and crystal chandeliers, her expression serene yet confident. The camera smoothly follows her from a medium distance, subtly panning to capture her poised posture and elegant stride. Rich textures, ambient reflections, and slow-motion movement enhance the luxurious atmosphere, making every gesture feel graceful and intentional.	Fig. 7