TSD: FEW-STEP VIDEO DIFFUSION WITH TEMPORALLY DECOUPLED DISTRIBUTION MATCHING DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable video generation is central to interactive applications and digital content creation. Although distribution matching distillation (DMD) enables few-step video generation, balancing visual fidelity with coherent motion remains challenging during prolonged training, and temporal variation may even collapse. We present Temporal Subspace Distillation (TSD), a general supervision framework designed to make better use of teacher knowledge during continued training. Our key idea is to decompose video latents into a temporal mean replicated across frames and zero-mean deviations. These components lie in orthogonal subspaces, enabling separate supervision of persistent visual content and temporal variation. Building on this decomposition, we introduce an identity anchor to transfer time-constant content from the teacher and one-sided deviation floors to penalize insufficient temporal variation in the student. We further extract motion saliency from the same teacher predictions to guide where variation is protected without prescribing a specific motion trajectory. Experiments show that TSD improves video quality over baseline methods, producing videos with high visual fidelity and coherent, plausible motion. TSD retains the student architecture and inference procedure and adds no inference cost.

1 INTRODUCTION

Video diffusion models synthesize detailed, temporally coherent content, but their iterative sampling remains costly for interactive applications. Distribution Matching Distillation (DMD) has emerged as a leading framework for accelerating diffusion generation by reducing the number of sampling steps (Yin et al., 2024b;a). It has been widely adopted in recent autoregressive and real-time video generation systems (Yin et al., 2025; Huang et al., 2026; Zhu et al., 2026). However, DMD training can be unstable, limiting the benefits of continued training. Existing video distillation recipes use relatively short DMD stages, with Self Forcing reporting 600 training iterations and Causal Forcing using 750 (Huang et al., 2026; Zhu et al., 2026). We observed that continued distillation often improves the visual fidelity of individual frames, but motion may degrade. This degradation can appear as nearly static outputs, temporal discontinuities, or implausible movements and interactions (Figure 1). Our goal is to retain the visual benefits of continued training while maintaining coherent and plausible motion.

Prior work combines adaptive regression with temporal regularization to improve appearance and preserve temporal variation (You et al., 2026). However, its regression objective acts on complete video predictions, so the supervision intended to improve appearance can also alter temporal variation and potentially suppress desired motion. Its temporal regularizer encourages overall variation without distinguishing intended motion from unwanted changes in regions that should remain stable, potentially rewarding spurious fluctuations rather than meaningful motion. An explicit decomposition into persistent and time-varying components would allow these objectives to target their respective components, providing more direct control over video distillation.

We therefore propose Temporal Subspace Distillation (TSD), which decomposes each video latent into its temporal mean replicated across frames and the deviations from that mean. These components lie in two orthogonal subspaces: the subspace of signals that are constant across frames and the

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.

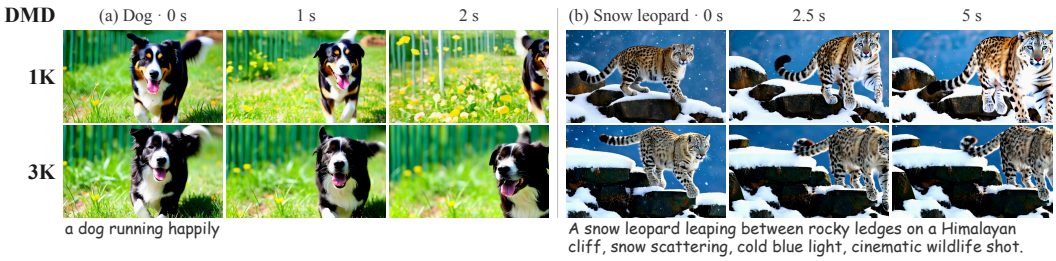

Figure 1: **Continued distillation.** We compare DMD at 1K and 3K training iterations. The 3K samples show finer details, but the motion becomes less apparent and physically implausible.

subspace of signals with zero temporal mean. This decomposition enables targeted auxiliary supervision for the two components: distilling time-averaged content from the teacher while protecting the student's own temporal variation.

Building on this decomposition, we introduce two complementary forms of auxiliary supervision. To transfer persistent content without prescribing a teacher motion trajectory, the *Identity Anchoring Loss* (IA) aligns the student's temporal mean with that of the clean teacher prediction already computed by DMD. To counter the loss of temporal variation, the one-sided *Motion Floor* (MF) and *Velocity Floor* (VF) losses penalize insufficient temporal variance and low-pass temporal change, respectively. The decomposition places IA's latent output gradient in the mean subspace and those of MF and VF in the orthogonal deviation subspace, allowing targeted supervision of each component alongside the original DMD update. We combine these terms with light temporal smoothing to discourage rapid fluctuations. To guide where variation is protected, an extended configuration uses teacher-derived motion saliency to give greater weight to regions where the teacher predicts temporal change. This design makes better use of teacher knowledge by combining time-averaged content supervision with spatial guidance for temporal variation. To complement these latent-space constraints, we extend the same floor formulation to optical flow computed from decoded frames, providing an optional objective that directly regularizes pixel-space displacement.

We evaluate our default configuration, TSD-B (Balanced), and its extension, TSD-D (Dynamic). Experiments show that TSD achieves higher overall scores on VBench-1 and VBench-2 than vanilla DMD and temporal-variance regularization baselines. Qualitative comparisons illustrate high visual fidelity, coherent actions, and consistent object relationships under both short and detailed prompts (Figure 3). Component ablations demonstrate the benefits of combining mean anchoring with deviation protection. Evaluations with CF++ initialization and two-step generation further support the applicability of the framework across different distillation settings. TSD achieves these gains through auxiliary supervision during training, while the student architecture, inference procedure, and inference cost remain unchanged.

Overall, our contributions are:

- We propose Temporal Subspace Distillation (TSD), a framework that explicitly decomposes video latents into two orthogonal temporal subspaces for targeted supervision of persistent content and temporal variation.
- We introduce an identity anchor and a family of one-sided latent deviation floors to improve content learning and mitigate motion degradation during continued distillation.
- We develop a supervision strategy that makes better use of existing teacher predictions for both content learning and temporal regularization. Motion-saliency weighting (MSW) extracts spatial motion cues to guide where temporal variation is protected.

2 Related Work

2.1 Video Generation

Video diffusion models learn the joint spatial and temporal structure through iterative denoising (Ho et al., 2022). Recent latent diffusion transformers scale this approach to high-quality text-to-video generation (Ma et al., 2024; Yang et al., 2025b; Kong et al., 2024; Wan et al., 2025). These bidirec-

tional models denoise an entire clip jointly, using context from both temporal directions. In particular, Wan2.1 (Wan et al., 2025) combines a spatiotemporal VAE with full spatiotemporal attention and provides the pretrained models used in our study.

Autoregressive video models extend generation through successive frames or chunks conditioned on preceding context (Chen et al., 2024; Teng et al., 2025; Chen et al., 2025). Diffusion Forcing (Chen et al., 2024) connects autoregressive prediction with diffusion by training causal sequence models with independent noise levels across tokens. MAGI-1 (Teng et al., 2025) combines bidirectional attention within chunks with causal attention across chunks for scalable video generation. SkyReels-V2 (Chen et al., 2025) uses frame-specific noise schedules to support autoregressive generation and long-video continuation. LongLive (Yang et al., 2025a) and LongLive-2.0 (Chen et al., 2026) further develop streaming long-video generation, supporting interactive prompt changes and improving efficiency. These approaches provide complementary ways to organize temporal context and extend generation beyond a fixed clip.

2.2 Diffusion Distillation

Diffusion distillation compresses iterative sampling into one or a few steps through trajectory approximation, consistency learning, or distribution matching (Salimans & Ho, 2022; Song et al., 2023; Yin et al., 2024b). DMD (Yin et al., 2024b) provides a distribution match objective based on a pretrained teacher score and an online estimate of the student's distribution score, together with a paired regression regularizer. DMD2 (Yin et al., 2024a) removes the paired-regression requirement, stabilizes training with two-time-scale updates, and supports multi-step generation.

For video generation, distribution matching has enabled few-step autoregressive students (Yin et al., 2025; Huang et al., 2026; Zhu et al., 2026). CausVid (Yin et al., 2025) applies asymmetric distribution matching to distill a bidirectional video teacher into a causal student. Self Forcing (Huang et al., 2026) trains with autoregressive rollouts on self-generated context and video-level distribution matching, reducing the gap between training and inference. Causal Forcing (Zhu et al., 2026) uses an autoregressive diffusion teacher for causal ODE distillation, followed by DMD with a bidirectional teacher. We build on this final distribution-matching stage.

Motion Consistency Model (MCM) (Zhai et al., 2024) combines motion distillation with image adversarial learning, while AVD (You et al., 2026) pairs adaptive regression with temporal-variance regularization. However, neither confines the resulting output gradients to the temporal mean. Appearance refinement can therefore also modify frame-to-frame variation, even with a separate temporal objective. TSD anchors the student's temporal mean to the teacher and regularizes deviations in an orthogonal subspace, separating the auxiliary output gradients while preserving the original DMD update.

3 Method

In this section, we present TSD (Figure 2), which aims to balance visual fidelity with coherent and plausible motion during continued distillation by making fuller use of teacher knowledge. Starting from the video DMD formulation (Sec. 3.1), we develop two complementary forms of supervision: temporal mean distillation to transfer persistent content (Sec. 3.2), and one-sided deviation protection to regulate temporal variation (Sec. 3.3). We then describe how TSD-D spatially guides deviation protection and shares its reference across training ranks (Sec. 3.4), before combining the components into a unified training objective and specifying both configurations (Sec. 3.5).

3.1 Video Distribution Matching

DMD matches a generator's distribution to a pretrained teacher through a score difference motivated by reverse-KL minimization (Yin et al., 2024b;a). In our setting, a few-step causal generator G_θ, with parameters θ, maps noise and text to a latent video $\hat{\boldsymbol{x}}_0 \in \mathbb{R}^{F\times C\times H\times W}$, with F frames, C channels, height H, and width W. We omit text conditioning from the notation below.

At a noise level τ drawn from the training distribution π, we obtain the noised video $\boldsymbol{x}_\tau = \alpha_\tau \hat{\boldsymbol{x}}_0 + \sigma_\tau \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is standard Gaussian noise and α_τ, σ_τ are the signal and noise coefficients. The frozen teacher, parameterized by ϕ, supplies the guided target score s_ϕ. An online diffusion critic,

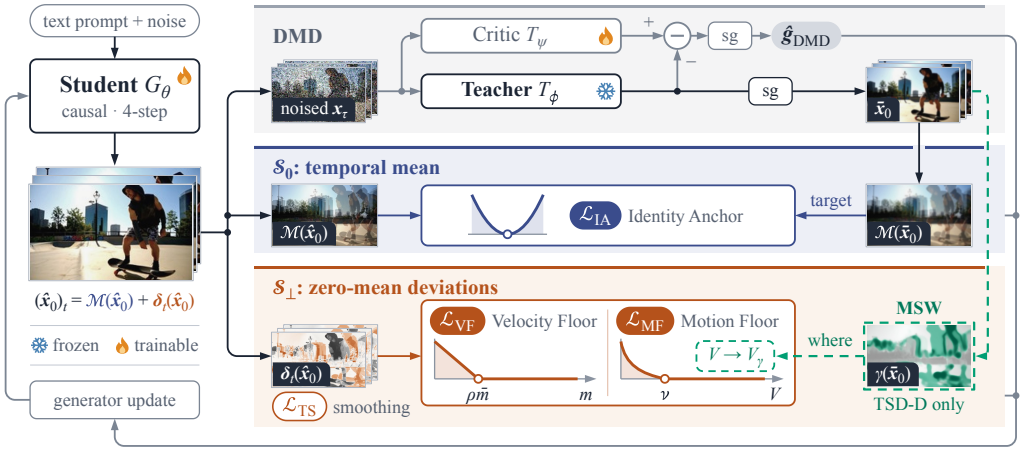

Figure 2: **Overview of TSD.** The few-step student generates a latent video, and the inherited DMD update acts on both temporal subspaces. The same detached teacher prediction supplies the temporal-mean target for the Identity Anchor (IA). In the zero-mean deviation subspace, Motion Floor (MF) and Velocity Floor (VF) protect the student's own variation, while Temporal Smoothing (TS) provides light smoothing. Solid paths are shared by TSD-B and TSD-D. The green path denotes TSD-D's motion-saliency weighting (MSW): teacher-derived spatial weights replace the temporal variance V with its weighted form V_γ in MF, indicating where to protect variation without prescribing a trajectory. The DMD direction and auxiliary loss gradients are then combined along the grey path to update the student.

parameterized by ψ, learns to denoise student-generated samples to estimate the student's score s_ψ. Scores are log-density gradients with respect to the noised input. The teacher and critic produce clean latent video estimates $T_\phi(\boldsymbol{x}_\tau, \tau)$ and $T_\psi(\boldsymbol{x}_\tau, \tau)$, respectively. Both denoise the same $\boldsymbol{x}_\tau$, while $\hat{\boldsymbol{x}}_0$ denotes the student-generated sample before corruption. We abbreviate these predictions as T_ϕ and T_ψ below. Following the DMD stage of Causal Forcing (Zhu et al., 2026), the normalized DMD direction $\widehat{\boldsymbol{g}}_{\text{DMD}}$ for one generated clip is

$$\widehat{\boldsymbol{g}}_{\text{DMD}} = \frac{T_\psi - T_\phi}{n} = \frac{\sigma_\tau^2}{\alpha_\tau n}\left[s_\psi - s_\phi\right], \qquad n = \frac{\|\hat{\boldsymbol{x}}_0 - T_\phi\|_1}{FCHW}, \tag{1}$$

where n is the mean absolute discrepancy between the student sample and the teacher estimate, and $\|\cdot\|_1$ sums absolute entries. The second equality follows from the score conversion given in Appendix A, which also specifies the guidance convention. We incorporate this normalized direction into the stop-gradient surrogate (Yin et al., 2024b;a)

$$\mathcal{L}_{\text{DMD}} = \frac{1}{2FCHW}\left\|\hat{\boldsymbol{x}}_0 - \text{sg}[\hat{\boldsymbol{x}}_0 - \widehat{\boldsymbol{g}}_{\text{DMD}}]\right\|_2^2, \qquad \nabla_{\hat{\boldsymbol{x}}_0}\mathcal{L}_{\text{DMD}} = \frac{\widehat{\boldsymbol{g}}_{\text{DMD}}}{FCHW}, \tag{2}$$

where $\text{sg}[\cdot]$ detaches its argument and $\|\cdot\|_2$ is the Frobenius norm. The entire surrogate target is held fixed, including the score predictions and n, so the per-sample scaling is applied only through $\widehat{\boldsymbol{g}}_{\text{DMD}}$. The resulting output gradient is backpropagated through G_θ. This video-level update provides joint supervision for persistent content and temporal variation, with limited explicit control over how each component changes during continued distillation. This motivates auxiliary objectives that transfer persistent content while protecting the student's temporal variation. TSD retains the DMD update and reuses T_ϕ for this targeted supervision.

3.2 Temporal Subspaces and Mean Distillation

Our goal is to transfer content shared across frames while leaving the student's temporal trajectory unconstrained by a framewise teacher target. We therefore use the temporal mean for content supervision and handle the remaining temporal variation separately. For a latent video $\boldsymbol{x}$ with $F \geq 2$ frames, let $\boldsymbol{x}_t$ denote its t-th frame. Its temporal mean $\mathcal{M}(\boldsymbol{x})$ and deviations $\boldsymbol{\delta}_t(\boldsymbol{x})$ are

$$\mathcal{M}(\boldsymbol{x}) = \frac{1}{F}\sum_{t=1}^{F}\boldsymbol{x}_t, \qquad \boldsymbol{\delta}_t(\boldsymbol{x}) = \boldsymbol{x}_t - \mathcal{M}(\boldsymbol{x}), \qquad \sum_{t=1}^{F}\boldsymbol{\delta}_t(\boldsymbol{x}) = \boldsymbol{0}.$$

Broadcasting the mean to all frames gives a time-constant video in the mean subspace $\mathcal{S}_0$. The full deviation sequence $\boldsymbol{\delta}(\boldsymbol{x}) = (\boldsymbol{\delta}_t(\boldsymbol{x}))_{t=1}^F$ lies in its orthogonal complement $\mathcal{S}_\perp$, the subspace of videos with zero temporal mean. Let Π_0 denote the mean-broadcasting projection and $\Pi_\perp = I - \Pi_0$ the deviation projection, where I is the identity operator.

To anchor persistent content in the mean subspace, we reuse the clean teacher prediction already computed for DMD. We make the student and teacher roles explicit by writing

$$\boldsymbol{x}_0^{(\mathrm{stu})} \equiv \hat{\boldsymbol{x}}_0, \qquad \boldsymbol{x}_0^{(\mathrm{tch})} \equiv \bar{\boldsymbol{x}}_0 = \mathrm{sg}[T_\phi(\boldsymbol{x}_\tau, \tau)]\,.$$

The *Identity Anchoring Loss* (IA) matches the student's temporal mean to that of the detached teacher target:

$$\mathcal{L}_{\mathrm{IA}}(\theta) = \frac{1}{CHW}\left\|\mathcal{M}\big(\boldsymbol{x}_0^{(\mathrm{stu})}\big) - \mathcal{M}\big(\boldsymbol{x}_0^{(\mathrm{tch})}\big)\right\|_2^2. \tag{3}$$

Because IA depends only on the temporal mean, every latent frame receives the same gradient. The output gradient therefore lies in $\mathcal{S}_0$ and has no component in the deviation subspace:

$$\Pi_\perp \nabla_{\boldsymbol{x}_0^{(\mathrm{stu})}} \mathcal{L}_{\mathrm{IA}} = \mathbf{0}.$$

IA thus provides targeted supervision of the mean component, but does not protect temporal deviations from other training updates. In particular, the unchanged DMD update still acts on both subspaces and can reduce the student's temporal variation. We therefore introduce separate objectives to resist this reduction in the next section.

3.3 Deviation Protection

We use one-sided floors to protect temporal variation. Each floor penalizes a chosen statistic only below its reference level, with zero penalty once that level is reached.

Motion Floor (MF). Mean supervision alone does not prevent temporal variation from shrinking during continued DMD training. MF applies a one-sided floor to the student's temporal variance, penalizing only values below a prescribed reference:

$$V(\boldsymbol{x}) = \frac{1}{FCHW}\sum_{t=1}^{F}\|\boldsymbol{\delta}_t(\boldsymbol{x})\|_2^2\,, \qquad \mathcal{L}_{\mathrm{MF}}(V_*) = \max\{0, -\log(V_* + \varepsilon) - \tau_{\mathrm{v}}\}\,. \tag{4}$$

Here $V(\boldsymbol{x})$ is the global temporal variance, V_* is the statistic used by a configuration, and $V_* = V(\hat{\boldsymbol{x}}_0)$ for the global floor. The constant $\varepsilon > 0$ ensures numerical stability. The hyperparameter τ_{v} is a threshold on negative log-variance, shared across prompts. Equivalently, it sets the variance reference $\nu = e^{-\tau_{\mathrm{v}}}$, with the penalty active only when $V_* + \varepsilon < \nu$. When active, an output-space descent step on the global MF term expands existing nonzero deviations along their current direction. Otherwise, its gradient vanishes.

Velocity Floor (VF). MF constrains the overall amount of temporal variation, but rapid frame-to-frame fluctuations can satisfy its variance reference even when sustained temporal change is weak. VF therefore protects changes that remain after temporal smoothing, measured across a longer frame interval. Let $\mathcal{B}$ be a linear temporal blur that averages neighboring latent frames with normalized Gaussian weights and replicated endpoints. The weights sum to one, so a time-constant signal remains unchanged. For a frame stride $1 \le s < F$, we measure the low-pass change of the deviation sequence as

$$m(\hat{\boldsymbol{x}}_0) = \frac{1}{s(F-s)CHW}\sum_{t=1}^{F-s}\left\|\big(\mathcal{B}\,\boldsymbol{\delta}(\hat{\boldsymbol{x}}_0)\big)_{t+s} - \big(\mathcal{B}\,\boldsymbol{\delta}(\hat{\boldsymbol{x}}_0)\big)_t\right\|_1. \tag{5}$$

The time-constant mean cancels in frame differences, so this expression is equivalent to using $\mathcal{B}\hat{\boldsymbol{x}}_0$ and depends only on temporal deviations. We estimate a reference $\bar{m}$ from detached student statistics using an exponential moving average with momentum 0.9. This reference follows early training and is frozen after T_0 generator updates so that it cannot decrease with later contraction. VF penalizes change below a fraction $\rho \in (0, 1]$ of this reference:

$$\mathcal{L}_{\mathrm{VF}}(\theta) = [\rho\bar{m} - m(\hat{\boldsymbol{x}}_0)]_+\,. \tag{6}$$

Here $[z]_+ = \max(z, 0)$ denotes the positive part of a scalar z. MF guards overall variation, while VF guards its low-pass component. Both operate on latent statistics, which do not directly measure visible displacement (Appendix D.1).

Temporal Smoothing (TS). VF does not directly penalize the rapid fluctuations removed by its temporal blur. TS complements the floor by penalizing this residual component:

$$\mathcal{L}_{\mathrm{TS}} = \frac{1}{FCHW} \left\| (I - \mathcal{B})\hat{\boldsymbol{x}}_0 \right\|_2^2 . \tag{7}$$

We use a small weight because this penalty can suppress fast coherent motion as well as flicker.

Pixel-Flow Floor (PFF). Variation in latent space does not necessarily translate into visible motion in decoded frames. As an extension, PFF applies this one-sided protection to pixel displacement estimated by optical flow. Let $\boldsymbol{y}$ be a sampled short window of frames decoded from $\hat{\boldsymbol{x}}_0$ using the frozen VAE decoder, and Φ a frozen RAFT optical-flow network (Teed & Deng, 2020). Let $\mathcal{P}$ contain the valid frame-pair indices in this window at frame interval s'. For each $t \in \mathcal{P}$, define $f_t = \text{mean top}_q \|\Phi(\boldsymbol{y}_t, \boldsymbol{y}_{t+s'})\|$, where the norm gives per-pixel Euclidean flow magnitudes and top_q retains their largest fraction $q \in (0, 1]$. With a displacement reference $\tau_{\mathrm{p}} > 0$ in pixels, PFF is

$$\mathcal{L}_{\mathrm{PFF}}(\theta) = \frac{1}{\tau_{\mathrm{p}} \left|\mathcal{P}\right|} \sum_{t \in \mathcal{P}} \left[\tau_{\mathrm{p}} - f_t\right]_+ , \tag{8}$$

where $|\mathcal{P}|$ is the number of frame pairs. The loss averages their displacement shortfalls normalized by τ_{p}. A pair contributes no penalty once its flow statistic reaches this reference. Gradients pass through the frozen decoder and flow network to the selected latent frame.

3.4 Teacher Guidance and Reference Synchronization

To make better use of teacher predictions and improve training stability, we extend the preceding formulation with two mechanisms. MSW uses teacher-derived motion cues to spatially weight MF, while CRS synchronizes the reference used by VF across training ranks. We call the base formulation TSD-B (*Balanced*) and this extension TSD-D (*Dynamic*).

Teacher Motion-Saliency Weighting (MSW). A video may contain moving subjects alongside regions that should remain static. Global MF weights all locations equally, so fluctuations in otherwise static regions can help satisfy its floor. MSW therefore uses temporal differences in the teacher prediction as a soft spatial mask, giving greater weight to regions where the teacher predicts change. Averaging these weights over time guides where variation is protected without prescribing its direction or timing. For spatial location (h, w) and frame $t \geq 2$, let $d_{t,h,w}$ be the channel-averaged teacher change, with $d_{1,h,w} = d_{2,h,w}$. We construct the detached spatial weight $\gamma_{h,w}$ as

$$\begin{aligned} d_{t,h,w}(\bar{\boldsymbol{x}}_0) &= \frac{1}{C} \left\| (\bar{\boldsymbol{x}}_0)_{t,:,h,w} - (\bar{\boldsymbol{x}}_0)_{t-1,:,h,w} \right\|_1 , \qquad \hat{d} = \frac{1}{FHW} \sum_{t,h,w} d_{t,h,w}(\bar{\boldsymbol{x}}_0), \\ \gamma_{h,w} &= \frac{1}{F} \sum_{t=1}^{F} \mathrm{sigmoid}\left(\kappa \left[\frac{d_{t,h,w}(\bar{\boldsymbol{x}}_0)}{\hat{d} + \varepsilon_d} - b \right] \right) . \end{aligned} \tag{9}$$

The colon indexes all channels, $\hat{d}$ is the mean change over frames and locations, and the sigmoid maps normalized change to weights between zero and one. The scalars $\kappa > 0$ and b control weight sharpness and selectivity, and $\varepsilon_d > 0$ stabilizes the normalization. These weights define the saliency-weighted variance and its corresponding MF loss:

$$\begin{aligned} V_\gamma(\hat{\boldsymbol{x}}_0) &= \frac{\sum_{t,c,h,w} \gamma_{h,w} \boldsymbol{\delta}_{t,c,h,w}(\hat{\boldsymbol{x}}_0)^2}{(F-1)C \left(\sum_{h,w} \gamma_{h,w} + \varepsilon_\gamma \right)}, \\ \mathcal{L}_{\mathrm{MF}}\big(V_\gamma(\hat{\boldsymbol{x}}_0)\big) &= \max\big\{0, -\log\big(V_\gamma(\hat{\boldsymbol{x}}_0) + \varepsilon\big) - \tau_{\mathrm{v}}\big\} . \end{aligned} \tag{10}$$

Here $\boldsymbol{\delta}_{t,c,h,w}(\hat{\boldsymbol{x}}_0)$ denotes an entry of the deviation tensor, $\varepsilon_\gamma > 0$ stabilizes the weighted normalization, and $F - 1$ follows the sample-variance convention used by MSW. The weighted floor replaces the global MF term in TSD-D (Sec. 3.5), with implementation details in Appendix A.

Cross-Rank Synchronization (CRS). Different data-parallel ranks observe different samples during warm-up and can estimate different VF references. In TSD-D, CRS averages their detached statistics before the update of $\bar{m}$, giving every rank the same reference. TSD-B retains a separate reference on each rank. Appendix A specifies the averaging rule.

Table 1: Main comparison on Wan2.1-T2V, 14B teacher and 1.3B student. Results are mean$_{\pm\text{sample std}}$ over three training seeds. ΔVB1 is the change in VBench-1 mean from 1000 to 3000 iterations. INIT evaluates the shared starting checkpoint once.

Method	VBench-1 ↑ 1K	VBench-1 ↑ 3K	VBench-2 ↑ 1K	VBench-2 ↑ 3K	ΔVB1	dyn. degree 1K	dyn. degree 3K
INIT (step 0, n=1)	84.31		44.87		—	0.82	
DMD	$86.43_{\pm 0.61}$	$80.82_{\pm 0.62}$	$50.40_{\pm 1.41}$	$51.98_{\pm 1.33}$	-5.61	$0.76_{\pm 0.21}$	$0.17_{\pm 0.09}$
AVD	$81.80_{\pm 1.86}$	$81.39_{\pm 2.50}$	$48.17_{\pm 2.01}$	$50.12_{\pm 1.44}$	-0.41	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$
AVD-OFFICIAL	$84.17_{\pm 3.67}$	$82.95_{\pm 1.54}$	$49.73_{\pm 3.22}$	$51.44_{\pm 3.33}$	-1.22	$0.99_{\pm 0.01}$	$0.73_{\pm 0.28}$
TSD-B	$87.00_{\pm 0.42}$	$84.97_{\pm 1.13}$	$\mathbf{52.69}_{\pm 1.71}$	$\mathbf{53.93}_{\pm 1.04}$	-2.03	$0.98_{\pm 0.02}$	$0.71_{\pm 0.10}$
TSD-D	$\mathbf{87.22}_{\pm 0.20}$	$\mathbf{86.86}_{\pm 0.48}$	$52.50_{\pm 0.73}$	$52.05_{\pm 1.48}$	-0.36	$0.99_{\pm 0.01}$	$0.85_{\pm 0.05}$

3.5 Unified Training Objective

The generator combines distribution matching, temporal mean anchoring, and deviation protection:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{DMD}} + \lambda_{\text{IA}}\mathcal{L}_{\text{IA}} + \lambda_{\text{MF}}\mathcal{L}_{\text{MF}}(V_*) + \lambda_{\text{VF}}\mathcal{L}_{\text{VF}} + \lambda_{\text{TS}}\mathcal{L}_{\text{TS}} + \lambda_{\text{PFF}}\mathcal{L}_{\text{PFF}}. \tag{11}$$

Each floor penalizes a statistic below its reference, but optimizing the combined objective does not enforce that reference as a hard constraint. The DMD loss is unchanged from Section 3.1.

Configurations. Both variants use Eq. (11), with TSD-B as our default. Their choices of MF statistic and VF reference are

$$(V_*, \bar{m}) = \begin{cases} \big(V(\hat{\boldsymbol{x}}_0), \bar{m}_{\text{local}}\big), & \text{for TSD-B,} \\ \big(V_\gamma(\hat{\boldsymbol{x}}_0), \bar{m}_{\text{shared}}\big), & \text{for TSD-D,} \end{cases}$$

where $\bar{m}_{\text{local}}$ denotes the VF reference maintained separately on each data-parallel rank, and $\bar{m}_{\text{shared}}$ denotes the common reference estimated using CRS. Thus TSD-D uses teacher-derived spatial cues to guide where variation is protected and synchronizes the VF reference across ranks. Both configurations use the same auxiliary weights. All components operate during training, and both configurations retain the same student architecture and inference procedure.

4 Experiments

4.1 Experimental Setup

We use Wan2.1-T2V-14B (Wan et al., 2025) as a frozen bidirectional teacher and a 1.3B causal student initialized from the stage-2 ODE checkpoint of Causal Forcing (Zhu et al., 2026). The student generates videos autoregressively with 4 denoising steps per latent frame, producing 81 frames at 16 fps and a resolution of 832×480. Training uses filtered text prompts from VidProM (Wang & Yang, 2024). All methods share the teacher, student initialization, and prompt set.

We train for 3000 iterations with a total batch size of 8 across eight GPUs, updating the critic at every iteration and the student once every five iterations. We use AdamW with learning rates of 2×10^{-6} for the student and 4×10^{-7} for the critic, and maintain a student exponential moving average with decay 0.99. The teacher uses guidance scale 3.0 under the Causal Forcing convention. Both TSD variants use $\lambda_{\text{IA}} = 0.075$, $\lambda_{\text{MF}} = 0.05$, $\lambda_{\text{VF}} = 0.1$, and $\lambda_{\text{TS}} = 0.02$ in Eq. (11). TSD-D additionally enables MSW and CRS. Appendix A provides the remaining hyperparameters and implementation details.

We evaluate with VBench-1 (Huang et al., 2024) and VBench-2 (Zheng et al., 2025) using their official prompts and sampling protocols. The baselines are vanilla DMD and two prompt-only adaptations of AVD (You et al., 2026): always-on temporal-variance regularization (AVD) and its gated variant (AVD-OFFICIAL).

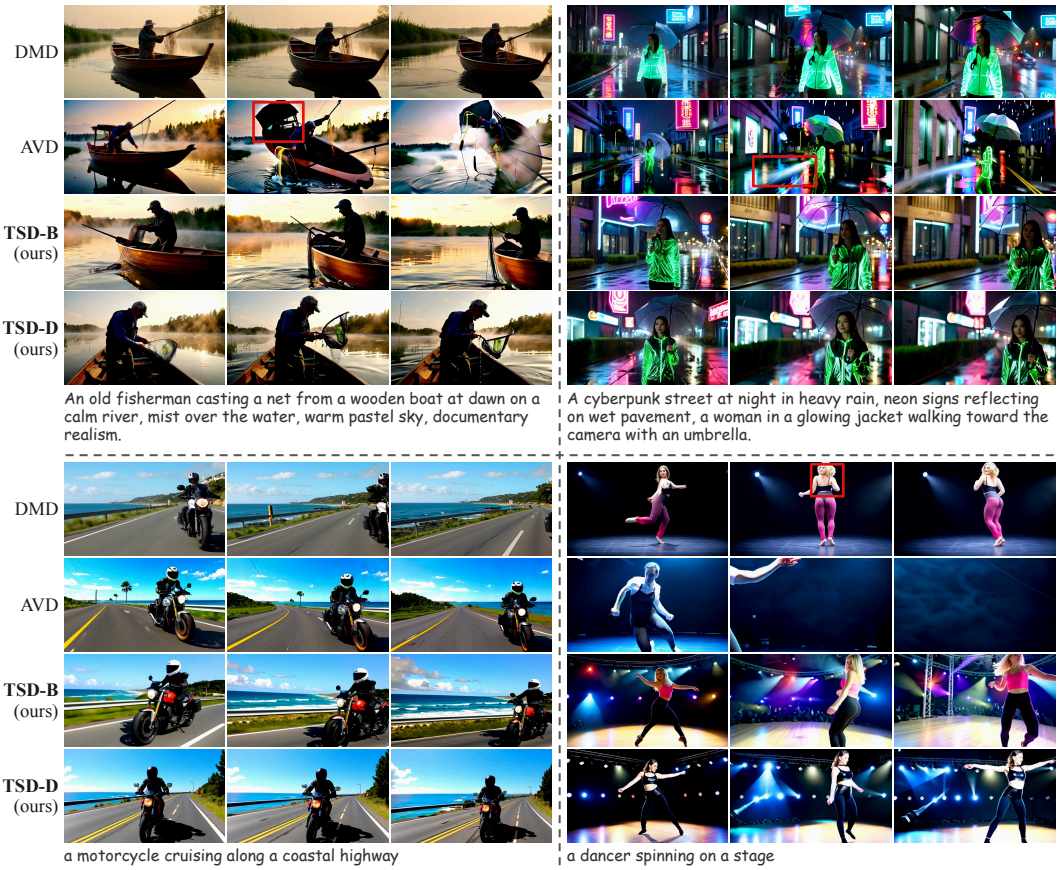

Figure 3: Qualitative comparisons with baseline methods. Each method row shows three frames in temporal order, with the corresponding prompt below each scene. The upper panels use long scene descriptions, and the lower panels use short action prompts.

4.2 Quantitative Evaluation

Table 1 compares TSD with vanilla DMD and the two AVD variants under the same training and inference settings. At 3K, both TSD configurations achieve higher scores than all three baselines on VBench-1 and VBench-2. From 1K to 3K, TSD improves visual fidelity without noticeable motion degradation. In contrast, AVD maintains a high dynamic degree but can introduce spurious motion even in regions that should remain static. These results demonstrate the effectiveness of combining distribution matching with separate supervision of persistent content and temporal variation.

4.3 Qualitative Evaluation

Figure 3 compares visual quality and motion behavior under both long and short prompts. In the fisherman example, vanilla DMD produces an almost static video, with little change in the fisherman's pose. AVD generates larger changes but introduces visible distortions in the fisherman and boat. In contrast, TSD produces coherent movement while preserving their structure. For the long cyberpunk prompt, TSD keeps the umbrella moving coherently with the person walking toward the camera, preserving their relative position as the scene evolves.

The short motorcycle and dancer prompts further illustrate the quality of the generated actions. In the motorcycle example, TSD generates plausible forward motion while maintaining consistent lane markings across frames. The dancer example involves complex human motion, where vanilla DMD keeps the face directed toward the camera even when the torso turns away. TSD instead maintains coordinated head and body movements, producing a more natural turning motion. Together, these examples show that TSD handles both detailed scene descriptions and concise instructions while preserving coherent subject structure and relationships between moving objects.

Table 2: **Ablation.** Three training seeds with ODE initialization.

Configuration	VB1↑	VB2↑
DMD	$80.82_{\pm 0.62}$	$51.98_{\pm 1.33}$
+ IA	$82.54_{\pm 1.62}$	$50.55_{\pm 0.6}$
+ MF	$83.81_{\pm 3.69}$	$52.80_{\pm 0.6}$
+ IA + MF	$\mathbf{85.65}_{\pm 1.16}$	$52.28_{\pm 1.2}$
+ IA + MF + VF	$83.64_{\pm 2.36}$	$51.78_{\pm 1.2}$
+ IA + MF + TS	$81.70_{\pm 4.65}$	$50.64_{\pm 5.0}$
TSD-B	$84.97_{\pm 1.13}$	$\mathbf{53.93}_{\pm 1.04}$

Table 3: **Transfer across budgets.** CF++ initialization with four and two denoising steps.

	4 steps		2 steps	
Method	**VB1↑**	**VB2↑**	**VB1↑**	**VB2↑**
DMD	$82.47_{\pm 2.17}$	$50.73_{\pm 1.1}$	$78.24_{\pm 3.12}$	$52.31_{\pm 1.1}$
AVD	$73.57_{\pm 7.96}$	$47.20_{\pm 5.5}$	$81.89_{\pm 0.20}$	$49.33_{\pm 3.1}$
AVD-OFFICIAL	$80.03_{\pm 4.73}$	$52.04_{\pm 2.6}$	$83.56_{\pm 1.18}$	$49.44_{\pm 1.1}$
TSD-B	$\mathbf{85.36}_{\pm 0.39}$	$52.48_{\pm 1.3}$	$\mathbf{84.29}_{\pm 1.36}$	$52.32_{\pm 0.6}$
TSD-D	$83.97_{\pm 0.34}$	$\mathbf{54.08}_{\pm 2.1}$	$82.78_{\pm 2.31}$	$\mathbf{52.75}_{\pm 2.1}$

4.4 Generality

TSD is a general supervision framework that can augment different video distillation settings. We examine this flexibility by applying the same auxiliary objectives to the consistency-distillation initialization (CF++) and evaluating four- and two-step inference. As shown in Table 3, both TSD configurations achieve higher VBench-1 and VBench-2 scores than vanilla DMD at both sampling budgets. These results support the general applicability of TSD across different distillation settings.

4.5 Ablation

Table 2 evaluates the main components of TSD. Combining IA and MF improves VBench-1 over either term alone, while jointly using VF and TS outperforms adding either one alone on both benchmarks. A critic-free control also degrades visual quality, supporting our choice to retain the original DMD update (Appendix D.3). Additional MSW and CRS ablations are reported in Appendix B.1.

4.6 Longer Training

Table 4 reports results at 3K, 5K, and 10K iterations for one run of each configuration. Beyond 3K, the metrics fluctuate without consistent improvement. TSD-B changes little overall, while TSD-D improves at 5K before declining at 10K. Visual inspection likewise reveals no obvious improvement in generated video quality beyond 3K. These observations suggest diminishing returns from extending training under the current distillation setup.

Table 4: **Longer training.** Each cell reports VBench-1 / VBench-2 / dynamic degree.

Method	3K	5K	10K
TSD-B	86.20 / 52.68 / 0.87	86.07 / 52.61 / 0.92	86.69 / 52.28 / 0.90
TSD-D	86.49 / 52.93 / 0.93	86.26 / 54.08 / 0.98	83.13 / 51.62 / 0.73

5 Conclusion

We presented Temporal Subspace Distillation (TSD), which augments DMD with separate supervision of persistent content and temporal variation in orthogonal temporal subspaces. Experiments on Wan2.1 show that TSD produces videos with high visual fidelity and coherent motion, with qualitative comparisons further illustrating natural actions and consistent object relationships. These benefits require no changes to student inference.

Limitations. Training stability remains a limitation of TSD, inherited from the underlying DMD formulation. Potential contributors include the mode-seeking tendency of reverse-KL matching (Xu et al., 2025) and errors in the online critic as it tracks the changing student distribution (Yin et al., 2024a). The coupled student and critic updates pose an optimization challenge analogous to GAN training. Stabilizing this coupled optimization while preserving the benefits of distribution matching remains an open direction.

AI USE STATEMENT

We used generative AI tools to polish the manuscript text and to identify recent related work, helping us avoid overlooking relevant studies. We also used Qwen2.5-7B-Instruct to draft prompt variants for the prompt-detail experiment. These tools also assisted with implementing parts of the evaluation pipeline and benchmark code, as well as setting up and migrating complex experimental software environments across computing systems. The authors conceived the research ideas and wrote the core research code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We will publicly release our code upon acceptance. Please also see the supplementary material for more details.

ETHICS STATEMENT

None.

REFERENCES

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 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.

Yukang Chen, Luozhou Wang, Wei Huang, Shuai Yang, Bohan Zhang, Yicheng Xiao, Ruihang Chu, Weian Mao, Qixin Hu, Shaoteng Liu, et al. Longlive-2.0: An nvfp4 parallel infrastructure for long video generation. *arXiv preprint arXiv:2605.18739*, 2026.

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. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

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.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

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. *arXiv preprint arXiv:2303.01469*, 2023.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

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.

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.

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*, 2025a.

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. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025b.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 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 *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. *arXiv preprint arXiv:2603.21864*, 2026.

Yuanhao Zhai, Kevin Lin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Chung-Ching Lin, David Doermann, Junsong Yuan, and Lijuan Wang. Motion consistency model: Accelerating video diffusion with disentangled motion-appearance distillation. *Advances in Neural Information Processing Systems*, 37:111000–111021, 2024.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

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*, 2026.

APPENDIX

A TRAINING DETAILS

Backbone and distillation. We follow the stage-3 DMD recipe of Causal Forcing (Zhu et al., 2026) on Wan2.1-T2V (Wan et al., 2025). The frozen teacher is the bidirectional 14B model, while the causal student and online critic are both 1.3B. The student generates one frame per causal block in four denoising steps $\{1000, 750, 500, 250\}$ with timestep shift 5.0. The latent shape is $F \times C \times H \times W = 21 \times 16 \times 60 \times 104$, in frame, channel, height, and width order, corresponding to approximately 81 decoded frames at 480×832 resolution.

DMD scores and guidance. The teacher and critic scores in Eq. (1) are $s_\phi = (\alpha_\tau T_\phi - \boldsymbol{x}_\tau)/\sigma_\tau^2$ and $s_\psi = (\alpha_\tau T_\psi - \boldsymbol{x}_\tau)/\sigma_\tau^2$. The teacher uses classifier-free guidance $T_\phi = T_c + g(T_c - T_{\text{neg}})$ with $g = 3$, where T_c and T_{neg} are clean predictions conditioned on the input and negative prompts, respectively. The critic uses the input prompt without guidance. Both scores and the normalizer n are detached during the generator update.

Optimization. We train with AdamW using $(\beta_1, \beta_2) = (0, 0.999)$. The generator and critic learning rates are 2×10^{-6} and 4×10^{-7}. The generator EMA begins at step 200 with weight 0.99.

Objective configuration. Both configurations use $\lambda_{\text{IA}} = 0.075$, $\lambda_{\text{MF}} = 0.05$, $\lambda_{\text{VF}} = 0.1$, and $\lambda_{\text{TS}} = 0.02$. MF uses $\tau_{\text{v}} = 0.8$ and $\varepsilon = 10^{-6}$. VF uses $\rho = 0.9$, a warm-up of $T_0 = 200$ generator updates, stride $s = 4$, and a temporal Gaussian kernel of size 5 with width $\sigma_{\mathcal{B}} = 1.5$ frames. TSD-D additionally enables MSW with $\kappa = 3.0$ and $b = 1.1$, together with CRS. The prompt-only AVD (You et al., 2026) baselines use $-\log(V + \varepsilon)$ with weight 0.05. AVD-OFFICIAL gates this term at $\tau_{\text{v}} = 0.8$, and neither variant includes real-video regression.

Relation to AVD. Global MF reformulates AVD's gated temporal-variance regularizer (You et al., 2026) as a hinge penalty. Let $a = -\log(V + \varepsilon)$. The official AVD implementation applies a when $a \geq \tau_{\text{v}}$ and zero otherwise, whereas our formulation uses $\max(0, a - \tau_{\text{v}})$. The hinge approaches zero continuously as the variance reaches its reference, while the gated loss drops discontinuously to zero once the reference is exceeded. TSD combines this floor with teacher supervision restricted to the temporal mean, low-pass deviation protection, and temporal smoothing. TSD-D further uses teacher-derived spatial weighting and a synchronized VF reference. The key distinction therefore lies in how TSD organizes and extends auxiliary supervision across temporal subspaces.

VF warm-up reference. The VF loss in Eq. (6) uses a reference $\bar{m}$ estimated during warm-up and frozen afterward. Let $m_k^{(r)}$ be the low-pass temporal change statistic in Eq. (5) at generator update k on rank r. In TSD-D, CRS averages the detached statistics across R ranks to update the shared reference:

$$\begin{aligned} m_{\text{pool},k} &= \frac{1}{R}\sum_{r=1}^{R} \text{sg}\left[m_k^{(r)}\right], \\ \bar{m}_1 &= m_{\text{pool},1}, \\ \bar{m}_k &= \beta \bar{m}_{k-1} + (1-\beta) m_{\text{pool},k}, \qquad 2 \leq k \leq T_0, \\ \bar{m}_k &= \bar{m}_{T_0}, \qquad k > T_0. \end{aligned} \tag{12}$$

Here $m_{\text{pool},k}$ is the cross-rank average, $\bar{m}_k$ is the reference at generator update k, $\text{sg}[\cdot]$ denotes stop-gradient, and $\beta = 0.9$ is the EMA coefficient. TSD-B applies the same recursion independently on each rank using $\text{sg}\left[m_k^{(r)}\right]$. The VF loss uses the current $\bar{m}_k$, including during warm-up.

Numerical details. The MSW variance in Eq. (10) uses the sample-variance normalization $F - 1$. Uniform weights recover $\frac{F}{F-1} V(\hat{\boldsymbol{x}}_0)$ up to the denominator stabilizer, so at the same τ_{v} the weighted floor activates at a global variance about 5% lower for $F = 21$. Static teacher predictions produce uniform weights and do not disable MF. We set $\varepsilon_\gamma = \varepsilon/C = 10^{-6}/C$ and $\varepsilon_d = 10^{-6}$. The first frame difference is copied from the second, giving $d_{1,h,w} = d_{2,h,w}$ at each spatial location (h, w). Temporal Gaussian blurs use replicate padding and leave clips shorter than the kernel unchanged.

B Additional Experiments

B.1 Ablation of TSD-D

We examine how the two additions in TSD-D contribute to its performance. MSW uses teacher-derived motion saliency to guide where temporal variation is protected, while CRS shares the VF reference across data-parallel ranks. Table 5 compares enabling these components individually and jointly on top of TSD-B, using the same ODE initialization and auxiliary weights, with three training seeds per configuration.

With CRS enabled, adding MSW improves both benchmark scores. MSW alone does not yield the same benefit, supporting its joint use with CRS in TSD-D. The combined configuration achieves the highest VBench-1 score in this ablation.

Table 5: Ablation of MSW and CRS on the shared ODE initialization at 3K iterations. Scores are averaged over three training seeds and reported with sample standard deviations. The TSD-B row repeats Table 2.

Configuration	**VBench-1** ↑	**VBench-2** ↑
TSD-B	84.97 ± 1.13	53.93 ± 1.04
TSD-B + MSW	83.71 ± 2.50	52.16 ± 2.2
TSD-B + CRS	85.89 ± 0.15	51.60 ± 0.7
TSD-B + MSW + CRS = TSD-D	86.86 ± 0.48	52.05 ± 1.48

B.2 Per-seed Results

Table 6 reports the individual VBench-1 and VBench-2 scores for the three training runs of each method at 1K and 3K iterations. At 3K, all three runs of each TSD configuration exceed every vanilla DMD run on VBench-1, showing that this advantage is shared across the reported runs.

Table 6: Per-seed VBench-1 and VBench-2 scores underlying Table 1. Each row lists the three training runs in the same order (s0/r1/r2) at both checkpoints.

Method	**step**	**VBench-1** (s0/r1/r2)	**VBench-2** (s0/r1/r2)
DMD	3000	81.45 / 80.81 / 80.21	51.87 / 50.70 / 53.36
AVD	3000	81.24 / 83.96 / 78.97	50.26 / 51.48 / 48.62
AVD-OFFICIAL	3000	81.95 / 84.73 / 82.18	53.58 / 53.14 / 47.60
TSD-B	3000	86.22 / 84.66 / 84.03	55.04 / 52.97 / 53.79
TSD-D	3000	87.20 / 87.08 / 86.31	53.13 / 50.36 / 52.67
DMD	1000	85.73 / 86.72 / 86.85	51.39 / 48.78 / 51.03
AVD	1000	81.73 / 79.98 / 83.70	46.03 / 48.45 / 50.03
AVD-OFFICIAL	1000	87.76 / 84.34 / 80.42	52.60 / 50.35 / 46.25
TSD-B	1000	86.83 / 86.70 / 87.48	51.31 / 52.15 / 54.60
TSD-D	1000	87.08 / 87.44 / 87.13	53.21 / 52.54 / 51.76

C Detailed Benchmark Results

C.1 VBench-1

Table 7 shows that both TSD configurations achieve higher Quality scores than the baselines at 3K iterations, contributing to higher overall VBench-1 scores. Scores are multiplied by 100 for reporting.

Table 7: **VBench-1 category scores.**

Method	**Quality** ↑	**Semantic** ↑	**Total** ↑
DMD	83.94	**68.36**	80.82
AVD	87.57	56.65	81.39
AVD-OFFICIAL	87.51	64.71	82.95
TSD-B	89.28	67.72	84.97
TSD-D	**91.56**	68.06	**86.86**

C.2 VBench-2

Table 8 reports scores for all 18 VBench-2 dimensions, averaged over three training seeds. Dynamic Attribute, Mechanics, and Thermotics are higher than for vanilla DMD in every configuration with a variance term, although the differences are not uniform across the remaining dimensions. Always-on AVD, the configuration with the highest dynamic degree, also has the lowest multi-view consistency, while TSD-B nearly matches vanilla DMD on this dimension and TSD-D lies between them. This pattern exposes an empirical tension in the tested configurations, but it does not establish a universal anti-correlation between motion and consistency.

Table 8: Dimension-level VBench-2 results (0 to 100) at step 3000, averaged over three training seeds. MVC denotes Multi-View-Consistency. AVD-OFF. is AVD-OFFICIAL. No entry is bolded.

Dimension	DMD	AVD	AVD-OFF.	TSD-B	TSD-D
Camera Motion	20.0	27.2	25.6	25.8	25.0
Complex Landscape	18.7	18.1	17.0	18.1	19.1
Complex Plot	11.3	8.9	10.0	10.6	10.6
Composition	36.0	38.1	36.1	37.2	36.3
Diversity	58.8	65.2	70.4	69.7	61.6
Dynamic Attribute	4.8	15.6	9.6	9.5	9.6
Dynamic Spatial Rel.	29.3	19.8	30.4	29.1	33.5
Human Anatomy	86.1	84.9	84.4	83.7	86.7
Human Clothes	99.8	93.2	92.2	99.7	98.9
Human Identity	94.3	77.5	80.9	88.0	88.9
Human Interaction	49.7	39.2	46.8	57.7	53.4
Instance Preservation	97.7	79.1	87.7	92.2	92.2
Material	32.7	38.1	36.3	36.6	32.3
Mechanics	41.9	59.7	54.4	52.6	49.0
Motion Order Underst.	9.0	11.4	7.9	7.9	10.3
Motion Rationality	21.3	27.4	23.6	24.3	22.4
MVC	38.0	7.9	22.0	37.7	27.7
Thermotics	44.2	55.9	53.0	52.4	48.7

D Temporal Subspace Analysis

D.1 Subspace properties

We first establish how the auxiliary objectives act on the temporal subspaces defined in Section 3.2. For video latents $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{F\times C\times H\times W}$, the mean-broadcasting projector Π_0 and deviation projector $\Pi_\perp = I - \Pi_0$ satisfy

$$\begin{aligned} \boldsymbol{x} &= \Pi_0\boldsymbol{x} + \Pi_\perp\boldsymbol{x}, \\ \langle \Pi_0\boldsymbol{x}, \Pi_\perp\boldsymbol{y}\rangle &= 0, \\ \|\boldsymbol{x}\|_2^2 &= F\|\mathcal{M}(\boldsymbol{x})\|_2^2 + \|\boldsymbol{\delta}(\boldsymbol{x})\|_2^2. \end{aligned} \tag{13}$$

Here F is the number of latent frames, C the number of latent channels, and H, W the latent height and width. The operator I is the identity, $\mathcal{M}(\boldsymbol{x})$ is the temporal mean, and $\boldsymbol{\delta}(\boldsymbol{x})$ is the zero-mean deviation sequence. The factor F accounts for repeating the mean across all frames. All norms and inner products here are Frobenius. The same decomposition applies to the sampled DMD direction in Eq. (1):

$$\widehat{\boldsymbol{g}}_{\mathrm{DMD}} = \Pi_0\widehat{\boldsymbol{g}}_{\mathrm{DMD}} + \Pi_\perp\widehat{\boldsymbol{g}}_{\mathrm{DMD}}. \tag{14}$$

Gradients of the auxiliary objectives. For the student sample $\hat{\boldsymbol{x}}_0$ and detached teacher prediction $\bar{\boldsymbol{x}}_0$, IA and the active global MF term have output gradients

$$\begin{aligned} \nabla_{\hat{\boldsymbol{x}}_0}\mathcal{L}_{\mathrm{IA}} &= \frac{2}{FCHW}\Pi_0(\hat{\boldsymbol{x}}_0 - \bar{\boldsymbol{x}}_0), \\ \nabla_{\hat{\boldsymbol{x}}_0}\mathcal{L}_{\mathrm{MF}} &= -\frac{2\boldsymbol{\delta}(\hat{\boldsymbol{x}}_0)}{FCHW\big(V(\hat{\boldsymbol{x}}_0)+\varepsilon\big)}. \end{aligned} \tag{15}$$

The first lies in the time-constant subspace $\mathcal{S}_0$, while the second lies in the zero-mean subspace $\mathcal{S}_\perp$. The variance V and stabilizer ε are defined in Eq. (4). An output-space descent step on active MF expands existing deviations, but its gradient is zero at an exactly time-constant video.

The same subspace property extends to MSW, VF, and TS. Hold the teacher-derived weights and warm-up references fixed, as in the implementation, and let $\mathbf{c} \in \mathcal{S}_0$ be any offset shared by all frames. Then $\boldsymbol{\delta}(\hat{\boldsymbol{x}}_0 + \mathbf{c}) = \boldsymbol{\delta}(\hat{\boldsymbol{x}}_0)$, so global and saliency-weighted MF and VF are unchanged. Since the temporal blur $\mathcal{B}$ preserves time-constant signals, $(I - \mathcal{B})\mathbf{c} = 0$, and TS is also unchanged. Thus each of these latent objectives $\mathcal{J}$ satisfies

$$\mathcal{J}(\hat{\boldsymbol{x}}_0 + \mathbf{c}) = \mathcal{J}(\hat{\boldsymbol{x}}_0), \qquad \Pi_0\nabla_{\hat{\boldsymbol{x}}_0}\mathcal{J} = \mathbf{0}. \tag{16}$$

The gradient identity follows by differentiating along every direction in $\mathcal{S}_0$ and also holds for the subgradients used at hinges and absolute values. These objectives therefore have output gradients orthogonal to IA.

Connection to score distillation. Replacing the critic's clean prediction T_ψ by $\hat{\boldsymbol{x}}_0$ in the surrogate of Section 3.1 gives the critic-free output gradient $\boldsymbol{g}_{\mathrm{CF}}$. Its direction corresponds to the clean-prediction form of score distillation (Poole et al., 2022), and

$$\boldsymbol{g}_{\mathrm{CF}} = \frac{\hat{\boldsymbol{x}}_0 - \bar{\boldsymbol{x}}_0}{nFCHW}, \qquad \Pi_0\boldsymbol{g}_{\mathrm{CF}} = \frac{1}{2n}\nabla_{\hat{\boldsymbol{x}}_0}\mathcal{L}_{\mathrm{IA}}. \tag{17}$$

Here n is the per-sample normalizer in Eq. (1), and both the target and n are detached. IA supplies this mean-directed supervision with a fixed loss weight, while the original DMD term retains the critic correction in both subspaces.

Variation and motion. Any permutation of the frames leaves V unchanged, even when it disrupts a coherent action. A variance floor therefore cannot by itself enforce temporal order or plausible motion. VF measures ordered changes after temporal smoothing, while TS penalizes the fluctuations removed by that smoothing. These latent statistics still do not directly measure displacement in decoded frames, motivating the pixel-space extension in Section 3.3.

Keeping the DMD term unchanged does not preserve every stationary point of the combined objective. IA, active floors, and TS can all contribute additional gradients. The floors become inactive when their references are met, but they remain soft penalties rather than hard constraints.

D.2 PROMPT DETAIL

We examine how prompt detail affects generated motion, separating the effect of prompt length from that of explicit motion instructions. For the first 36 prompts of the VBench-1 dynamic-degree suite, Table 9 compares four versions. S is the original short prompt. R repeats the original sentence to approximately match the length of D-m. D-s adds appearance and scene detail while retaining the original action, without adding further motion or camera descriptions. D-m additionally specifies how the subject and camera move over time. We generate the detailed rewrites with Qwen2.5-7B-Instruct and manually correct one rewrite that switched language.

Each student checkpoint uses three sampling seeds per prompt, giving 108 clips per version. The trained checkpoints come from one training run per configuration on the second cluster, with the shared initialization also included for reference. The 14B teacher uses 50 sampling steps and guidance scale 5, with 108 clips for S and 36 clips using sampling seed 0 for each other version.

The table reports dynamic degree on a 0 to 1 scale. Flow measures displacement in pixels using RAFT on consecutive-frame pairs sampled every eight frames. For each clip, we average the top-5% flow magnitude across these pairs, then report the median across clips. The "cam." columns count clips with spatially widespread, directionally consistent flow. A clip is counted when its pair-averaged ratio of mean to top-5% flow exceeds 0.6 and its direction coherence exceeds 0.7. Direction coherence is the fraction of moving pixels oriented within approximately 20° of the median flow direction.

At every reported 3K student checkpoint, repetition and appearance detail increase dynamic degree relative to S, while D-m produces the highest score. Explicit motion instructions therefore contribute beyond simply lengthening the prompt. For example, D-m increases TSD-D's dynamic degree from 0.52 to 0.94 and its median flow from 4.5 to 21.0 pixels. The 1K and 3K comparisons also show that vanilla DMD and TSD-B lose flow under both S and D-m. Richer motion descriptions increase generated movement, but do not fully prevent its decline during training in these runs.

Table 9: **Effect of prompt detail on motion.**

	dynamic degree				flow (px)		cam.	
Checkpoint	S	R	D-s	D-m	S	D-m	S	D-m
14B teacher	0.75	0.83	0.69	0.94	20.9	35.5	11	5
INIT (shared, step 0)	0.89	0.94	0.94	1.00	21.2	44.8	16	11
DMD, 1K	0.93	0.93	0.93	1.00	14.7	46.8	9	3
DMD, 3K	0.37	0.57	0.61	0.88	3.9	14.7	40	19
TSD-B, 1K	1.00	1.00	1.00	1.00	11.2	22.5	17	7
TSD-B, 3K	0.41	0.56	0.59	0.85	4.8	8.3	62	30
TSD-D, 3K	0.52	0.66	0.74	0.94	4.5	21.0	24	0
TSD-B, CF++ init, 3K	0.87	0.94	0.98	1.00	10.0	49.2	5	1

D.3 CRITIC-FREE CONTROL

The coupled student and critic updates in DMD pose an optimization challenge analogous to GAN training. We examine whether removing the critic offers a viable alternative. In one TSD-B run on the second cluster, we replace T_ψ with the student's sample $\hat{x}_0$, disable critic training, and retain the auxiliary losses. Table 10 compares this control with a critic-equipped reference at 1K and 3K. Both imaging and aesthetic quality decline substantially at both checkpoints. These results support retaining the original DMD update alongside the targeted auxiliary supervision.

Table 10: **Critic-free control on TSD-B.** Imaging and aesthetic quality are evaluated with VBench-1 on the official prompts and reported on a 0 to 100 scale. Results use one training run per configuration on the second cluster.

	Imaging quality ↑		Aesthetic quality ↑	
Checkpoint	With critic	Critic-free	With critic	Critic-free
1K	69.5	28.1	66.0	37.1
3K	71.3	35.0	66.7	42.3

E CONTINUED DISTILLATION

We encourage readers to watch demo.mp4 and the full clips in the videos/ folder of the supplementary material when assessing continued distillation. These videos show changes in visual fidelity, action coherence, and object interactions over time that VBench-1 and VBench-2 do not fully capture. The benchmark scores serve as complementary references and should be interpreted alongside these video comparisons.

Figure 4 compares 1K and 3K checkpoints across four scenes. The TSD comparisons illustrate improved visual detail together with coherent subject motion and scene evolution. Benchmark scores provide complementary evidence, but do not fully characterize these perceptual changes. In particular, dynamic degree measures the presence of motion above a flow threshold, and flow magnitude measures displacement rather than action coherence or physical plausibility. The decreases reported in Tables 1 and 9 indicate reduced movement, which can coexist with improved visual detail. Our objective is to retain the appearance benefits of continued distillation while mitigating nearly static outputs, temporal discontinuities, and implausible interactions.

The 1K DMD in Table 1 provides an early-checkpoint reference and achieves a competitive VBench-1 score. However, a high aggregate score does not establish that further training offers no visual benefit. Figure 4 illustrates finer appearance detail at 3K, motivating supervision that preserves coherent temporal behavior as appearance improves. We refer readers to the supplementary videos to assess these changes over complete sequences.

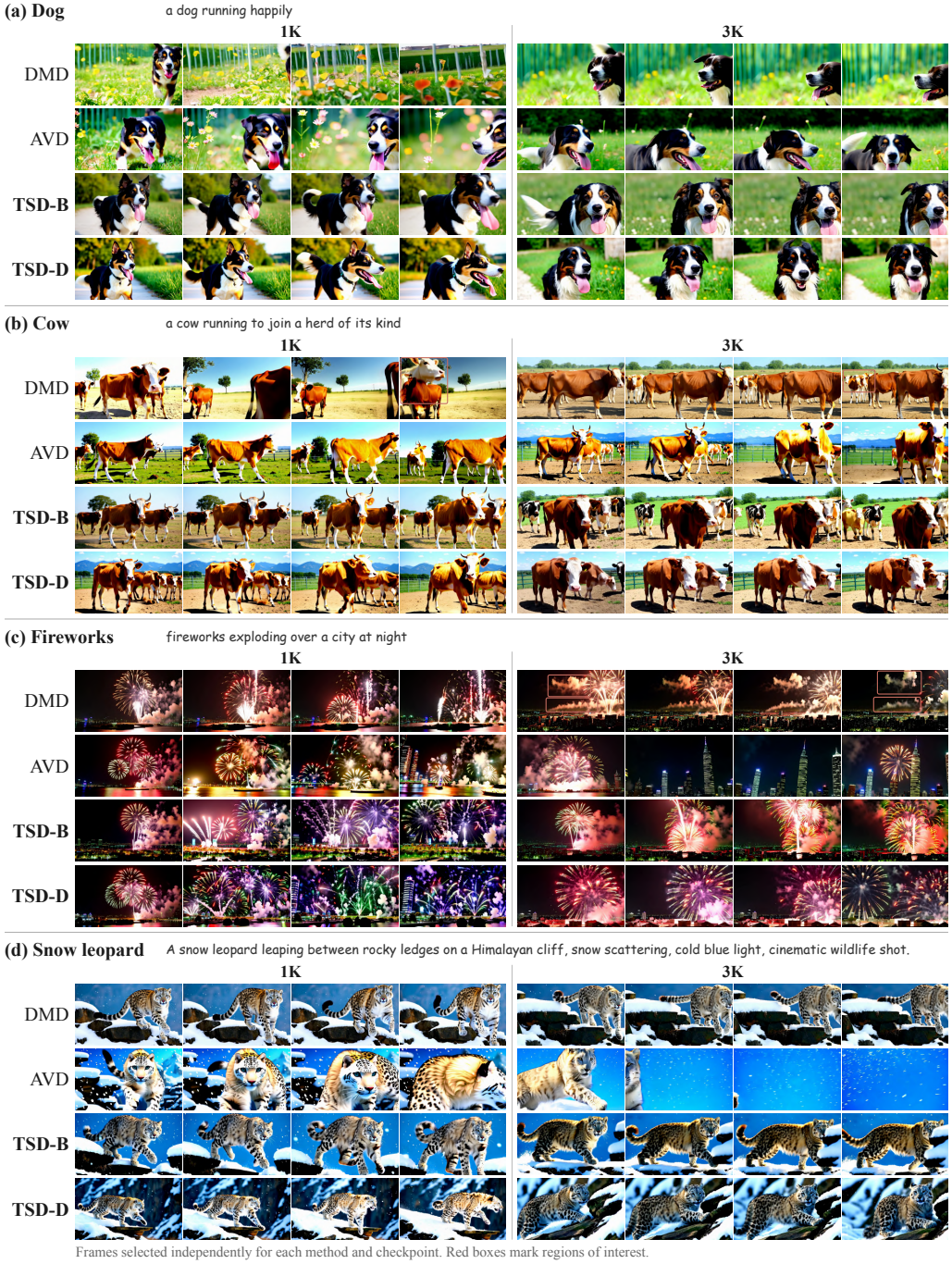

Figure 4: **Continued distillation.** DMD, AVD, TSD-B, and TSD-D at 1K (left) and 3K (right). Red boxes highlight head-body overlap and persistent smoke bands.