REVISITING SCORE DISTILLATION FOR VIDEO GENERATION: A TEMPORAL-FREQUENCY PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Score distillation is a leading paradigm for compressing a many-step video diffusion or flow teacher into a few-step student. Existing practice is organized around two divergences that expose a semantics–dynamics trade-off: reverse KL divergence anchors semantics but attenuates dynamics, whereas Fisher divergence retains dynamics but shifts semantics. In this work, we show the trade-off is intrinsic to the divergences, for they respond differently to two errors in mode selection and variance retention: Fisher charges the former weakly, and reverse KL the latter. The two errors, however, separate along the temporal-frequency axis: the former is settled in the lower band that carries global semantics, while the latter is recorded in the upper band that carries local dynamics. These findings dictate the design of Temporal-Frequency Routed Distillation (TFRD), which routes the lower band through reverse KL and the upper band through Fisher, and uses a gradient projector to keep the two bands' updates spectrally disjoint. Applied to the 4-step distillation of Wan2.1-T2V-1.3B, TFRD consistently improves the total, quality, semantic, and dynamic scores on VBench over prior methods.

1 INTRODUCTION

Diffusion and flow models have achieved unprecedented fidelity and diversity in image and video synthesis, underpinned by stable and scalable training paradigms (Dhariwal & Nichol, 2021; Karras et al., 2022; Ho et al., 2022; Rombach et al., 2022; Esser et al., 2024; Brooks et al., 2024; Wan et al., 2025; Gao et al., 2025). Their principal drawback is inference cost: sampling typically requires tens to hundreds of neural network evaluations for iterative denoising. Few-step distillation reduces this cost by compressing a many-step teacher into a one- to four-step student. Representative methods include progressive distillation (Salimans & Ho, 2022; Meng et al., 2023), consistency distillation (Song et al., 2023; Song & Dhariwal, 2024), score distillation (Wang et al., 2023b; Luo et al., 2023b; Yin et al., 2024b;a; Salimans et al., 2024; Zhou et al., 2024), and adversarial distillation (Sauer et al., 2024b;a; Lin et al., 2024; 2025).

Among these, score distillation is a leading paradigm for few-step video generation, and current methods differ in the divergences they descend. Distribution Matching Distillation (DMD) (Yin et al., 2024b;a) minimizes the reverse Kullback–Leibler (KL) divergence from student to teacher. Its gradient takes the form of the score gap between the auxiliary fake model, which tracks the student's evolving distribution, and the frozen teacher, so it never differentiates through the fake model. Score Gradient Matching Distillation (SGMD) (Wu et al., 2026) instead descends the Fisher divergence between the two score fields, pulling the gradient back through the fake model's Jacobian. As shown in Figure 1, the two methods fall on opposite sides of a *semantics–dynamics trade-off*.

This trade-off is intrinsic to the divergences, not an artifact of how DMD and SGMD estimate them. Reverse KL is *mode-seeking*: it penalizes student mass where the teacher assigns low density, pulling the student away from the tails, where large and diverse motions reside. The student then keeps only a fraction of the teacher's temporal variation (*variance attenuation*) and dampens the dynamics the teacher would have produced. The Fisher divergence reads the two distributions only through their score fields: a mode's mixing weight enters the log density as an additive constant and vanishes under differentiation, leaving those fields nearly identical wherever one component dominates. The student can therefore reweight the teacher's modes at little penalty (*mode-mass drift*) and return a well-made video not aligned with the prompt. Neither divergence alone is adequate for both errors.

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.

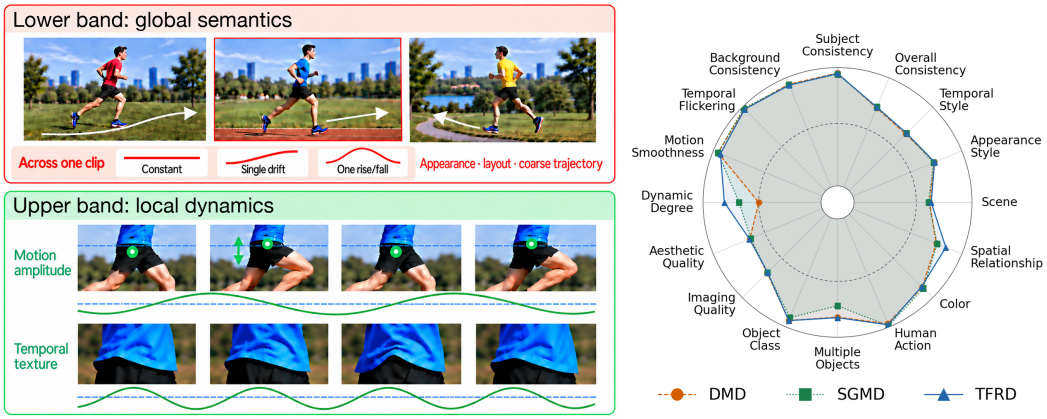

Figure 1: **Upper left:** The lower band captures the global semantics, which vary slowly over a clip, illustrated by alternative subjects, scenes, and low-order temporal profiles: a constant, a single drift, and one rise and fall. **Lower left:** The upper band captures the local dynamics, which oscillate repeatedly within a clip, illustrated by the motion amplitude of the runner's body and the temporal texture of fabric flutter. **Right:** On the VBench radar, DMD and SGMD exhibit a semantics–dynamics trade-off, which TFRD alleviates to achieve the best semantic and dynamic scores.

The two errors, however, separate along the *temporal-frequency* axis. As illustrated in Figure 1, we apply the discrete cosine transform (DCT) along the temporal dimension and split the spectrum at a cutoff into a *lower band* and an *upper band*. What distinguishes the teacher's modes–global semantics such as subject, layout, and coarse trajectory–varies slowly over a clip, so the lower band settles *mode selection*. How much each of them moves–local dynamics such as motion amplitude and temporal texture–oscillates within a clip, so the upper band records *variance retention*.

Each divergence, in turn, charges one of these errors and responds weakly to the other. In the lower band, we model teacher and student as mixtures that share components and differ only in their weights. Under local component dominance, the two scores all but coincide, while reverse KL still separates the mixtures through a log-ratio of the weights, so it charges the mode-mass drift that Fisher does not, and is *mode-mass aware*. In the upper band, we model them instead as centered Gaussians diagonal in the temporal DCT basis, the student retaining some fraction of the teacher's variance in each bin. As a bin's retention ratio falls, the reverse-KL cost grows only logarithmically in its inverse while the Fisher cost grows linearly, so Fisher exerts the stronger restoring force and is *attenuation-sensitive*. One cutoff therefore separates the two errors and the two sensitivities at once.

The analysis dictates the band routing, and we propose Temporal-Frequency Routed Distillation (TFRD) to carry it out: reverse KL below the cutoff and the Fisher divergence above it, each implemented through the estimator that approaches it. The lower-band update already lies in its band, whereas the Jacobian pullback above the cutoff can couple DCT bins and carry the upper-band update out of its own. We confine it with an orthogonal projector after the pullback, keeping the forward pass at full context and the two sample-space updates spectrally disjoint at injection.

We instantiate TFRD on the open-source video diffusion model Wan2.1-T2V-1.3B (Wan et al., 2025), distilling a 50-step teacher with classifier-free guidance (CFG) into a 4-step student without guidance. This reduces the number of network evaluations (NFE) from 100 to 4. On VBench, TFRD achieves a total score of 84.74, compared with 83.19 for DMD and 83.72 for SGMD. It also achieves the highest quality, semantic, and dynamic scores among the three methods.

Our contributions can be summarized as follows:

- We take a temporal-frequency perspective on score distillation and show that Fisher responds weakly to mode-mass drift in lower band, and reverse KL to variance attenuation in upper band.
- We propose TFRD, a band-routed distillation method that assigns each band its appropriate divergence and keeps the two bands' updates spectrally disjoint via a gradient projector.
- We validate TFRD on the 4-step distillation of Wan2.1-T2V-1.3B, where it consistently surpasses prior score distillation methods on VBench in the total, quality, semantic, and dynamic scores.

2 Related Work

Score distillation has developed along two divergences. The first is the reverse KL divergence from the student distribution to the data distribution. VSD (Wang et al., 2023b) formulates the gradient of this objective as the difference between the score of the student's distribution and that of the data, so a student can be trained with two auxiliary networks: a fake model that tracks its evolving distribution and a frozen teacher that stands in for the data. Diff-Instruct (Luo et al., 2023b) integrates the same quantity along the diffusion process to obtain a data-free objective for single-step generators. DMD (Yin et al., 2024b) realizes the gradient as a per-sample normalized, stop-gradient displacement that requires no Jacobian of either network. DMD2 (Yin et al., 2024a) introduces an adversarial term and a multi-step student, and designs a two-timescale update rule with five fake-model updates per student iteration. CausVid (Yin et al., 2025) and Self-Forcing (Huang et al., 2025) adapt DMD to distill a bidirectional teacher into an autoregressive student for real-time interactive video generation. The liability of this thread is intrinsic to the divergence: reverse KL is mode-seeking, which in video manifests as attenuated dynamics. The second thread replaces reverse KL with the Fisher divergence, which constrains the score field pointwise rather than the mode masses. SiD (Zhou et al., 2024) is among the first to back-propagate through the score networks rather than treat their outputs as constants. SIM (Luo et al., 2024) trades the intractable gradient of the score for estimable score values. SGMD (Wu et al., 2026) brings this thread to video: it descends a teacher stop-gradient Fisher divergence and stands in for the omitted score-field term with a correction built from its tracking residual, which reduces the fake-model update schedule from five per student iteration to one. This thread improves dynamics but leaves semantics unprotected, so each divergence is strong where the other is weak. Our method reconciles them by decomposing videos along the temporal-frequency axis and assigning each divergence to its appropriate band. We defer the discussion of consistency models and adversarial distillation to Section A.

3 Preliminaries

Score distillation couples three denoisers: the *student* G_θ generates a sample $x = G_\theta(z; c)$ from a latent z and a prompt c; the *fake model* v_ψ tracks the student's distribution q; and the frozen *teacher* v_ϕ represents the data distribution p. We suppress conditioning on c below and regard each sample as a vector in $\mathbb{R}^D$, where D is the number of scalar elements it contains.

3.1 Reverse KL and DMD Estimator

The reverse KL divergence measures the expected log-density ratio from student to teacher:

$$\mathrm{KL}(q\|p) = \mathbb{E}_{x\sim q}\left[\log q(x) - \log p(x)\right]. \tag{1}$$

Let $s_{\mathrm{fake}}(x) = \nabla_x \log q(x)$ and $s_{\mathrm{real}}(x) = \nabla_x \log p(x)$ denote the clean student and teacher scores. The pathwise gradient of the reverse KL is:

$$\nabla_\theta \mathrm{KL}(q\|p) = \mathbb{E}_z\left[\big(s_{\mathrm{fake}}(x) - s_{\mathrm{real}}(x)\big)\frac{\partial x}{\partial \theta}\right]. \tag{2}$$

The clean scores in Equation (2) cannot be evaluated directly because v_ψ and v_ϕ predict velocities at noisy inputs. To obtain a tractable surrogate direction, DMD (Yin et al., 2024b;a) samples $t \in (0, 1)$ and evaluates the denoisers on a renoised sample x_t, yielding clean-sample predictions $\widehat{x}_\psi$ and $\widehat{x}_\phi$:

$$x_t = (1-t)x + t\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \quad \widehat{x}_j = x_t - t v_j(x_t, t), \quad j \in \{\psi, \phi\}. \tag{3}$$

Writing $\Delta = \widehat{x}_\psi - \widehat{x}_\phi$ for denoiser discrepancy and sg for stop-gradient, the practical gradient is:

$$\nabla_\theta \mathrm{KL}(q\|p) \simeq \mathbb{E}_{z,t,\varepsilon}\left[\frac{\Delta}{N}\frac{\partial x}{\partial \theta}\right], \quad N = \mathrm{sg}\left[\frac{\|x - \widehat{x}_\phi\|_1}{D}\right], \tag{4}$$

where the detached scalar N normalizes the displacement across noise levels. The corresponding gradient-defining loss and its sample-space gradient are:

$$\mathcal{L}_{\mathrm{DMD}} = \frac{1}{2D}\left\|x - \mathrm{sg}\left[x - \frac{\Delta}{N}\right]\right\|^2, \quad D\nabla_x \mathcal{L}_{\mathrm{DMD}} = \frac{\Delta}{N}. \tag{5}$$

This displacement is *Jacobian-free*: it reaches the student without differentiating through v_ψ.

3.2 FISHER AND SGMD ESTIMATOR

The Fisher divergence measures the squared difference between the clean student and teacher scores:

$$\mathcal{F}(q\|p) = \mathbb{E}_{x\sim q} \left\| s_{\text{fake}}(x) - s_{\text{real}}(x) \right\|^2 . \tag{6}$$

SGMD (Wu et al., 2026) adopts the clean-sample predictions in Equation (3) to construct a teacher stop-gradient surrogate for the Fisher divergence, with per-sample matching loss:

$$\mathcal{L}_{\text{match}} = \frac{\|\widehat{x}_\psi - \text{sg}[\widehat{x}_\phi]\|^2}{2DN}, \tag{7}$$

which SGMD treats as an estimate of $\mathcal{F}(q\|p)/2D$.

During each generator update, ψ and N are held fixed. With K denoting the Jacobian of $\widehat{x}_\psi$ with respect to x, the chain rule along $x \to x_t \to \widehat{x}_\psi$ gives:

$$K = \frac{\partial \widehat{x}_\psi}{\partial x} = (1-t)\frac{\partial \widehat{x}_\psi}{\partial x_t}, \quad D\nabla_x \mathcal{L}_{\text{match}} = \frac{K^\top \Delta}{N}. \tag{8}$$

This path carries the same discrepancy Δ as Equation (5), but *Jacobian-weighted*: SGMD's surrogate is a function of $\widehat{x}_\psi$, so its gradient must be pulled back through the fake model.

A separate issue concerns what holding ψ fixed omits. As the student distribution changes, the fake model must be refitted, so the total derivative of $\mathcal{L}_{\text{match}}$ would contain an additional implicit term induced by this refitting. The fixed-ψ update does not evaluate this term.

To compensate for the omission, SGMD adds a negative-residual (NR) correction to the generator update. Writing $r = x - \widehat{x}_\psi$ for the tracking residual, the loss and its sample-space gradient are:

$$\mathcal{L}_{\text{NR}} = -\frac{1}{2Dt} \left\| \text{sg}[x] - \widehat{x}_\psi \right\|^2, \quad D\nabla_x \mathcal{L}_{\text{NR}} = \frac{K^\top r}{t}. \tag{9}$$

SGMD combines the two terms with correction weight β as $\mathcal{L}_{\text{SGMD}} = \mathcal{L}_{\text{match}} + \beta \mathcal{L}_{\text{NR}}$.

4 METHODOLOGY

Two failures survive distillation for different reasons. The student may have learned many of the teacher's modes but still weight them wrongly under a given prompt, returning a well-made video that is not the one asked for. It can also attenuate the dynamics the teacher would have produced. The two divergences of Section 3 charge these errors in opposite directions: reverse KL keeps the mode weights in its objective but treats attenuation cheaply, whereas the Fisher divergence reacts sharply to attenuation yet is nearly blind to drifting mode weights. Neither is adequate alone, but the two errors separate along the temporal-frequency axis. What follows analyzes the divergences themselves; renoising, the detached scales, and the Jacobian are devices the estimators use to approach them. Section 4.1 sets up the decomposition axis, Sections 4.2 and 4.3 establish the two halves of the asymmetry, and Section 4.4 routes each band to the divergence that charges its error.

4.1 TEMPORAL-FREQUENCY DECOMPOSITION

As illustrated in Figure 2, we split the sample space along the temporal-frequency axis. Let $U \in \mathbb{R}^{T\times T}$ be the orthonormal DCT-II matrix applied along the temporal dimension, where T is the number of latent temporal positions. For any length-T signal z at a fixed spatial position and channel,

$$Z = Uz, \quad \|z\|^2 = \sum_{k=0}^{T-1} |Z_k|^2. \tag{10}$$

With an integer cutoff $k_c \in \{1, \dots, T-1\}$, we split the spectrum into a lower band $\mathcal{B}_\ell$ and an upper band $\mathcal{B}_u$ with orthogonal projectors P_ℓ and P_u:

$$\begin{aligned} &\mathcal{B}_\ell = \{0, \dots, k_c - 1\}, \quad \mathcal{B}_u = \{k_c, \dots, T-1\}, \\ &P_b = U^\top \text{diag}(\mathbf{1}_{k\in\mathcal{B}_b}) U, \quad b \in \{\ell, u\}. \end{aligned} \tag{11}$$

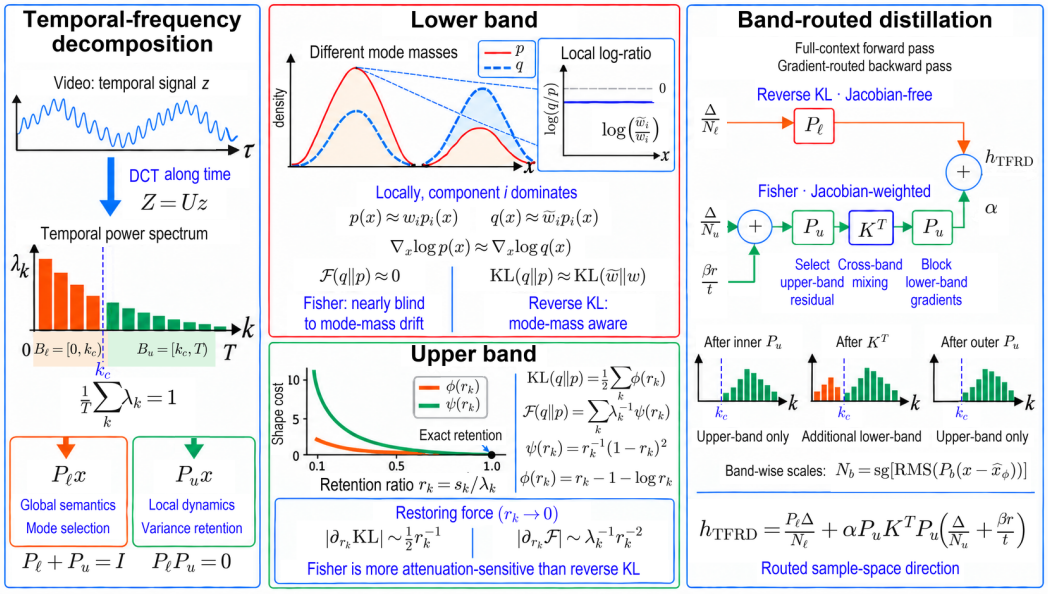

Figure 2: Overview of TFRD. **Left:** Temporal-frequency decomposition splits the video into a lower band, carrying the global semantics, and an upper band, carrying the local dynamics. **Upper middle:** Under mode-mass drift, the Fisher divergence is nearly blind, while reverse KL is mode-mass aware. **Lower middle:** Under variance attenuation, the Fisher divergence is attenuation-sensitive, while reverse KL is only weakly so. **Right:** Each band is routed to the divergence that charges its error: reverse KL on the lower, Fisher on the upper. A projector confines the upper gradient to its band.

We also use P_b for the operator on $\mathbb{R}^D$ that applies this $T \times T$ projector to the temporal fiber at every channel and spatial position, leaving those indices unchanged. The resulting operator is an orthogonal projector, with $P_\ell + P_u = I$ and $P_\ell P_u = 0$. For a teacher sample x, let X_k collect its k-th DCT coefficient over all spatial positions and channels, and set:

$$\lambda_k = \frac{\mathbb{E}\|X_k\|^2}{T^{-1}\mathbb{E}\|x\|^2}, \quad \frac{1}{T}\sum_{k=0}^{T-1} \lambda_k = 1. \tag{12}$$

Video latents are dominated by their low temporal frequencies: λ_k decays with k, so the majority of the teacher's variance sits in $\mathcal{B}_\ell$, and $\mathcal{B}_u$ holds the residual minority that carries the dynamics. What distinguishes the teacher's modes–the subject, the layout, the coarse trajectory–varies slowly in time and is therefore settled in the lower band, while the upper band records how much each of them moves–the motion amplitude, the temporal texture. Section 4.2 takes up the lower band's control of mode selection, and Section 4.3 the decay of λ_k.

4.2 Mode Selection in the Lower Band

The first failure is one of *mode selection*–which of the teacher's modes a given sample realizes–and by Section 4.1 it is settled in the lower band. Its distributional signature is *mode-mass drift*: the student's mixing weights over these modes depart from the teacher's. Here we temporarily restore the suppressed conditioning on c to the weights alone, where that drift lives,

$$p(x \mid c) = \sum_i w_i(c)\, p_i(x), \quad q(x \mid c) = \sum_i \widetilde{w}_i(c)\, p_i(x), \tag{13}$$

with positive, normalized weights $w_i(c), \widetilde{w}_i(c)$ and smooth components p_i, one per alternative at c. Teacher and student render each alternative the same way, so their disagreement is confined to the weights–and it is the weights that answer the prompt. The student's error is therefore $\widetilde{w}(c) \neq w(c)$: it holds the teacher's alternatives but weights them as though answering a different prompt. We fix c and suppress it again below. Suppose component i dominates both mixtures throughout most of the mass associated with mode i, so locally $p(x) \approx w_i p_i(x)$ and $q(x) \approx \widetilde{w}_i p_i(x)$. The two mixture

scores then nearly coincide, and the divergences part company:

$$\nabla_x \log p(x) \approx \nabla_x \log p_i(x) \approx \nabla_x \log q(x),$$
$$\mathcal{F}(q\|p) \approx 0, \quad \mathrm{KL}(q\|p) \approx \sum_i \widetilde{w}_i \log \frac{\widetilde{w}_i}{w_i}. \tag{14}$$

For the Fisher divergence, the weights are removed by the gradient: inside mode i the mixing weight multiplies p_i, so it enters the logarithm as an additive constant and vanishes under differentiation, leaving both scores at $\nabla_x \log p_i$ irrespective of the weights. Scores can therefore tell the two mixtures apart only where modes overlap, and the q-weighted average in $\mathcal{F}$ gives those regions little weight, since they hold little of the student's mass. A student can move substantial mass onto alternatives the teacher favored less, returning the wrong video for the prompt, and still leave the Fisher divergence almost unchanged. Reverse KL admits no such cancellation and is instead *mode-mass aware*: the weights survive the logarithm as $\log(\widetilde{w}_i/w_i)$, and averaging over q returns them weighted by the student's mode masses–the reverse KL $\mathrm{KL}(\widetilde{w}\|w)$ between the two weight vectors.

Section C makes this quantitative once the components overlap by a controlled amount, and states both sides on the lower band: the Fisher field obeys an upper bound that does not depend on the mode-mass drift, while the lower-band marginals of q and p obey a reverse-KL lower bound that grows with it. Modes that overlap too heavily to be told apart fall outside the controlled-overlap assumption, but moving mass between them carries no alignment cost, since a prompt does not distinguish them either. The pairs whose reweighting does misalign a prompt differ in what the prompt names, and therefore overlap little and satisfy the assumption.

4.3 Variance Retention in the Upper Band

The second failure is one of *variance retention*–how much of the teacher's temporal variation survives inside a sample–and by Section 4.1 it is recorded in the upper band. We model it as a scale mismatch. Let the distributions be centered Gaussians diagonal in the temporal DCT basis of Equation (10): $p = \mathcal{N}(0, \Sigma)$ and $q = \mathcal{N}(0, S)$, with diagonal entries λ_k (the spectrum of Equation (12)) and s_k. The retention ratio $r_k = s_k/\lambda_k$ is the fraction of the teacher's variance retained in bin k. *Variance attenuation* in this model corresponds to $r_k \ll 1$ on $\mathcal{B}_u$, while $r_k = 1$ is exact retention. The score field is linear (Equation (29)), and $\{r_k\}$ parameterizes the scale errors. Evaluating the two divergences gives:

$$\mathrm{KL}(q\|p) = \frac{1}{2}\sum_k \phi(r_k), \qquad \mathcal{F}(q\|p) = \sum_k \frac{1}{\lambda_k}\psi(r_k),$$
$$\phi(r) = r - 1 - \log r, \qquad \psi(r) = (1-r)^2/r. \tag{15}$$

The difference follows from how each divergence scales with the student's variance. As bin k is attenuated, $r_k \to 0$, $\phi(r_k) = \log(1/r_k) + O(1)$ whereas $\psi(r_k) = 1/r_k + O(1)$. Differentiating the explicit costs gives $|\partial_{r_k}\mathrm{KL}| \sim \frac{1}{2}r_k^{-1}$ and $|\partial_{r_k}\mathcal{F}| \sim \lambda_k^{-1} r_k^{-2}$, so the objectives exert different restoring forces against severe attenuation under a finite optimization budget. The two costs also treat the direction of the error differently: the Fisher cost is symmetric, $\psi(r) = \psi(1/r)$, charging a loss and a gain of a bin's variance by the same factor alike, whereas $\phi(1/r) > \phi(r)$ for $r < 1$ makes the loss the cheaper of the two under reverse KL, which therefore favors attenuation over amplification–the form its mode-seeking tendency takes in this model. Section D proves Equation (15) and gives conditions under which the same cost rates extend beyond the Gaussian model.

The factor λ_k^{-1} in the Fisher restoring force is what ties the comparison to a band, and reverse KL cannot carry it: its per-bin cost depends on the retention ratio alone, and its invariance under any invertible reparameterization of the sample makes this structural rather than an artifact of the model. As noted in Section 4.1, the teacher's variance decays with k, so λ_k^{-1} is largest exactly where that variance is smallest, and at equal retention the Fisher divergence is most *attenuation-sensitive* in the upper band and least so in the lower one. Aggregated over the full spectrum, that weighting reaches the conclusion of Section 4.2 by a second, independent route: there the mixture score cancels the mode weights, here the large λ_k of the lower band divides down whatever residual survives. Reverse KL has the complementary profile, charging the mode weights while discounting a loss of variance, which is what the routing in Section 4.4 exploits.

4.4 BAND-ROUTED DISTILLATION

The two analyses assign a divergence to each band: reverse KL below the cutoff, Fisher above it. We implement each through the estimator Section 3 attaches to it, the DMD-style displacement of Equation (5) for the lower band and the SGMD-style matching and tracking terms of Equations (7) and (9) for the upper band. We replace their full-band scale N with separate detached band scales:

$$N_b = \mathrm{sg}\left[\frac{\|P_b(x - \widehat{x}_\phi)\|}{\sqrt{D}}\right], \quad b \in \{\ell, u\}. \tag{16}$$

We use the root-mean-square form here in place of the mean-absolute form of Equation (4) so that the two band scales compose orthogonally, $N_\ell^2 + N_u^2 = \mathrm{sg}[\|x - \widehat{x}_\phi\|^2/D]$; neither band is then rescaled by the other's residual. With Δ and r as in Section 3, the loss terms are:

$$\mathcal{L}_\ell = \frac{1}{2D}\left\| x - \mathrm{sg}\left[x - \frac{P_\ell \Delta}{N_\ell}\right]\right\|^2, \tag{17}$$

$$\mathcal{L}_u = \frac{1}{2DN_u}\left\|P_u\big(\widehat{x}_\psi - \mathrm{sg}[\widehat{x}_\phi]\big)\right\|^2 - \frac{\beta}{2D}\left\|\widehat{x}_\psi - \mathrm{sg}\left[\widehat{x}_\psi + \frac{P_u r}{t}\right]\right\|^2, \tag{18}$$

and the generator objective is $\mathcal{L}_{\mathrm{TFRD}} = \mathcal{L}_\ell + \alpha\mathcal{L}_u$, where α weights the upper-band loss against the lower-band one. The lower-band displacement and tracking correction are gradient-defining surrogates in the sense of Equations (5) and (9): $D\nabla_x\mathcal{L}_\ell = P_\ell\Delta/N_\ell$, and the subtracted tracking term in $\mathcal{L}_u$ reproduces the NR direction of Equation (9) in band-restricted form, $+\beta K^\top P_u r/t$, before the outer projector and weight α are applied. The Jacobian-free lower-band gradient $P_\ell\Delta/N_\ell$ already lies in its assigned band. The upper-band gradient has no such guarantee: the fake model's Jacobian can couple DCT bins, so $K^\top P_u\Delta$ may leave the band it was selected from. We apply a second P_u after the differentiation through $\widehat{x}_\psi$: the inner projector selects the upper-band component of the discrepancy, the outer one confines the pulled-back gradient to that band, and the forward pass is left at full context throughout. This outer projector acts on the gradient rather than on the loss, so the direction injected into the student is:

$$h_{\mathrm{TFRD}} = \frac{P_\ell\Delta}{N_\ell} + \alpha P_u K^\top P_u\left(\frac{\Delta}{N_u} + \frac{\beta r}{t}\right). \tag{19}$$

Since $P_\ell + P_u = I$ and $P_\ell P_u = 0$, each temporal bin belongs to exactly one generator path, and the two sample-space gradient fields are spectrally disjoint at injection. Section B states the fake-model loss and summarizes the complete training procedure in Algorithm 1.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Models and Evaluation. We distill the open-source Wan2.1-T2V-1.3B teacher (Wan et al., 2025) into a 4-step student with the same architecture. Both the student and fake model are initialized from the pretrained teacher. The teacher uses 50 denoising steps with CFG, requiring $50 \times 2 = 100$ NFEs; each student uses 4 NFEs without CFG. Videos contain 81 frames at 16 fps and have a resolution of 832×480. We compare TFRD with DMD (Yin et al., 2024b;a), implemented as multi-step DMD2 without its adversarial term, and SGMD (Wu et al., 2026). We report the total, quality, semantic, and dynamic scores using the official VBench evaluation (Huang et al., 2024).

Implementation Details. All methods share a codebase and train on the public, prompt-only corpus adopted by DMD (Yin et al., 2024b;a), differing only in their distillation objectives and fake-model update schedules. DMD performs five fake-model updates per student iteration, whereas SGMD and TFRD perform one. The fake model is trained on student samples with the standard flow-matching velocity loss. We fine-tune the student and fake model with AdamW at learning rates of 2×10^{-6} and 4×10^{-7}, respectively. We set $k_c = 3$ and $\alpha = 1$ for TFRD, and $\beta = 0.1$ for both SGMD and TFRD. Training uses a batch size of 64 for 1,000 student iterations, with FSDP2, gradient checkpointing, and BF16 mixed precision. For sampling, the student uses four timesteps $(1000, 937.5, 833.3, 625)$, followed by the clean endpoint at 0. With $T = 21$ temporal positions, the resulting lower/upper split of $3/18$ has a direct physical reading: the k-th DCT basis in Equation (10) has angular frequency

Table 1: Main comparison on Wan2.1-T2V-1.3B. Ratio denotes the number of fake-model updates per student iteration. VBench scores are reported on a 0–100 scale, with higher values preferred.

Method	NFE ↓	Ratio ↓	Total ↑	Quality ↑	Semantic ↑	Dynamic ↑
DMD (Yin et al., 2024b;a)	4	5	83.19	84.03	79.84	52.78
SGMD (Wu et al., 2026)	4	1	83.72	84.97	78.70	69.44
TFRD (ours)	4	1	**84.74**	**85.64**	**81.12**	**81.94**



Figure 3: Qualitative comparison of 4-step generated videos from DMD, SGMD, and TFRD.

$\omega_k = \pi k/T$, so bin k completes $k/2$ periods over the clip. The lower band therefore holds exactly the components that complete at most one full period: a constant level ($k = 0$), a single drift ($k = 1$), and one complete rise and fall ($k = 2$). We take these to carry the global semantics, and every upper-band bin ($k \geq 3$), oscillating through more than one cycle, to encode the local dynamics.

5.2 Main Results

Quantitative Comparison. Table 1 splits the two failures across the full-band baselines. DMD applies reverse KL to every bin and leads SGMD on semantics; SGMD applies Fisher to every bin and leads DMD on dynamics. Each scores lowest on the metric that registers the error its divergence charges weakly: variance attenuation for reverse KL and mode-mass drift for Fisher. TFRD instead exceeds both baselines on both axes, with the highest total and quality scores. An interpolation between the two objectives would land inside the segment they span, so clearing both endpoints indicates that the two errors are separable along the temporal-frequency axis and that routing recovers what full-band application gives up.

Qualitative Comparison. Figure 3 shows what the two failures look like in the samples. In the left example, SGMD returns a well-formed clip that realizes different content from what the prompt names, while DMD and TFRD realize the prompt-aligned content: the failure is one of mode selection, not of fidelity, and it appears in what the lower band carries–the subject, layout, and coarse trajectory. In the right example, DMD keeps the prompt-aligned content but flattens the motion of camera and subject, whereas SGMD and TFRD retain the motion amplitude the prompt requests: the failure is one of variance retention, not of mode selection, and it appears in what the upper band records–how much of that content moves. Richer examples are available on our project page.[1]

[1] https://tfrdvideo.pages.dev/

Table 2: User study. Each entry is the win rate of TFRD over the comparator in the column header. Ties contribute half a win, so 50% indicates equal preference and larger values favor TFRD.

Win rate of TFRD ↑	**vs. DMD (Yin et al., 2024b;a)**	**vs. SGMD (Wu et al., 2026)**
Preference	83.2%	75.7%

Table 3: Ablations of the band cutoff k_c, loss weight α, and gradient routing. Every other column differs from the "Default" column only in the setting given in its header.

Metric	**Default**	$k_c = 2$	$k_c = 4$	$\alpha = 0.5$	$\alpha = 2$	**w/o** P_u
Total ↑	**84.74**	84.04	84.02	84.33	84.39	84.30
Quality ↑	**85.64**	85.05	84.69	85.12	85.28	85.31
Semantic ↑	81.12	80.00	**81.36**	81.19	80.85	80.24
Dynamic ↑	81.94	**88.89**	72.22	76.39	81.94	80.56

User Study. We complement automatic metrics with a blinded pairwise comparison of TFRD against DMD and SGMD, involving 20 participants and 30 prompts. In each trial, a participant views two clips in randomized order without seeing the method identities, then chooses a preferred clip or a tie. Table 2 reports the win rate $(n_{\text{win}} + 0.5\, n_{\text{tie}})/n_{\text{all}}$ over n_{all}=600 judgments per pair. TFRD is preferred in both pairings, at 83.2% against DMD and 75.7% against SGMD.

5.3 Ablation Study

Table 3 ablates the band cutoff, loss weight, and gradient routing under the same training and evaluation settings, each varied from the default configuration ($k_c = 3$, $\alpha = 1$, P_u enabled).

Band Cutoff. We sweep the band cutoff $k_c \in \{2, 3, 4\}$, with each step moving a single bin between the two paths. Handing bin $k = 2$ to the Fisher divergence drops the semantic score to 80.00 and lifts the dynamic score to 88.89; handing bin $k = 3$ to reverse KL moves both the other way, to 81.36 and 72.22. Whichever divergence receives the bin gains on the error it charges and loses on the one it charges weakly, as Sections 4.2 and 4.3 predict. Both exchanges lower the total score; at $k_c = 2$, the table's highest dynamic score comes with its lowest semantic one.

Loss Weight. We vary the scale α of the upper-band terms relative to the lower-band term over $\{0.5, 1, 2\}$. Halving α costs 5.55 on the dynamic score and returns 0.07 on the semantic one, while doubling it returns nothing on the former and gives up 0.27 on the latter. This sweep moves the total score by at most 0.41, against 0.72 in the cutoff sweep, because the band scales of Equation (16) already size each path from its own residual and leave α to set only their relative weight. What TFRD gains is therefore attributable to *where* each divergence is applied rather than how strongly.

Gradient Routing. We compare TFRD with a variant that removes the outer gradient projector P_u of Equation (19). The semantic score falls by 0.88 to 80.24 and takes the total down to 84.30. The lower band is where the leak lands: the fake model's Jacobian can couple DCT bins, and the component it returns below the cutoff can oppose the lower-band displacement, the only term charging mode-mass drift. This suggests that the gradient must be confined to its band after the pullback through $\widehat{x}_\psi$, not merely selected in the upper band before it.

6 Conclusion

We analyze two errors intrinsic to the divergences: reverse KL leads to variance attenuation, while Fisher admits mode-mass drift. The two errors, however, separate along the temporal-frequency axis. Each divergence charges the other's error and responds weakly to its own, making reverse KL mode-mass aware and Fisher attenuation-sensitive. TFRD therefore routes each band to the divergence that charges its error and uses a gradient projector to keep the two bands' updates spectrally disjoint. On Wan2.1-T2V-1.3B, it distills a 100-NFE teacher into a 4-NFE student and consistently outperforms prior methods on VBench in the total, quality, semantic, and dynamic scores. We hope to leave the community with an insight: the divergence for score distillation need not be applied uniformly–where to apply it can matter as much as which one to choose.

AI USE STATEMENT

We used generative AI tools to polish the manuscript, identify and summarize relevant literature, and write and edit software code. We reviewed all AI-assisted work: we tested AI-assisted code, verified AI-suggested references against the original papers, and checked all derivations, claims, and results. The remaining required-disclosure tasks are not applicable to this work. We take responsibility for the final content, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. OpenAI Technical Report, 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems*, 2025.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

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.

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

Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

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

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023a.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In *Advances in Neural Information Processing Systems*, 2023b.

Weijian Luo, Zemin Huang, Zhengyang Geng, J. Zico Kolter, and Guo-jun Qi. One-step diffusion distillation through score implicit matching. In *Advances in Neural Information Processing Systems*, 2024.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik P. Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. In *Advances in Neural Information Processing Systems*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your flow: Scaling continuous-time flow map distillation. In *Advances in Neural Information Processing Systems*, 2025.

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

Tim Salimans, Thomas Mensink, Jonathan Heek, and Emiel Hoogeboom. Multistep distillation of diffusion models via moment matching. In *Advances in Neural Information Processing Systems*, 2024.

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

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, 2024b.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, 2024.

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

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

Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion. In *International Conference on Learning Representations*, 2023a.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Advances in Neural Information Processing Systems*, 2023b.

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.

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *International Conference on Learning Representations*, 2022.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In *International Conference on Learning Representations*, 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.

A Extended Related Work

Few-step distillation originates with Salimans & Ho (2022), which repeatedly halves the sampling steps by training a student to match two teacher steps at once. Meng et al. (2023) extends it to classifier-free guided models by absorbing the guidance scale into a conditional student. As each halving is a separate run whose error enters the next, later work turns to objectives that target the few-step regime directly, yielding score distillation, consistency models, and adversarial distillation.

Consistency models distill the *trajectory* of the probability-flow ODE. CM (Song et al., 2023) imposes a self-consistency constraint requiring all points on one ODE trajectory to map to the same endpoint, so a single evaluation suffices. LCM (Luo et al., 2023a) moves to the latent space of diffusion models with guidance folded into the objective, and CTM (Kim et al., 2024) generalizes the map to arbitrary anytime-to-anytime transitions so that quality continues to improve with additional steps. These schemes compare network outputs across a finite step, and shrinking that step to zero turns the self-consistency constraint into a differential condition along the probability-flow ODE. This continuous-time limit removes the discretization bias but must be evaluated through Jacobian-vector products, which sCM (Lu & Song, 2025) stabilizes and scales to large models. MeanFlow (Geng et al., 2025) and AYF (Sabour et al., 2025) can be understood as combining sCM with CTM under the rectified-flow schedule. Because these objectives behave like forward divergences, they are mode-covering: they retain diversity but tend to lose detail. To recover it, rCM (Zheng et al., 2026) adds a mode-seeking reverse-KL score-distillation regularizer on top of sCM.

Adversarial distillation supplements or replaces the teacher's score with a learned discriminator. DDGAN (Xiao et al., 2022) and Diffusion-GAN (Wang et al., 2023a) first couple adversarial training with the diffusion noising process. ADD (Sauer et al., 2024b) combines a score-distillation term with a pretrained-feature discriminator for one-step synthesis. Its latent variant LADD (Sauer et al., 2024a) reuses the teacher's own features as the discriminator backbone to reach high resolution, and Hyper-SD (Ren et al., 2024) and SDXL-Lightning (Lin et al., 2024) interleave segment-wise consistency or progressive distillation with adversarial refinement. Adversarial terms supply a sharp, data-driven signal, but they reintroduce real data, an extra network, and the usual instability of GAN training. Our method uses neither a consistency constraint nor a discriminator. It operates purely on the score and is in principle compatible with either family as an additional term.

B Training Algorithm

The fake model tracks the student's full output distribution without band projection. After each student update, we draw a fresh minibatch of prompts c', latents z', noise levels t', and Gaussian noise ε'. With $x' = \mathrm{sg}[G_\theta(z'; c')]$ and $x'_{t'} = (1 - t')x' + t'\varepsilon'$, the rectified-flow velocity target is $\varepsilon' - x'$. We update only ψ by minimizing:

$$\mathcal{L}_{\text{fake}}(\psi) = \mathbb{E}\left[\frac{1}{2D}\left\|v_\psi(x'_{t'}, t'; c') - (\varepsilon' - x')\right\|^2\right]. \tag{20}$$

Algorithm 1 Temporal-Frequency Routed Distillation

Require: Student model G_θ, fake model v_ψ, frozen teacher v_ϕ, cutoff k_c, weights α, β

1: Construct the band projectors using Equation (11).
2: **for** each training iteration **do**
3: Sample prompts c and latent z; generate $x \leftarrow G_\theta(z; c)$.
4: Sample t, ε; evaluate fake and teacher denoisers by Equation (3).
5: Form $\Delta \leftarrow \widehat{x}_\psi - \widehat{x}_\phi$, $r \leftarrow x - \widehat{x}_\psi$, and the detached band scales by Equation (16).
6: Form h_{TFRD} by Equation (19), using a vector–Jacobian product for $K^\top$.
7: Update θ with $\frac{1}{D}(\partial x/\partial\theta)^\top h_{\text{TFRD}}$.
8: Draw a fresh minibatch $(c', z', t', \varepsilon')$ for Equation (20).
9: Update ψ by minimizing Equation (20).
10: **end for**

C Sensitivity to Mode-Mass Drift

This appendix makes Equation (14) quantitative for overlapping smooth components. Throughout, $p = \sum_{i=1}^{m} w_i p_i$ and $q = \sum_{i=1}^{m} \widetilde{w}_i p_i$ are the mixtures of Equation (13) at a fixed prompt, whose conditioning we continue to suppress, and we write

$$s_i = \nabla_x \log p_i, \quad \pi_i = \frac{w_i p_i}{p}, \quad \widetilde{\pi}_i = \frac{\widetilde{w}_i p_i}{q}, \quad \delta = \|\widetilde{w} - w\|_1 \tag{21}$$

for the component scores, the posterior component probabilities under p and under q, and the mode-mass drift, respectively.

Assumption 1 (Local dominance with controlled overlap)**.** There exist pairwise disjoint measurable sets $A_1, \ldots, A_m$, with $A = \bigcup_i A_i$, and constants $\eta \in [0, 1)$, $G < \infty$ and $\kappa \geq 0$ such that

(A1) $\int_{A_i} p_i(x)\,\mathrm{d}x \geq 1 - \eta$ for every i,

(A2) $\pi_i(x) \geq 1 - \eta$ and $\widetilde{\pi}_i(x) \geq 1 - \eta$ for every i and every $x \in A_i$,

(A3) $\|s_i(x) - s_j(x)\| \leq G$ for every i, j and every $x \in A$,

(A4) $\mathbb{E}_q\big[\max_{i,j} \|s_i(x) - s_j(x)\|^2\, \mathbf{1}\{x \notin A\}\big] \leq \kappa$.

Conditions (A1) and (A3) constrain the components of p alone, whereas (A2) and (A4) also depend on the student weights $\widetilde{w}$. All four concern the full sample. The band-restricted bound below requires one further assumption.

Assumption 2 (Lower-band determination of dominance regions)**.** The dominance regions of Assumption 1 are determined by the lower band: there are pairwise disjoint measurable sets $B_1, \ldots, B_m$ in the range of P_ℓ such that $A_i = \{x : P_\ell x \in B_i\}$ for every i.

The upper band is modeled separately in Section 4.3.

Proposition 1. *Under Assumption 1,*

$$\mathcal{F}(q\|p) \leq 4\eta^2 G^2 + \kappa, \tag{22}$$

$$\mathrm{KL}(q\|p) \geq 2\left(\tfrac{\delta}{2} - 2\eta\right)_+^2, \tag{23}$$

where Equation (23) *uses (A1) alone. If Assumption 2 also holds, the same bound applies to the lower-band marginals $q_\ell = (P_\ell)_{\#}q$ and $p_\ell = (P_\ell)_{\#}p$:*

$$\mathrm{KL}(q_\ell\|p_\ell) \geq 2\left(\tfrac{\delta}{2} - 2\eta\right)_+^2. \tag{24}$$

Moreover, for every i and every $x \in A_i$,

$$\left|\log \frac{q(x)}{p(x)} - \log \frac{\widetilde{w}_i}{w_i}\right| \leq -\log(1 - \eta). \tag{25}$$

Proof. Both mixture scores are convex combinations of the component scores, $\nabla_x \log p = \sum_i \pi_i s_i$ and $\nabla_x \log q = \sum_i \widetilde{\pi}_i s_i$. Since $\sum_i \pi_i = \sum_i \widetilde{\pi}_i = 1$,

$$\nabla_x \log q - \nabla_x \log p = \sum_{i,j} \widetilde{\pi}_i \pi_j\, (s_i - s_j). \tag{26}$$

Fix i and $x \in A_i$. By (A2) the remaining posterior mass satisfies $\sum_{j \neq i} \pi_j(x) \leq \eta$, so (A3) gives

$$\|\nabla_x \log p(x) - s_i(x)\| = \Big\| \textstyle\sum_{j \neq i} \pi_j(x)\big(s_j(x) - s_i(x)\big) \Big\| \leq \eta G,$$

and the same bound holds with q and $\widetilde{\pi}$ in place of p and π. By the triangle inequality the two mixture scores differ by at most $2\eta G$ at every $x \in A$. For $x \notin A$, the coefficients $\widetilde{\pi}_i \pi_j$ in Equation (26) are nonnegative and sum to one, so the difference is at most $\max_{i,j} \|s_i(x) - s_j(x)\|$. Splitting the expectation that defines $\mathcal{F}$ over A and its complement and applying (A4),

$$\mathcal{F}(q\|p) \leq (2\eta G)^2\, q(A) + \kappa \leq 4\eta^2 G^2 + \kappa,$$

which is Equation (22).

For Equation (23), let g map $x \in A_i$ to i and every $x \notin A$ to a fixed index, and let $P = g_{\#}p$ and $Q = g_{\#}q$ be the pushforwards. Drawing $I \sim w$ and then $x \sim p_I$ realizes p, and $(g(x), I)$ is a coupling of P and w. Because the A_i are disjoint, (A1) gives $\Pr[g(x) \neq I] \leq \sum_i w_i\big(1 - \int_{A_i} p_i\big) \leq \eta$, hence $\mathrm{TV}(P, w) \leq \eta$. The same construction under q gives $\mathrm{TV}(Q, \widetilde{w}) \leq \eta$. Total variation contracts under the pushforward by g, and $\mathrm{TV}(\widetilde{w}, w) = \delta/2$, so by the triangle inequality

$$\mathrm{TV}(q, p) \geq \mathrm{TV}(Q, P) \geq \mathrm{TV}(\widetilde{w}, w) - \mathrm{TV}(Q, \widetilde{w}) - \mathrm{TV}(P, w) \geq \tfrac{\delta}{2} - 2\eta.$$

Total variation is nonnegative, so the right-hand side may be replaced by its positive part. Pinsker's inequality $\mathrm{KL}(q\|p) \geq 2\,\mathrm{TV}(q, p)^2$ with logarithms in nats then yields Equation (23).

For Equation (24), note that under Assumption 2 the map g factors through the lower band. Let $\widetilde{g}$ send B_i to i and the remainder of the range of P_ℓ to the fixed index, so that $g = \widetilde{g} \circ P_\ell$ and the pushforwards of the previous paragraph read $Q = \widetilde{g}_{\#}q_\ell$ and $P = \widetilde{g}_{\#}p_\ell$. Total variation contracts under $\widetilde{g}$, so the chain above may be routed through the marginals,

$$\mathrm{TV}(q_\ell, p_\ell) \geq \mathrm{TV}(Q, P) \geq \tfrac{\delta}{2} - 2\eta,$$

and Pinsker applied to q_ℓ and p_ℓ gives Equation (24).

Finally, for $x \in A_i$ the definitions in Equation (21) give $p(x) = w_i p_i(x)/\pi_i(x)$ and $q(x) = \widetilde{w}_i p_i(x)/\widetilde{\pi}_i(x)$, whence

$$\log \frac{q(x)}{p(x)} - \log \frac{\widetilde{w}_i}{w_i} = \log \frac{\pi_i(x)}{\widetilde{\pi}_i(x)}.$$

By (A2) both posteriors lie in $[1 - \eta, 1]$, so the right-hand side lies in $[\log(1 - \eta), -\log(1 - \eta)]$, which is Equation (25). □

Remark 1 (Uniformity in the mode-mass drift). The right-hand side of Equation (22) is independent of δ: every pair (p, q) satisfying Assumption 1 with the same constants obeys the same Fisher bound, regardless of its mode-mass drift. Because (A2) constrains both mixtures, the bound covers reweighting among the alternatives that p and q both support, and for well-separated components it accommodates δ of order one, beyond the threshold 4η at which Equation (23) becomes informative. The Fisher divergence therefore obeys an overlap-controlled upper bound over the same range on which the reverse-KL lower bound grows with δ, which is the precise sense in which it is weakly sensitive to mode-mass drift. Finally, $\|P_\ell v\| \leq \|v\|$ for any orthogonal projector, so Equation (22) also bounds the lower-band quantity $\mathbb{E}_q\|P_\ell(\nabla_x \log q - \nabla_x \log p)\|^2$ discussed in Section 4.2.

Remark 2 (Band restriction of the two divergences). The divergences admit different forms of band restriction. Write $f = \log(q/p)$ for the log-density ratio, so that $\nabla f = \nabla_x \log q - \nabla_x \log p$ is the score gap. The Fisher divergence decomposes orthogonally: $\|\nabla f\|^2 = \|P_\ell \nabla f\|^2 + \|P_u \nabla f\|^2$ pointwise, so Equation (22) also bounds either projected field (Remark 1). Reverse KL does not split this way because f is a scalar. Its available decomposition is hierarchical rather than orthogonal, given by the chain rule over the lower-band marginals,

$$\mathrm{KL}(q\|p) = \mathrm{KL}(q_\ell\|p_\ell) + \mathbb{E}_{q_\ell}\big[\mathrm{KL}\big(q(\cdot \mid y)\|p(\cdot \mid y)\big)\big], \quad y = P_\ell x. \tag{27}$$

Under Assumption 2 the hierarchical form is enough: by Equation (24) the leading term alone already charges the drift. The comparison in Section 4.2 is then band-matched: a δ-independent upper bound on the lower-band Fisher field against a δ-growing lower bound on the lower-band reverse KL. The method follows the same distinction: the upper path projects the Fisher discrepancy, while the lower path projects the sample-space direction derived from global reverse KL.

D Sensitivity to Variance Attenuation

This appendix proves Equation (15) and the properties of ϕ and ψ used in Section 4.3. As in that section, let $p = \mathcal{N}(0, \Sigma)$ and $q = \mathcal{N}(0, S)$, where Σ and S are positive definite and diagonal in the temporal DCT basis of Equation (10). Their diagonal entries are $\lambda_k > 0$ and $s_k > 0$, and $r_k = s_k/\lambda_k$ is the retention ratio. Write $f = \log(q/p)$ and

$$\phi(r) = r - 1 - \log r, \quad \psi(r) = \frac{(1 - r)^2}{r} = \frac{1}{r} - 2 + r. \tag{28}$$

Proposition 2. *With the notation above, the potential and the field are*

$$f(x) = -\frac{1}{2}\sum_k \Big(\frac{1}{s_k} - \frac{1}{\lambda_k}\Big)x_k^2 - \frac{1}{2}\sum_k \log r_k, \quad \partial_k f(x) = -\frac{1}{\lambda_k}\cdot\frac{1-r_k}{r_k}\,x_k, \tag{29}$$

and the two divergences evaluate to

$$\mathrm{KL}(q\|p) = \mathbb{E}_q[f] = \frac{1}{2}\sum_k \phi(r_k), \tag{30}$$

$$\mathcal{F}(q\|p) = \mathbb{E}_q\|\nabla f\|^2 = \sum_k \frac{1}{\lambda_k}\,\psi(r_k). \tag{31}$$

As $r \to 0^+$ the summands and their slopes separate as

$$\begin{aligned} \phi(r) &= \log\tfrac{1}{r} + O(1), \quad \psi(r) = \tfrac{1}{r} + O(1), \\ \phi'(r) &= 1 - \tfrac{1}{r}, \quad \psi'(r) = 1 - \tfrac{1}{r^2}, \end{aligned} \tag{32}$$

so that the costs diverge as $\log\frac{1}{r}$ against $\frac{1}{r}$ and their slopes as $\frac{1}{r}$ against $\frac{1}{r^2}$. Moreover $\psi(1/r) = \psi(r)$, whereas $\phi(1/r) > \phi(r)$ for $r \in (0,1)$.

Proof. Both distributions are centered Gaussians diagonal in the same basis, so $\log q(x) = -\frac{1}{2}\sum_k x_k^2/s_k - \frac{1}{2}\sum_k \log s_k - \frac{D}{2}\log 2\pi$ and likewise for p. Subtracting gives the first expression in Equation (29), and the second follows from $s_k^{-1} - \lambda_k^{-1} = \lambda_k^{-1}(1-r_k)/r_k$.

Taking $\mathbb{E}_q$ of the first expression with $\mathbb{E}_q[x_k^2] = s_k$,

$$\mathbb{E}_q[f] = -\frac{1}{2}\sum_k s_k\Big(\frac{1}{s_k} - \frac{1}{\lambda_k}\Big) - \frac{1}{2}\sum_k \log r_k = \frac{1}{2}\sum_k \big(r_k - 1 - \log r_k\big),$$

which is Equation (30). Squaring the second and taking $\mathbb{E}_q$ with $\mathbb{E}_q[x_k^2] = \lambda_k r_k$,

$$\mathbb{E}_q\|\nabla f\|^2 = \sum_k \frac{1}{\lambda_k^2}\cdot\frac{(1-r_k)^2}{r_k^2}\cdot\lambda_k r_k = \sum_k \frac{1}{\lambda_k}\cdot\frac{(1-r_k)^2}{r_k},$$

which is Equation (31). The Fisher expression squares the score-field slope while the expectation contributes one factor of student variance. In the KL expression, the quadratic part of the potential contains only one factor of each.

Equation (32) follows from Equation (28): as $r \to 0^+$ the logarithm dominates ϕ and the term r^{-1} dominates ψ, and both derivatives are immediate, using $\psi(r) = \frac{1}{r} - 2 + r$ for ψ'. The symmetry $\psi(1/r) = \psi(r)$ follows from the same expansion, and $\phi(1/r) - \phi(r) = \frac{1}{r} - r + 2\log r$ vanishes at $r = 1$ with derivative $-(1-r)^2/r^2 \le 0$, hence is positive on $(0,1)$. □

Remark 3 (Budgeted attenuation). Suppose the contribution of bin k to either objective is at most ϵ. Under Equation (30), this gives $\phi(r_k) \le 2\epsilon$ and hence $\log(1/r_k) \le 1 + 2\epsilon$. Under Equation (31), it gives $\psi(r_k) \le \lambda_k\epsilon$. Setting $a = \lambda_k\epsilon$ and solving the resulting quadratic yields

$$r_k \;\ge\; \tfrac{1}{2}\Big[(2+a) - \sqrt{(2+a)^2 - 4}\Big] = \begin{cases} 1 - \sqrt{a} + O(a), & a \to 0, \\ a^{-1} + O(a^{-2}), & a \to \infty. \end{cases}$$

The KL bound permits attenuation that deepens exponentially with ϵ, whereas the Fisher bound permits only polynomial deepening. The latter is especially restrictive when the teacher variance makes a small.

Remark 4 (Beyond the Gaussian model). The cost rates in Equation (32) extend beyond Gaussian distributions when the teacher satisfies suitable integrability conditions against the degenerating student. For $u \in (0,1]$, let q_u be the distribution obtained by multiplying the k-th coordinate of $x \sim q$ by $\sqrt{u}$. Assume q has finite Fisher information, nonzero in that coordinate, and let p be smooth. Writing ∇^2 for the Laplacian, require

$$\sup_{0<u\le 1}\mathbb{E}_{q_u}\big\|\nabla\log p\big\|^2 < \infty, \quad \sup_{0<u\le 1}\mathbb{E}_{q_u}\big|\nabla^2\log p\big| < \infty, \quad \sup_{0<u\le 1}\mathbb{E}_{q_u}\big|\log p\big| < \infty. \tag{33}$$

Then $\partial_k \log q_u$ gains a factor $u^{-1/2}$ under the coordinate contraction, giving the exact identity $\mathbb{E}_{q_u}\big[(\partial_k \log q_u)^2\big] = u^{-1}\mathbb{E}_q\big[(\partial_k \log q)^2\big]$. The differential entropy changes by only $\frac{1}{2}\log u$. Expanding $\mathcal{F}(q_u\|p)$, integration by parts turns the cross term $\mathbb{E}_{q_u}[\nabla \log q_u \cdot \nabla \log p]$ into $-\mathbb{E}_{q_u}[\nabla^2 \log p]$. The first two conditions in Equation (33) therefore leave a u^{-1} leading term. The third gives logarithmic growth for $\mathrm{KL}(q_u\|p)$.

Remark 5 (Scope and relation to Assumption 1). The two models address complementary errors. Assumption 1 allows mixing weights to drift while holding the components fixed; with one component, it forces $q = p$. The Gaussian model instead has one component whose scale may drift, with no mode weights to change. Together, the models distinguish errors in mode selection from errors in variance retention.

Because Σ, S, and P_u are diagonal in the same basis, restricting the sums in Equations (30) and (31) to $k \in \mathcal{B}_u$ gives band-restricted costs term by term. For general centered Gaussians, Equation (30) still holds with r_k equal to the eigenvalues of $\Sigma^{-1}S$, because reverse KL is invariant under an invertible linear change of variables. The Fisher expression does not have this invariance: in general,

$$\mathcal{F}(q\|p) = \mathrm{tr}\big[(S^{-1} - \Sigma^{-1})^2 S\big], \tag{34}$$

which depends on Σ and S separately rather than only on the eigenvalues of $\Sigma^{-1}S$. It reduces to Equation (31) when the two covariance matrices commute. This dependence on coordinates is what produces the factor λ_k^{-1} in Section 4.3.

E Limitations and Future Work

TFRD decomposes the video along a single temporal axis and splits it at one cutoff into two bands, while spatial frequencies and finer-grained band structures are not considered. The cutoff is held fixed throughout distillation rather than adapted to the prompt. Empirically, we study one teacher at a single resolution and clip length, and transfer to larger backbones and longer clips remains to be evaluated. These are properties of the present instantiation rather than constraints of the routing itself, which requires only an orthogonal decomposition of the video and a divergence assigned to each component. This formulation therefore suggests extensions to spatial frequencies, finer-grained partitions, and prompt-dependent cutoffs, at the cost of recomputing the corresponding projectors.