CoDMD: Copula-aware Distribution Matching Distillation for Fast Video Generation

Wenhu Zhang[1] **Kun Cheng**[2*] **Yiting Wang**[3] **Shiyao Li**[5] **Yuechen Zhang**[4]
Wenbo Li[4] **Jiajun Zha**[1] **Jingyi Zhang**[2] **Kang Zhao**[2] **Jiaya Jia**[16†]
[1]HKUST [2] Alibaba Token Hub, Alibaba Group [3]THU [4]CUHK [5] XDU [6] SmartMore
Project page: `https://andrew-zhang98.github.io/CoDMD_page`

Abstract

Few-step distillation for video diffusion models has attracted significant attention, driven by the urgent demand for efficient deployment in real-world scenarios. However, Distribution Matching Distillation (DMD), a leading paradigm, tends to degrade under limited NFE budgets, manifesting in video generation as layout instability, oversaturation, and broken motion dynamics. Although DMD implicitly targets the joint score, its practical sample-wise objective provides no explicit constraint on cross-sample or cross-frame relational geometry. Under a few-step budget, this can make dependence transfer inefficient and exacerbate reverse-KL mode seeking. This motivates explicit supervision of dependency-sensitive relational statistics. Motivated by this insight, we propose Copula-aware DMD (CoDMD), a lightweight copula-motivated relational regularizer that reuses predictions already produced by the frozen teacher and the online fake model to construct pairwise relation matrices across samples and frames. These are matched through a supplementary distributional objective that requires no additional networks, datasets, or sampling trajectories. On the Wan-2.1-T2V model series at 1.3B & 14B scales, CoDMD distills 50-step teachers into 4-step students, achieving an approximate 25$\times$ speed-up while attaining VBench scores of 84.46 & 84.87, outperforming prior trajectory-based (rCM 82.81 & 84.05) and distribution-based (DMD* 83.38 & 83.81) methods.

1 Introduction

Diffusion-based generative models have recently emerged as a leading paradigm for both image and video synthesis [2, 4, 5, 9, 22, 30, 35], driven by their favorable scaling behavior and superior perceptual fidelity. However, their practical deployment is hampered by increasing spatial resolutions, longer clip durations, and the iterative denoising process, which typically requires 50–100 forward passes per sample. To bridge this gap, a growing body of work pursues few-step distillation, which compresses the number of function evaluations (NFE) required at inference time and aims to deliver an efficient yet high-fidelity media generator.

Among existing few-step distillation paradigms, Distribution Matching Distillation (DMD) [42] aligns the student-induced distribution with that of the teacher through a distributional objective, rather than enforcing strict timestep-by-timestep trajectory consistency. When the number of inference steps is aggressively reduced (e.g., 4 steps or fewer), however, DMD exhibits a pronounced tendency toward mode collapse, manifesting in the image domain as oversaturation and reduced sample diversity. Several recent efforts augment DMD with auxiliary training paradigms to recover the lost quality. DMD2 [41] introduces a GAN loss to sharpen details. DMDR [8] combines DMD with reinforcement learning to improve overall fidelity. rCM [47] jointly trains DMD with consistency models to mitigate

*Project leader,†Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

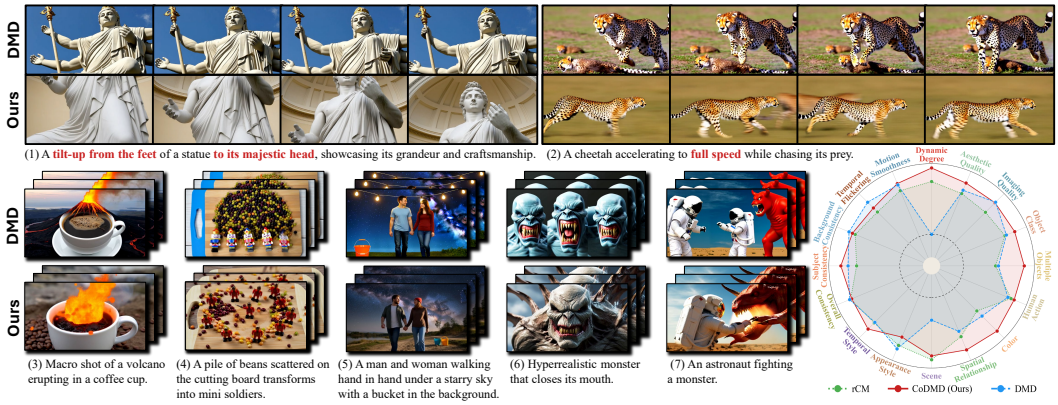


Figure 1: Sampled 4-step videos. DMD suffers from (1)*failed camera motion*; (2)*super slow action*; (3,4)*broken layouts*; (5,6,7)*oversaturation* and (4,7)*instruction misalignment*. Labels "DMD" in the figures denote DMD*, defined in Section 4.1. On the VBench radar, **CoDMD (84.46)** outperforms previous methods, i.e., **DMD (83.38)** and **rCM (82.81)**, on most of the dimensions, with the largest gains on *dynamic degree*, *color* and *multiple objects*. Best zoomed in.

mode collapse. DP-DMD [37] introduces a two-stage pipeline with v-prediction to pursue diversity. While these approaches partially relieve the collapse symptoms, they have primarily been designed for the general or image-domain generation. The substantially more challenging regime of few-step video distillation, in contrast, remains relatively underexplored from a DMD perspective.

Applying DMD to video generation amplifies image-domain defects into more conspicuous failures. As shown in Fig. 1, DMD often exhibits (i) compromised motion fidelity (failed camera motion, overly slow actions), (ii) structural inconsistencies (broken layouts, oversaturation), and (iii) semantic misalignment with the input prompt under the 4-step budget.

To articulate this deficiency, we invoke Sklar's theorem [18, 25], which decomposes a continuous joint distribution into its marginal distributions and a copula that encodes their dependence structure. From this perspective, high-quality video generation requires not only realistic local appearance but also faithful dependence among frames, objects, layouts, and motions. Motion strength, camera trajectory, object interaction, and layout persistence depend on relations that isolated per-frame statistics do not determine. Exact joint scores contain both marginal and dependence information. Our claim instead concerns practical optimization: DMD transfers the joint-score residual through student-distribution samples and the generator Jacobian, without explicitly matching cross-sample or cross-frame relational statistics. In video generation, where quality depends critically on coherent relations among appearance, motion, and layout, explicit supervision of such statistics provides a complementary way to guide structural knowledge transfer.

This motivation is reinforced by the mode-seeking tendency of reverse-KL [1]. Reverse-KL penalizes the student for assigning probability mass to regions weakly supported by the teacher, but need not symmetrically penalize missed teacher-supported modes. Under stringent NFE budgets, the student's expressive capacity is constrained, and a restricted generator family can favor overly concentrated approximations that retain dominant patterns while discarding other variations. Our restricted-mixture analysis illustrates this behavior under its stated assumptions; it does not assert that every DMD-trained generator collapses. Together, these considerations motivate copula-motivated relational regularization, while retaining the distinction between implicit joint-score supervision and explicit relational constraints. The cosine matrices serve as dependency-sensitive statistics rather than formal copula estimators.

Guided by this diagnosis, we propose Copula-aware DMD (CoDMD), a lightweight copula-aware regularizer that augments DMD without introducing any auxiliary networks, datasets, or sampling trajectories. Our key observation is that, within every DMD iteration, the score functions of both the frozen teacher and the online fake model are already evaluated for the standard distribution-matching loss. These existing predictions can therefore also provide relational supervision. Concretely, CoDMD reuses these score estimates to construct pairwise relation matrices at two complementary granularities: across batch elements and across temporal frames within each video. A supplementary

distributional objective then uses the teacher–fake relational residual to supervise the student's relation matrices explicitly. By encouraging the student to preserve the teacher's relational variability across both batch and frame, this objective discourages degenerate concentration on a narrow subset of outputs, thereby empirically counteracting the mode-seeking bias of reverse-KL.

We evaluate CoDMD on Wan-2.1-T2V diffusion model series at both 1.3B & 14B scales. CoDMD distills 50-step teachers into 4-step students while maintaining strong visual quality, achieving an approximate 25$\times$ acceleration. On VBench, it attains total scores of 84.46 & 84.87, outperforming trajectory-based methods such as rCM (82.81 & 84.05) and distribution-based algorithms such as DMD* (83.38 & 83.81), with notable gains in dynamic degree (+15), color (+2.8), and multiple objects (+3.16). Our contributions are as follows:

- We provide a copula perspective on few-step video distillation that distinguishes dependence information in joint scores from its implicit transfer through practical DMD optimization, motivating explicit relational supervision alongside reverse-KL distribution matching.
- We propose Copula-aware DMD (CoDMD), a lightweight copula-aware regularizer for DMD that supervises pairwise relations using the teacher–fake residual across batch elements and temporal frames using score estimates already computed in standard training, requiring no extra networks, data, or sampling trajectories.
- Across two scales of the Wan-2.1-T2V model series, CoDMD consistently improves few-step video distillation, achieving $\sim$25$\times$ speed-up while outperforming prior trajectory-based and distribution-based methods on the VBench total score and across most dimensions.

2 Related Work

Video diffusion and few-step acceleration. Modern video generators rest on diffusion and flow-matching formulations [5, 11, 12, 30] realized by spatio-temporal transformer backbones [2, 10, 24, 35, 40]. They reach strong fidelity, but require tens to hundreds of denoising steps per clip, which is prohibitive for interactive deployment. Training-free samplers such as DDIM [27] and DPM-Solver [15, 45] shorten the trajectory but still need many steps to remain visually satisfactory. Training-based distillation falls into two families: trajectory-matching methods that align the student along the sampling path [3, 13, 14, 23, 28, 29], and distribution-matching methods that align student and teacher at the data-distribution level. Our work builds on the latter.

Distribution Matching Distillation and its remedies. DMD [41, 42] updates the student via a score-as-target surrogate whose gradient is given by the score difference between a frozen teacher and a learnable fake model [36]. DMD preserves fine-grained details well in single-image distillation, but its naive few-step use often suffers from oversaturation and mode collapse [14, 41, 42]. These issues become more severe in video generation, where motion weakens, layouts become unstable, and sampled clips grow mutually similar. Recent remedies improve DMD by combining it with additional paradigms, such as distribution-shaping rewards [8], consistency training [29, 47], and two-stage pipelines [37]. While effective, these approaches add extra models, trajectories, or training stages on top of DMD, without directly matching batch- or frame-level relational statistics. In contrast, CoDMD strengthens the DMD objective itself rather than wrapping it with auxiliary machinery.

A parallel line of work extends DMD-style training to autoregressive long-video generation [6, 39, 43]. These methods are largely orthogonal to our few-step setting, but their emphasis on temporal error accumulation further highlights the importance of improving the underlying DMD objective, since weak supervision can compound over long horizons.

Relational supervision under computation budgets. Relational supervision has long been recognized as a powerful tool for knowledge transfer especially under strict computation budgets, as it captures the structural geometry, e.g., pairwise distances, angles, and affinity matrices between samples rather than just individual predictions [20, 33]. Relational Knowledge Distillation [19] matches pairwise distances and triplet angles between teacher and student features. Relation Networks [31] and similarity-preserving distillation [33] align sample affinity matrices. Contrastive representation distillation [32] transfers structure through mutual-information bounds. These foundational works motivate our use of similarity matrices as effective carriers of relational information. However, under the NFE-constrained setting in diffusion, the conventional single-teacher paradigm with feature-space

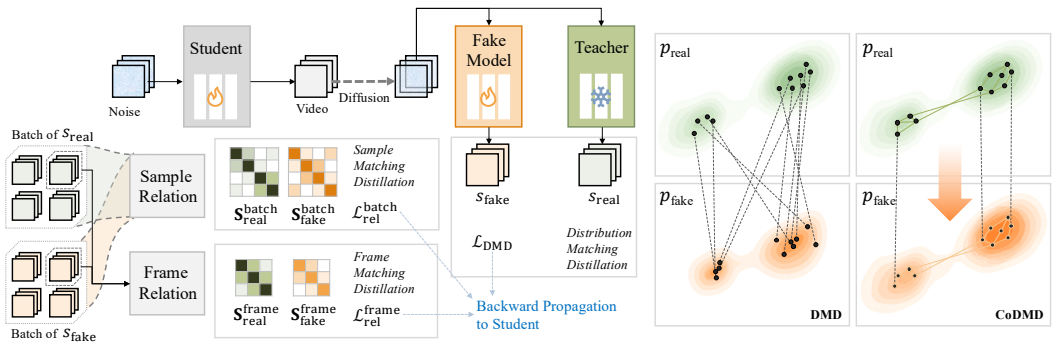


Figure 2: *Left:* Training pipeline of CoDMD. Beyond standard $\mathcal{L}_{\text{DMD}}$, we reuse the real and fake scores to build relational matrices at batch and frame granularities, yielding two supplementary relational losses $\mathcal{L}_{\text{rel}}^{\text{batch}}$ and $\mathcal{L}_{\text{rel}}^{\text{frame}}$. *Right:* The right panel illustrates an example of mismatched pairwise geometry after distribution matching. CoDMD additionally preserves the real structure among these elements.

matching becomes inadequate. Generative score distillation is intrinsically a triplet-network problem, driven by the dynamic residual gap between a frozen teacher and a fake distribution. Furthermore, to seamlessly integrate with DMD, the supervision must be adapted from the conventional feature space to the score distribution space. CoDMD addresses this by lifting the teacher–fake residual from point-wise coordinates to pairwise copula-aware relations. Our distinction is the DMD-specific teacher–fake residual target, rather than direct matching of static teacher representations.

3 Method

We first review DMD as a *score-as-target surrogate* (§3.1), then diagnose why this coordinate-wise recipe leaves the relational geometry of the teacher's score field less constrained (§3.2). Guided by this analysis, we propose Copula-aware DMD (§3.3), which adds relational supervision to the surrogate, lifting it from coordinates to pairwise similarity matrices at two complementary granularities. Figure 2 gives an overview.

3.1 Preliminaries: DMD as a coordinate-wise score-as-target surrogate

Notation. Let $\mathbf{x}_0 \sim p_{\text{real}}$ denote a clean video and $q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I})$ a variance-preserving Gaussian diffusion process indexed by $t \in [0, 1]$. A pretrained teacher denoiser $\mu_{\text{real}}(\mathbf{x}_t, t)$ provides the perturbed score $s_{\text{real}}(\mathbf{x}_t, t) = -(\mathbf{x}_t - \alpha_t \mu_{\text{real}})/\sigma_t^2$. A few-step student generator $G_\theta : \mathbb{R}^{d_z} \to \mathbb{R}^{F \times C \times H \times W}$ produces $\mathbf{x} = G_\theta(\mathbf{z})$ from $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and an auxiliary denoiser $\mu_{\text{fake}}^{\phi}(\mathbf{x}_t, t)$ is trained online with $\mathcal{L}_{\text{denoise}}^{\phi} = \|\mu_{\text{fake}}^{\phi}(\mathbf{x}_t, t) - \mathbf{x}\|_2^2$ to track the outputs of students and provide the corresponding fake score s_{fake}.

Distribution-matching gradient. DMD minimizes the reverse-KL divergence $D_{\text{KL}}(p_{\text{fake}} \,\|\, p_{\text{real}})$, whose gradient admits an expectation over coordinate-wise score differences [36]:

$$\nabla_\theta D_{\text{KL}} \simeq \mathbb{E}_{\mathbf{z},t}\Big[w_t \alpha_t \big(s_{\text{fake}}(\mathbf{x}_t, t) - s_{\text{real}}(\mathbf{x}_t, t) \big) \cdot \frac{\partial G_\theta(\mathbf{z})}{\partial \theta} \Big], \tag{1}$$

where $\mathbf{x}_t$ denotes the diffused student output and w_t represents a noise-level weighting.

Score-as-target surrogate. In practice, Eq. (1) is realized through a reconstruction loss that avoids explicitly computing the student Jacobian during the forward pass. Define the update direction

$$\mathbf{\Delta}(\mathbf{x}) \triangleq w_t \alpha_t \big(s_{\text{fake}}(\mathbf{x}_t, t) - s_{\text{real}}(\mathbf{x}_t, t) \big), \tag{2}$$

and the *score-as-target surrogate* loss

$$\mathcal{L}_{\text{DMD}} = \frac{1}{2} \big\| \mathbf{x} - \text{stopgrad}\big(\mathbf{x} - \mathbf{\Delta}(\mathbf{x})\big) \big\|_2^2. \tag{3}$$

Since the target is treated as a constant, $\nabla_\theta \mathcal{L}_{\mathrm{DMD}} = \boldsymbol{\Delta}(\mathbf{x}) \cdot \partial_\theta G_\theta(\mathbf{z})$ exactly recovers Eq. (1): the KL gradient is realized by nudging $\mathbf{x}$ toward $\mathbf{x} - \boldsymbol{\Delta}(\mathbf{x})$. Each score component can depend on the full video; this surrogate is not independent marginal matching. Its practical update samples the residual under the current student distribution and projects it through the generator Jacobian, whereas CoDMD additionally supervises explicit relations among samples and frames.

3.2 Why coordinate-wise supervision is insufficient: a copula perspective

We distinguish the dependence information in exact joint scores from its transfer during finite-capacity optimization. Sklar's theorem motivates relational statistics; restricted-family reverse-KL and pairwise examples illustrate the motivation without asserting an unconditional guarantee for our implemented loss.

Sklar's theorem and the marginal–copula factorization. Sklar's theorem [18, 25] states that any continuous joint $p(\mathbf{y})$ on $\mathbb{R}^D$ admits a unique decomposition

$$p(\mathbf{y}) = \Big(\prod_{i=1}^{D} f^{(i)}(y_i)\Big) \cdot c\big(F^{(1)}(y_1), \ldots, F^{(D)}(y_D)\big), \tag{4}$$

where $f^{(i)}, F^{(i)}$ are the marginal PDF (Probability Density Function) and CDF (Cumulative Distribution Function) of coordinate i, and $c : [0,1]^D \to \mathbb{R}_{\geq 0}$ is the copula density that encodes all cross-coordinate dependence and its underlying structure is invariant to coordinate-wise transformations of the marginals. Taking $\nabla \log p$ yields

$$\nabla_{\mathbf{y}} \log p(\mathbf{y}) = \underbrace{\left[\frac{\partial}{\partial y_i} \log f^{(i)}(y_i)\right]_{i=1}^{D}}_{=:\, s^{\mathrm{marg}}(\mathbf{y})} + \underbrace{\nabla_{\mathbf{y}} \log c\big(F^{(1)}(y_1), \ldots, F^{(D)}(y_D)\big)}_{=:\, s^{\mathrm{cop}}(\mathbf{y})}. \tag{5}$$

Exact joint scores contain dependence information: equality of smooth positive joint-score fields on a connected domain identifies the normalized densities. Practical DMD uses student-weighted, Jacobian-projected updates with an estimated fake score. CoDMD adds an explicit, non-separable relation-space signal. The cosine matrices are copula-motivated relational statistics, not formal copula estimators.

Marginal matching alone does not identify relational structure. A two-frame video (X_1, X_2) with $U \sim \mathcal{N}(0,1)$ and $P : (X_1, X_2) = (U, U)$, $Q : (X_1, X_2) = (U, -U)$ has identical per-frame marginals under P and Q, yet $\mathbb{E}_P[X_1 X_2] = 1$ versus $\mathbb{E}_Q[X_1 X_2] = -1$. Any objective that depends only on per-coordinate distributions assigns the same value to P and Q, and is blind to which temporal relation the student learns (Appendix H). This example concerns marginal-only criteria: exact DMD distinguishes the two distributions. Appendix H gives the nonzero score residual and a mixture example in which the missing-branch signal is attenuated on typical student samples.

Reverse-KL amplifies the gap via mode-seeking behavior. DMD minimizes the *reverse*-KL divergence, which is inherently mode-seeking [1, 38]. For a teacher mixture $P = \sum_{k=1}^{K} \pi_k P_k$ with disjoint supports and a student restricted to single-component approximants $\{P_1, \ldots, P_K\}$, $D_{\mathrm{KL}}(P_j \,\|\, P) = -\log \pi_j$, so the optimum is the component with the largest mixture weight (Appendix H). More generally, $D_{\mathrm{KL}}(Q_j \,\|\, P) = D_{\mathrm{KL}}(Q_j \,\|\, P_j) - \log \pi_j$ for any Q_j on component j: reverse-KL trades fit against weight but never penalizes the absence of other modes. The single-component restriction is essential to this calculation; it does not establish collapse for every DMD-trained generator.

Relational matching penalizes collapsed solutions. Let $d(z, z')$ be a symmetric distance in the score space and $\mathcal{D}(Q) = \mathbb{E}_{Z,Z' \sim Q}[d(Z, Z')]$. If the teacher score distribution P has modes with intra-mode distance $\leq \epsilon$ and cross-mode distance $\geq \Delta$, then for any single-mode student $Q = P_j$,

$$\mathcal{D}(P) - \mathcal{D}(Q) \;\geq\; \Delta\Big(1 - \sum\nolimits_{k=1}^{K} \pi_k^2\Big) - \epsilon. \tag{6}$$

The factor $(1 - \sum_k \pi_k^2)$ measures the teacher's relational diversity, approaching 1 for many well-separated modes. A single-mode student therefore incurs a discrepancy that reverse-KL ignores but a pairwise objective directly measures. Eq. (6) is the quantitative motivation for CoDMD: a pairwise statistic can expose lost cross-mode variation when the lower bound is positive. This abstract discrepancy is not equal to our implemented cosine-KL objective, which uses teacher and fake predictions on student-generated inputs.

Algorithm 1: CoDMD training procedure.

Input: Pretrained teacher μ_{real}; denoising step list $\mathcal{T} = \{t_1, \ldots, t_K\}$; relational weights $\lambda_{\text{batch}}, \lambda_{\text{frame}}$.
Output: Trained few-step generator G_θ.

$G_\theta \leftarrow \text{copyWeights}(\mu_{\text{real}})$, $\mu^{\phi}_{\text{fake}} \leftarrow \text{copyWeights}(\mu_{\text{real}})$ `// Initialization`
while *not converged* **do**
  $\mathbf{x} \leftarrow G_\theta\big(\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ \mathcal{T}\big)$; `// few-step backward simulation`
  if is generator update step **then**
    $\mathcal{L}_{\text{DMD}} \leftarrow \text{DMDLoss}(\mu_{\text{real}}, \mu^{\phi}_{\text{fake}}, \mathbf{x})$; `// Eq. (3)`
    $\mathcal{L}^{\text{batch}}_{\text{rel}} \leftarrow \mathcal{L}^{general}_{\text{rel}}(\mu_{\text{real}}, \mu^{\phi}_{\text{fake}}, \mathbf{x}, \text{batch})$; `// Eq. (9)`
    $\mathcal{L}^{\text{frame}}_{\text{rel}} \leftarrow \mathcal{L}^{general}_{\text{rel}}(\mu_{\text{real}}, \mu^{\phi}_{\text{fake}}, \mathbf{x}, \text{frame})$; `// Eq. (10)`
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{DMD}} + \lambda_{\text{batch}} \mathcal{L}^{\text{batch}}_{\text{rel}} + \lambda_{\text{frame}} \mathcal{L}^{\text{frame}}_{\text{rel}}$; `// Eq. (11)`
    $G_\theta \leftarrow \text{update}(G_\theta, \mathcal{L}_{\text{total}})$
  Sample $t \sim \mathcal{U}(0, 1)$;;
  $\mathbf{x}_t \leftarrow \text{forwardDiffusion}(\text{stopgrad}(\mathbf{x}), t)$;
  $\mathcal{L}^{\phi}_{\text{denoise}} \leftarrow \|\mu^{\phi}_{\text{fake}}(\mathbf{x}_t, t) - \text{stopgrad}(\mathbf{x})\|_2^2$; `// update fake score every step`
  $\mu^{\phi}_{\text{fake}} \leftarrow \text{update}(\mu^{\phi}_{\text{fake}}, \mathcal{L}^{\phi}_{\text{denoise}})$

3.3 Copula-aware DMD (CoDMD)

The analysis above motivates a regularizer based on pairwise statistics. We instantiate it with *cosine similarity matrices* computed directly from the score-based denoiser predictions.

Relational analogue of $\mathbf{x}$: the similarity matrix. Every DMD training step evaluates three quantities that carry distribution-level information: the student output $\mathbf{x} = G_\theta(\mathbf{z})$, the teacher prediction $\mu_{\text{real}}(\mathbf{x}_t, t)$, and the fake model prediction $\mu^{\phi}_{\text{fake}}(\mathbf{x}_t, t)$. We construct pairwise cosine similarity matrices from these score-space outputs at two granularities.

At the *batch level*, each generated sample $\mathbf{x}^{(b)} \in \mathbb{R}^{F \times C \times H \times W}$ is averaged over its temporal and spatial dimensions (F, H, W), preserving the channel dimension to form a C-dimensional vector. This yields B vectors per quantity. After gathering across all GPUs to B_{g} vectors, we compute the $B_{\text{g}} \times B_{\text{g}}$ cosine similarity matrix, where pairwise similarities are measured along the C channels. At the *frame level*, only the spatial dimensions (H, W) of each sample are averaged per frame, while the channel dimension remains intact, yielding F vectors of size C per sample. An $F \times F$ cosine similarity matrix is then computed per sample to capture inter-frame relations. As shown in Fig. 2, we denote the resulting similarity matrices as $\mathbf{S}^{\text{batch}}_{\text{stu}}$, $\mathbf{S}^{\text{batch}}_{\text{real}}$, $\mathbf{S}^{\text{batch}}_{\text{fake}}$ (batch level) and $\mathbf{S}^{\text{frame}}_{\text{stu}}$, $\mathbf{S}^{\text{frame}}_{\text{real}}$, $\mathbf{S}^{\text{frame}}_{\text{fake}}$ (frame level, per sample b), built respectively from $\mathbf{x}$, $s_{\text{fake}}(\mathbf{x}_t, t)$, and $s_{\text{real}}(\mathbf{x}_t, t)$.

Relational analogue of the score difference. In standard DMD, the perturbation $\mathbf{\Delta}(\mathbf{x})$ is the coordinate-wise score difference $s_{\text{fake}} - s_{\text{real}}$. Its relational counterpart is simply

$$\mathbf{\Delta}_{\mathbf{S}} \triangleq \mathbf{S}_{\text{fake}} - \mathbf{S}_{\text{real}}, \tag{7}$$

which measures how the fake-model's pairwise relations deviate from those of the teacher. This difference is computed separately for the batch-level and frame-level matrices.

Relational score-as-target surrogate. We form a detached target by nudging $\mathbf{S}_{\text{stu}}$ along $-\mathbf{\Delta}_{\mathbf{S}}$ and define the relational loss as the KL divergence between the softmax distributions:

$$\begin{aligned} \mathcal{L}^{general}_{\text{rel}} &= \text{KL}(\mathbf{P}_{\text{tgt}} \| \mathbf{P}_{\text{stu}}), \\ \mathbf{P}_{\text{tgt}} &= \text{softmax}\big(\tau^{-1}[\mathbf{S}_{\text{stu}} - \mathbf{\Delta}_{\mathbf{S}}]\big), \\ \mathbf{P}_{\text{stu}} &= \text{softmax}(\tau^{-1}\mathbf{S}_{\text{stu}}). \end{aligned} \tag{8}$$

Table 1: VBench results for Wan-2.1-T2V on 480p resolution. A full per-attribute breakdown is provided in Appendix D.1, where CoDMD leads on the majority of individual VBench dimensions.

Model	Resolution	NFE	TimeCost↓	Quality Score↑	Semantic Score↑	Total Score↑
Wan2.1-T2V-1.3B						
Teacher (Pretrained)	832×480×81	50×2	270s	84.43	80.73	83.69
DMD* [42]	832×480×81	4	11s	84.53	78.78	83.38
DCM [16]	832×480×81	6	18s	83.12	74.33	81.36
sCM [14]	832×480×81	6	18s	83.42	72.98	81.33
rCM [47]	832×480×81	4	11s	83.72	79.18	82.81
AVD [44]	832×480×81	4	11s	84.73	79.80	83.75
CoDMD (Ours)	832×480×81	4	11s	**85.50**	**80.27**	**84.46**
Wan2.1-T2V-14B						
Teacher (Pretrained)	832×480×81	50×2	710s	84.71	81.75	84.12
DMD* [42]	832×480×81	4	27s	84.56	80.84	83.81
sCM [14]	832×480×81	6	44s	83.48	74.01	81.58
rCM [47]	832×480×81	4	27s	84.97	80.35	84.05
AVD [44]	832×480×81	4	27s	84.28	**82.33**	83.89
CoDMD (Ours)	832×480×81	4	27s	**85.55**	82.12	**84.87**

where $\tau > 0$ is a temperature and the gradient through $\mathbf{P}_{\text{tgt}}$ is stopped. For the row-averaged loss, the exact gradient is $(P_{\text{stu},im} - P_{\text{tgt},im})/(N\tau)$ for any $\tau > 0$ (Appendix H.5). Chaining through pairwise cosine similarities couples each representation to its relation partners. A local expansion applies the softmax Jacobian to $\mathbf{\Delta_S}$; it is not generally a scalar multiple of that residual.

We instantiate the above construction at two granularities that correspond to the two under-constrained axes identified in §3.2.

Applying Eq. (8) to the batch-level similarity matrices over B_{g} videos gives

$$\mathcal{L}_{\text{rel}}^{\text{batch}} \;=\; \text{KL}\big(\mathbf{P}_{\text{tgt}}^{\text{batch}} \,\big\|\, \mathbf{P}_{\text{stu}}^{\text{batch}}\big). \tag{9}$$

This term injects batch-level relational supervision that is absent from the coordinate-wise DMD gradient. By aligning the pairwise relational geometry of the student to the teacher–critic residual on the same batch, it constrains how the student's score field varies across the samples drawn in each step—an axis that the practical DMD update supervises only implicitly.

Applying the same construction to the per-sample frame-level matrices over F frames gives

$$\mathcal{L}_{\text{rel}}^{\text{frame}} \;=\; \frac{1}{B}\sum_{b=1}^{B} \text{KL}\big(\mathbf{P}_{\text{tgt}}^{\text{frame},(b)} \,\big\|\, \mathbf{P}_{\text{stu}}^{\text{frame},(b)}\big). \tag{10}$$

No cross-GPU communication is required for this term. It directly counters the motion degradation diagnosed in §3.2 by enforcing that temporal similarity patterns follow the teacher–fake residual.

Final objective. As standard DMD, the fake model μ_{fake}^{ϕ} is trained with the denoising loss $\mathcal{L}_{\text{denoise}}^{\phi}$ to track the moving generator distribution. For the generator G_θ, the final objective is

$$\mathcal{L}_{\text{total}} \;=\; \mathcal{L}_{\text{DMD}} \;+\; \lambda_{\text{batch}}\,\mathcal{L}_{\text{rel}}^{\text{batch}} \;+\; \lambda_{\text{frame}}\,\mathcal{L}_{\text{rel}}^{\text{frame}}. \tag{11}$$

where $\mathcal{L}_{\text{DMD}}$ denotes the coordinate-wise DMD surrogate in Eq. (3), and $\lambda_{\text{batch}}, \lambda_{\text{frame}} \geq 0$ are weighting hyperparameters. Algorithm 1 summarizes the full training procedure. Since the similarity matrices are introduced only during training, inference remains identical to that of the original few-step generator, retaining its sampling schedule and per-step computation for video generation.



Figure 3: *Left:* Our CoDMD produces more faithful motion, stronger prompt alignment, and more realistic color than previous methods (i.e., DMD and rCM). Owing to the copula-aware relational objective, our output also matches the teacher's visual structure more closely, e.g., (1)screw turns inward; (2) dragon exhales fire. *Right:* VBench metrics across training iterations. CoDMD sustains high scores throughout training, while DMD undergoes a sharp collapse in dynamic degree.

4 Experiments

4.1 Experimental setup

Models & Data. We distill the official Wan-2.1-T2V teachers (1.3B and 14B) [35], both released as 50-step models with classifier-free guidance (50×2 NFE), into 4-step students. All clips are generated at 832×480, 81 frames (5 s at 16 fps). The distribution-matching loss uses text descriptions from a mixed corpus of public videos and never accesses the raw videos themselves, following [41, 42].

Training & Evaluation. We train all models using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a classifier-free guidance (CFG) scale of 3.5, and BF16 mixed-precision training on 32 GPUs. The relational temperature is set to $\tau = 0.1$, and the two relational weights are fixed at $\lambda_{\text{batch}} = \lambda_{\text{frame}} = 0.1$. Full training hyperparameters are provided in Table A.1 in the Appendix.

For evaluation, we follow the VBench protocol [7, 46] and report quality, semantic, and total scores, with $\text{Total} = (4\,\text{Quality} + \text{Semantic})/5$. All baselines are re-evaluated in a unified codebase under BF16 precision for a fair comparison. Specifically, we use the official checkpoints for rCM and DCM. Since the official weights for sCM and AVD are not publicly available, we evaluate our reproduced implementations of these methods based on their released codebases and paper descriptions. We denote the primary DMD-derived baseline by DMD*: it removes the original regression term, uses DMD2's two-time-scale schedule with five fake updates per generator update, and omits adversarial supervision. CoDMD adds only the relational terms to this recipe; Appendix B reports raw DMD and full DMD2. Labels "DMD" in the existing figures denote DMD*.

4.2 Main Results

Quantitative comparison. Table 1 reports VBench results against trajectory-based samplers (sCM [14], rCM [47]) and distribution-based methods (DMD [42], DCM [16], AVD [44]). Our CoDMD achieves the best total score at both model scales, reaching **84.46** on 1.3B and **84.87** on 14B. Compared with the DMD* baseline, the gains are about $+1.1$ points at both scales. The full per-attribute breakdown is provided in Appendix D.1, where CoDMD leads on the majority of individual VBench dimensions. Measured generation time decreases from 270 to 11 seconds on 1.3B and from 710 to 27 seconds on 14B, corresponding to $24.5\times$ and $26.3\times$ acceleration. The guided teacher uses two network evaluations per sampling step, whereas the student uses four evaluations in total. The improvement is more pronounced on the semantic axis than on the quality axis. This trend matches the diagnosis that the semantic aspects such as spatial layout, object persistence, and motion plausibility depend more strongly on relational geometry, which is explicitly constrained by our copula-aware supervision.

Qualitative comparison. We compare videos generated by DMD, rCM, and CoDMD in Fig. 1 and Fig. 3 (left). DMD shows several recurring failure modes, including static camera motion, unnaturally slow dynamics, structural breakdown, and oversaturation. rCM improves some of these issues but introduces its own artifacts, especially broken hand structures and semantic misalignments. In contrast, CoDMD produces more faithful motion, better semantic alignment, and more realistic color rendering.

Table 2: Complementary effects of batch- and frame-level relational supervision on Wan-2.1-T2V-1.3B. Coverage measures teacher-referenced coverage across samples, while Dynamic Degree measures temporal dynamics within videos. Quality, Semantic, and Total are VBench scores.

Method	**Coverage↑** (Sample)	**Dynamic Degree↑** (Frame)	**Quality↑**	**Semantic↑**	**Total↑**
DMD*	0.12	71.1	84.5	78.8	83.4
+ batch relation	**0.49**	83.1	85.1	79.4	84.0
+ frame relation	0.24	**93.3**	85.0	76.8	83.6
CoDMD (both)	0.48	86.1	**85.5**	**80.3**	**84.5**
Direct matching	0.33	88.6	84.7	79.0	83.8

Copula-aware loss analysis. Fig. 3 illustrates visual differences and tracks Dynamic Degree during training. The training curves track within-video motion, while teacher-referenced metrics characterize variation across generated samples. To quantify variation across samples, we measure inter-sample diversity and teacher-referenced Density/Coverage (Table C.1). Coverage measures the fraction of teacher-reference neighborhoods reached by at least one generated sample in a video-feature space. For each prompt, we compare ten student samples with ten teacher samples using frozen ViCLIP features; each teacher sample defines a neighborhood whose radius is its distance to the second-nearest teacher sample. Higher Coverage indicates that student outputs reach more of these teacher-supported neighborhoods. Thus, repeatedly generating a single plausible outcome can yield low Coverage even when each clip contains substantial motion. On 1.3B, Density improves from 0.18 to 0.44 and Coverage from 0.12 to 0.48 relative to DMD*. Table 2 reports Coverage alongside Dynamic Degree to separate sample-level variation from within-video motion.

User study. To complement the automated metrics, we conduct a blind pairwise user study with 20 evaluators. Each participant is presented with a pair of videos generated from the same prompt in randomised left–right order, and asked to select the one with better *visual quality* and *semantic alignment*. As shown in Fig. 4, CoDMD is preferred over DMD in 81.8% of comparisons and over rCM in 74.4%, confirming that our distilled model delivers perceptually meaningful improvements over existing methods. Against the 50-step teacher, the win rate reaches 51.3%, summarizing the observed pairwise preferences at this sampling budget. Details are available in Appendix E.

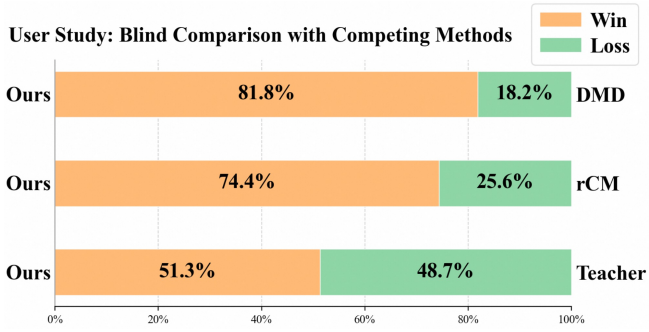


Figure 4: User study results. We report the win rate of CoDMD against others in pairwise blind comparisons.

4.3 Ablation studies

The two granularities are complementary. Table 2 isolates each component under the same setting. The DMD* baseline reaches 83.4 in total score with dynamic degree 71.1. Adding only $\mathcal{L}_{\text{rel}}^{\text{batch}}$ improves all metrics and raises the total to 84.0, while adding only $\mathcal{L}_{\text{rel}}^{\text{frame}}$ yields the largest motion gain (dynamic degree 93.3) but reduces semantic score by 2.0, indicating that temporal relational matching alone can over-emphasize motion at the expense of prompt alignment. Combining both losses gives the best total of 84.5, with dynamic degree settling at 86.1 and semantic score peaking at 80.3, consistent with the complementary roles analyzed in §3.2. Coverage distinguishes these effects

Table 3: Sensitivity to the shared relational loss weight λ on (1.3B, 4-step).

$\lambda_{\text{batch}}, \lambda_{\text{frame}}$	**Qual.↑**	**Sem.↑**	**Total↑**
0 (DMD)	84.53	78.78	83.38
0.1	**85.50**	80.27	**84.46**
0.2	85.04	**80.42**	84.12
0.5	84.79	80.40	83.91
1.0	83.87	78.00	82.70

Table 4: Alternative relation statistics on (1.3B, 4-step); DMD* uses no relation loss.

Relation statistic	Qual.↑	Sem.↑	Total↑
DMD*	84.53	78.78	83.38
CoDMD (cosine)	**85.50**	80.27	**84.46**
CoDMD (L1)	84.79	80.15	83.86
CoDMD (L2)	85.04	**80.42**	84.12
CoDMD (Mahalanobis)	85.19	80.32	84.22

across samples: the batch term raises it from 0.12 to 0.49, compared with 0.24 for the frame term. The combined model retains 0.48, supporting the complementary roles of the two granularities.

The score-as-target form is essential. Replacing the teacher–fake residual with direct matching between student and teacher score similarity matrices lowers the total score to 83.8, only slightly above DMD. As discussed in Appendix H.5, direct relational matching removes the real-versus-fake distribution-matching signal and therefore provides much weaker supervision. Unlike static teacher matching, this target incorporates the fake model's evolving estimate of the student distribution.

Sensitivity to relational loss weight. Table 3 varies the shared weight $\lambda = \lambda_{\text{batch}} = \lambda_{\text{frame}}$ from 0 (pure DMD) to 1.0. Values $\lambda \in [0.1, 0.5]$ improve over the DMD* baseline, with the best total score at λ=0.1 and the best semantic score at λ=0.2. Larger weights within this interval favor Semantic at the expense of Quality. At λ=1.0, both scores fall below the baseline, indicating a less favorable balance between relational and coordinate-wise supervision. We adopt λ=0.1 as the default throughout all other experiments.

Choice of relation statistic. Table 4 compares cosine, L1, L2, and Mahalanobis relation statistics. All four improve both Quality and Semantic over DMD*, with Total gains of 0.48–1.08 points. Cosine gives the best Quality and Total, whereas L2 gives the best Semantic score. Relative to L2, cosine gains 0.46 in Quality and gives up 0.15 in Semantic; the quality-weighted aggregation yields a 0.34-point Total advantage. Cosine compares normalized feature directions, making its relation matrices insensitive to feature magnitudes. Together, these results support cosine as an effective default while showing that the teacher–fake relational signal benefits several choices of geometry.

Causal 10-second generation. Starting from a Self-Forcing ODE checkpoint [6], we apply the same relation-space residual to obtain CoDMD-AR. Table 5 reports the causal 10-second comparison: CoDMD-AR improves Total by 0.88 over Self-Forcing and 2.41 over CausVid [43]. This extends the comparison from bidirectional five-second synthesis to sequential ten-second generation.

Table 5: Causal 10-second video generation.

Method	Quality↑	Semantic↑	Total↑
CausVid	81.90	74.00	80.32
Self-Forcing	82.38	79.73	81.85
CoDMD-AR	**83.31**	**80.40**	**82.73**

5 Conclusion

We identified that practical DMD transfers dependence information implicitly through its joint-score update, without directly supervising batch- or frame-level relational statistics. This motivates copula-motivated relational regularization alongside reverse-KL distribution matching. We addressed it with Copula-aware DMD, which lifts DMD's score-as-target surrogate from coordinates to pairwise similarity matrices at batch and frame granularities, reusing score predictions already computed at each step without extra networks, data, or inference overhead. On Wan-2.1-T2V series, CoDMD distills 50-step teachers into 4-step students that outperform prior methods on VBench across different attributes. Ablations confirm the two granularities are complementary and the teacher–fake residual formulation improves over direct matching. We discuss limitations and future work in Appendix I, including richer relational representations and adaptive loss weighting across samples and frames.

Acknowledgements and Disclosures

This work was supported in part by the Research Grants Council under the Areas of Excellence scheme grant AoE/E-601/22-R.

References

[1] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[2] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[3] P. Esser et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[4] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[5] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[6] X. Huang, Z. Li, G. He, M. Zhou, and E. Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion, 2025.

[7] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, Y. Wang, X. Chen, L. Wang, D. Lin, Y. Qiao, and Z. Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[8] D. Jiang, D. Liu, Z. Wang, Q. Wu, L. Li, H. Li, X. Jin, D. Liu, C. Lu, Z. Li, et al. Distribution matching distillation meets reinforcement learning. *arXiv preprint arXiv:2511.13649*, 2025.

[9] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

[10] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[11] Y. Lipman et al. Flow matching for generative modeling. In *ICLR*, 2023.

[12] X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[13] X. Liu, X. Zhang, J. Ma, J. Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023.

[14] C. Lu and Y. Song. Simplifying, stabilizing and scaling continuous-time consistency models. *arXiv preprint arXiv:2410.11081*, 2024.

[15] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

[16] Z. Lv, C. Si, T. Pan, Z. Chen, K.-Y. K. Wong, Y. Qiao, and Z. Liu. Dual-expert consistency model for efficient and high-quality video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14983–14993, 2025.

[17] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, 2020.

[18] R. B. Nelsen. *An introduction to copulas*. Springer, 2006.

[19] W. Park, D. Kim, Y. Lu, and M. Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019.

[20] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5007–5016, 2019.

[21] A. Polyak, A. Zohar, A. Brown, A. Tjandra, A. Sinha, A. Lee, A. Vyas, B. Shi, C.-Y. Ma, C.-Y. Chuang, D. Yan, D. Choudhary, I. Misra, J. Wang, J. Gray, K. Jagadeesh, L. Gao, M. Singh, M. R. Ronchi, M. E. Banani, P. Misra, Q. Gu, R. Bansal, R. Walsh, R. Kathuria, S. Purushwalkam, T.-H. Wu, W. Ping, Y. Wang, Z. He, et al. Movie gen: A cast of media foundation models, 2024.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[23] T. Salimans and J. Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[24] T. Seedance, D. Chen, L. Chen, X. Chen, Y. Chen, Z. Chen, Z. Chen, F. Cheng, T. Cheng, Y. Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

[25] M. Sklar. Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP*, volume 8, pages 229–231, 1959.

[26] I. Skorokhodov, S. Tulyakov, and M. Elhoseiny. StyleGAN-V: A continuous video generator with the price, image quality and perks of StyleGAN2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[27] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[28] Y. Song and P. Dhariwal. Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*, 2023.

[29] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency models. 2023.

[30] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

[31] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[32] Y. Tian, D. Krishnan, and P. Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

[33] F. Tung and G. Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1365–1374, 2019.

[34] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[35] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[36] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023.

[37] T. Wu, R. Li, L. Zhang, and K. Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. *arXiv preprint arXiv:2602.03139*, 2026.

[38] Y. Xu, W. Nie, and A. Vahdat. One-step diffusion models with f-divergence distribution matching. *arXiv preprint arXiv:2502.15681*, 2025.

[39] S. Yang, W. Huang, R. Chu, Y. Xiao, Y. Zhao, X. Wang, M. Li, E. Xie, Y. Chen, Y. Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

[40] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[41] T. Yin, M. Gharbi, T. Park, R. Zhang, E. Shechtman, F. Durand, and B. Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024.

[42] T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W. T. Freeman, and T. Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024.

[43] T. Yin, Q. Zhang, R. Zhang, W. T. Freeman, F. Durand, E. Shechtman, and X. Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22963–22974, 2025.

[44] Y. You, Y. Li, J. Li, Y. Mu, Q. Chen, and P. Jiang. Adaptive video distillation: Mitigating oversaturation and temporal collapse in few-step generation. *arXiv preprint arXiv:2603.21864*, 2026.

[45] Q. Zhang and Y. Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

[46] D. Zheng, Z. Huang, H. Liu, K. Zou, Y. He, F. Zhang, Y. Zhang, J. He, W.-S. Zheng, Y. Qiao, and Z. Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

[47] K. Zheng, Y. Wang, Q. Ma, H. Chen, J. Zhang, Y. Balaji, J. Chen, M.-Y. Liu, J. Zhu, and Q. Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. *arXiv preprint arXiv:2510.08431*, 2025.

Appendix

Copula-aware Distribution Matching Distillation

The supplementary material first describes training and baseline construction, then defines the sample-level metrics and presents additional quantitative results. User-study details and visual comparisons follow. The notation table introduces the theoretical analysis, and the final section discusses limitations and future directions.

1. Appendix A: Training configuration and compute.
2. Appendix B: DMD-family baselines.
3. Appendix C: Sample diversity and teacher-referenced distribution metrics.
4. Appendix D: Additional quantitative results.
5. Appendix E: User study.
6. Appendix F: Visual comparisons.
7. Appendix G: Notation.
8. Appendix H: Theoretical analysis.
9. Appendix I: Limitations and future work.

A Training Configuration and Compute

We use the training settings in Table A.1 for the two Wan-2.1 model scales. The DMD-family baseline recipes are specified in Appendix B.

Table A.1: Training configurations for Wan2.1 T2V.

Hyper-parameter	**1.3B**	**14B**
Batch Size	128	64
Learning Rate (student)	2e-6	2e-6
Learning Rate (fake score)	4e-7	4e-7
CFG Scale	3.5	3.5
Student Update Frequency	5	5
Maximal Simulation Steps	4	4

A.1 Compute resources

Each main training in the manuscript is conducted on 32 NVIDIA A100 GPUs with 80GB memory and takes approximately 1.5 days. We report the compute configuration for the experiments in the paper, while additional exploratory runs may require extra compute beyond the final reported settings.

B DMD-family Baselines

B.1 DMD baseline construction

The baseline denoted by DMD* uses the distribution-matching term from DMD but removes the original regression objective. It adopts DMD2's two-time-scale training schedule, with five fake-model updates per generator update, and omits adversarial supervision. CoDMD uses this same baseline recipe and adds only the batch and frame relational losses. Thus, the main comparison isolates the effect of relational supervision within a common training recipe.

Table B.1 additionally reports our reproductions of raw DMD and DMD2 on the 1.3B four-step setting. DMD2 includes adversarial supervision. CoDMD improves Total over DMD2 by 1.05 points, whereas the difference between DMD2 and DMD* is 0.03 points in this evaluation.

Table B.1: Detailed DMD-family comparison on Wan-2.1-T2V-1.3B.

Method	Quality↑	Semantic↑	Total↑
Raw DMD	83.28	74.01	81.43
DMD*	84.53	78.78	83.38
DMD2	84.59	78.67	83.41
CoDMD	**85.50**	**80.27**	**84.46**

C Sample Diversity and Teacher-Referenced Distribution Metrics

C.1 Shared sampling protocol

We sample 200 prompts from MovieGen Video Bench [21]. Each method and the teacher use the same prompts and ten shared seeds per prompt, yielding 2,000 videos per model. Feature extraction, preprocessing, and metric settings are identical across methods. For the per-prompt diversity and coverage evaluation, each prompt contributes equally to the reported average.

Inter-sample diversity. Inter-sample LPIPS measures perceptual appearance differences between outputs produced from different seeds of the same prompt. ViCLIP diversity measures variation in their high-level video representations. These measurements compare different generated samples, rather than frames within one video. We report teacher values as a reference: larger variation alone need not indicate higher generation quality.

Density and Coverage. We use a frozen ViCLIP-L/14 video encoder to extract one ℓ_2-normalized, 768-dimensional embedding per video, and compute teacher-referenced Density and Coverage following PRDC [17]. For prompt p, let $T_p = \{t_i\}_{i=1}^{N}$ and $S_p = \{s_j\}_{j=1}^{M}$ denote teacher and student embeddings, with $N = M = 10$. Let r_i be the Euclidean distance from t_i to its k-th nearest neighbor in $T_p \setminus \{t_i\}$, with $k = 2$. The metrics are

$$\text{Density}(S_p, T_p) = \frac{1}{kM} \sum_{j=1}^{M} \sum_{i=1}^{N} \mathbf{1}[\|s_j - t_i\|_2 \leq r_i], \tag{12}$$

$$\text{Coverage}(S_p, T_p) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\min_j \|t_i - s_j\|_2 \leq r_i]. \tag{13}$$

Density measures how many teacher neighborhoods contain student samples, normalized by k. Coverage measures the fraction of teacher neighborhoods reached by at least one student sample. We first compute these metrics separately for each prompt and then average over prompts. This estimates agreement with the finite teacher reference set in the chosen feature space.

C.2 Distributional metrics and teacher-referenced FVD

We compute Fréchet Video Distance (FVD) [34] with the PyTorch implementation used by StyleGAN-V [26]. A Kinetics-400-pretrained I3D network extracts video features before softmax. The 1.3B student and teacher distributions use the same 2,000 prompt–seed instances and identical preprocessing of the 81-frame videos. FVD compares feature means and covariances over the full sample set, using videos generated by the teacher as the reference.

Table C.1: Sample-level diversity and teacher-referenced distributional agreement on Wan-2.1-T2V-1.3B. LPIPS and ViCLIP diversity are interpreted relative to the teacher. Lower FVD is better.

Method	Inter-sample LPIPS	ViCLIP diversity	Density↑	Coverage↑	FVD↓
Teacher	0.5749	0.1360	–	–	–
DMD*	0.3976	0.0697	0.1800	0.1200	315.50
rCM	0.3699	0.0943	0.4000	0.2600	360.44
CoDMD	0.5864	0.1356	**0.4400**	**0.4800**	**259.22**

CoDMD's ViCLIP diversity is close to the teacher reference, and its LPIPS is slightly higher. Both baselines show lower inter-sample diversity under these measurements. CoDMD also obtains the highest Density and Coverage among the evaluated students and the lowest FVD. FVD supplies an overall distributional comparison, while per-prompt Coverage retains the prompt-conditioned structure of the sampling protocol.

D Additional Quantitative Results

This section complements the main comparisons with attribute-level results, generation-seed variability, batch composition, and reduced-step sampling.

D.1 Full VBench attribute breakdown

Table D.1 reports the per-attribute VBench scores for all compared methods at both the 1.3B and 14B scales. CoDMD achieves the highest total score at both scales and leads on the majority of individual attributes, with the most pronounced gains on *dynamic degree*, *color*, *multiple objects*, and *spatial relationship*. The teacher retains advantages on a few stability-oriented metrics such as *background consistency* and *temporal flickering*, which reflects the inherent trade-off between motion richness and frame-level steadiness under reduced NFE.

Table D.1: Full VBench attribute-level results on Wan-2.1-T2V at 1.3B and 14B scales. Bold marks the best value among the distilled models at each scale; teacher values are provided as a reference. Orange columns denote our CoDMD.

	Wan2.1-T2V-1.3B				*Wan2.1-T2V-14B*			
Attribute	Teacher	rCM	DMD*	**Ours**	Teacher	rCM	DMD*	**Ours**
Subject Consistency	93.46	94.11	95.18	**96.08**	93.58	94.87	**96.14**	95.92
Background Consistency	96.29	93.88	**94.65**	94.23	96.76	94.44	**95.44**	95.03
Temporal Flickering	99.05	97.53	**99.30**	98.26	98.90	97.66	**99.21**	98.37
Motion Smoothness	97.97	97.75	98.50	**98.67**	98.02	97.83	**98.54**	98.51
Dynamic Degree	67.22	83.06	71.11	**86.11**	68.61	**85.28**	64.44	81.39
Aesthetic Quality	68.94	64.60	64.89	**65.53**	69.28	66.78	66.36	**66.94**
Imaging Quality	67.58	66.80	**68.05**	67.99	67.82	**69.37**	68.04	68.62
Object Class	91.01	93.86	94.05	**95.11**	92.63	96.39	**96.58**	96.28
Multiple Objects	78.17	78.35	78.72	**81.88**	83.58	82.50	84.36	**86.08**
Human Action	96.80	97.20	96.80	**97.60**	98.40	99.20	**99.40**	**99.40**
Color	85.00	82.82	83.76	**86.52**	85.35	82.63	80.37	**85.19**
Spatial Relationship	81.12	76.86	76.24	**78.41**	81.96	78.19	79.30	**84.00**
Scene	56.56	**54.56**	50.86	54.20	53.82	53.66	53.75	**54.62**
Appearance Style	22.88	21.49	**21.60**	21.25	23.59	21.72	**22.49**	21.71
Temporal Style	25.90	25.08	25.05	**25.35**	25.73	25.19	25.23	**25.92**
Overall Consistency	27.09	26.60	**26.63**	26.57	27.34	26.85	26.99	**27.14**
Quality Score	84.43	83.72	84.53	**85.50**	84.71	84.97	84.56	**85.55**
Semantic Score	80.73	79.18	78.78	**80.27**	81.75	80.35	80.84	**82.12**
Total Score	83.69	82.81	83.38	**84.46**	84.12	84.05	83.81	**84.87**

D.2 Sensitivity to generation seeds

We repeat the complete 1.3B VBench evaluation under six generation-seed configurations, retaining the same checkpoints and inference settings. Table D.2 gives means and sample standard deviations across those configurations. Confidence intervals are two-sided 95% Student-t intervals for the seed-level mean. The average Total improvement is 1.10 points. These repetitions characterize inference-seed variability for fixed models, rather than variation across independently trained models.

Table D.2: VBench over six generation-seed configurations.

Method	Quality	Semantic	Total	Total 95% t-CI	Total min–max
DMD*	84.63 ± 0.17	78.05 ± 0.58	83.32 ± 0.21	[83.09, 83.54]	[83.09, 83.71]
CoDMD	85.53 ± 0.23	79.99 ± 0.49	84.42 ± 0.20	[84.21, 84.62]	[84.14, 84.64]

D.3 Batch size and prompt composition

We vary batch size and prompt composition while holding the number of optimization updates and the other hyperparameters fixed. Table D.3 reports random batches of size 64, 128, and 256, together with diverse and category-narrow prompt batches at size 128. The original setting is random sampling with batch size 128. Coverage increases with batch size in the random-batch comparison. At fixed

Table D.3: Batch support and sample-level coverage on Wan-2.1-T2V-1.3B. The same evaluation protocol is used for all rows.

Batch size	Prompt composition	LPIPS	Coverage↑	VBench Total↑
64	Random	0.5443	0.2600	84.15
128	Random (default)	0.5864	0.4800	84.46
256	Random	0.5697	**0.6400**	**84.48**
128	Diverse	0.6098	0.6000	84.42
128	Category-narrow	0.4653	0.1800	84.33

batch size, diverse prompts improve Coverage to 0.60, whereas category-narrow prompts reduce it to 0.18. These results indicate that the relation signal depends on the support represented by training batches. The batch-size study changes both the relation set and the number of samples seen at fixed update count, so it measures their combined effect. The batch loss constrains sampled minibatch geometry rather than guaranteeing global distributional coverage.

D.4 One- and two-step generation

Table D.4 evaluates the 1.3B model at one, two, and four steps. CoDMD improves Total by 1.96, 1.37, and 1.08 points, respectively. At one step, the gain is driven by Quality while Semantic decreases by 0.64 points; the two- and four-step settings improve both scores.

Table D.4: VBench at different sampling budgets.

Steps	Method	Quality↑	Semantic↑	Total↑
1	DMD*	80.82	73.63	79.38
1	CoDMD	83.42	72.99	**81.34**
2	DMD*	83.87	78.00	82.70
2	CoDMD	85.16	79.69	**84.07**
4	DMD*	84.53	78.78	83.38
4	CoDMD	85.50	80.27	**84.46**

E User Study Details

To complement the automated metrics, we conduct a blind pairwise user study to compare CoDMD with DMD, rCM, and the 50-step teacher. The study involves 20 evaluators. In each trial, a participant is shown two videos generated from the same text prompt, with the left–right display order randomized. The source methods are hidden from the participant. The participant is asked to judge which video is better with respect to visual quality and semantic alignment to the prompt.

We use the following instruction for all participants:

> You will be shown two videos generated from the same text prompt. Please compare the two videos and select the one that is better overall, considering both visual quality and semantic alignment to the prompt. The presentation order is randomized, and the source methods are not disclosed. Please make your judgment only based on the displayed videos and the prompt.

The aggregated pairwise preference rates are reported in Fig. 4.

This study is minimal risk. Participants are only asked to compare generated videos and no sensitive personal information is collected. Participation is voluntary, and participants may stop at any time.

Compensation. All participants received compensation according to local institutional practice.

Ethics review. No formal IRB or equivalent approval was obtained for this minimal-risk study.

Interface. The user study is conducted with a local side-by-side comparison interface that displays the prompt together with two generated videos. The interface supports both desktop and mobile layouts, with A/B video panels adapted to different screen sizes for convenient access across devices. An anonymized screenshot of the interface is shown in Fig. E.1.

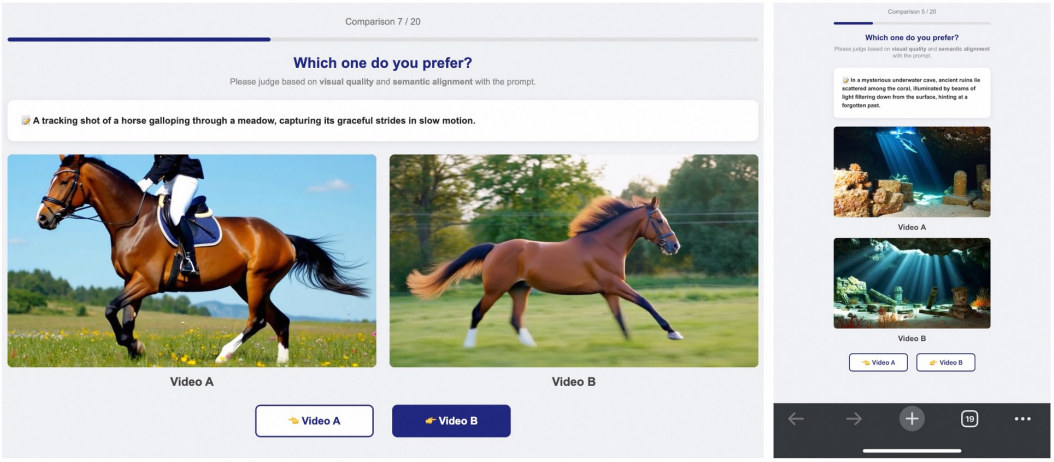


Figure E.1: Anonymized interface used in the blind pairwise user study. The interface displays the prompt together with two generated videos in randomized left–right order and supports both desktop and mobile layouts.

F Additional Visual Comparisons

The following visual comparisons illustrate motion, appearance, and prompt alignment across the evaluated methods.

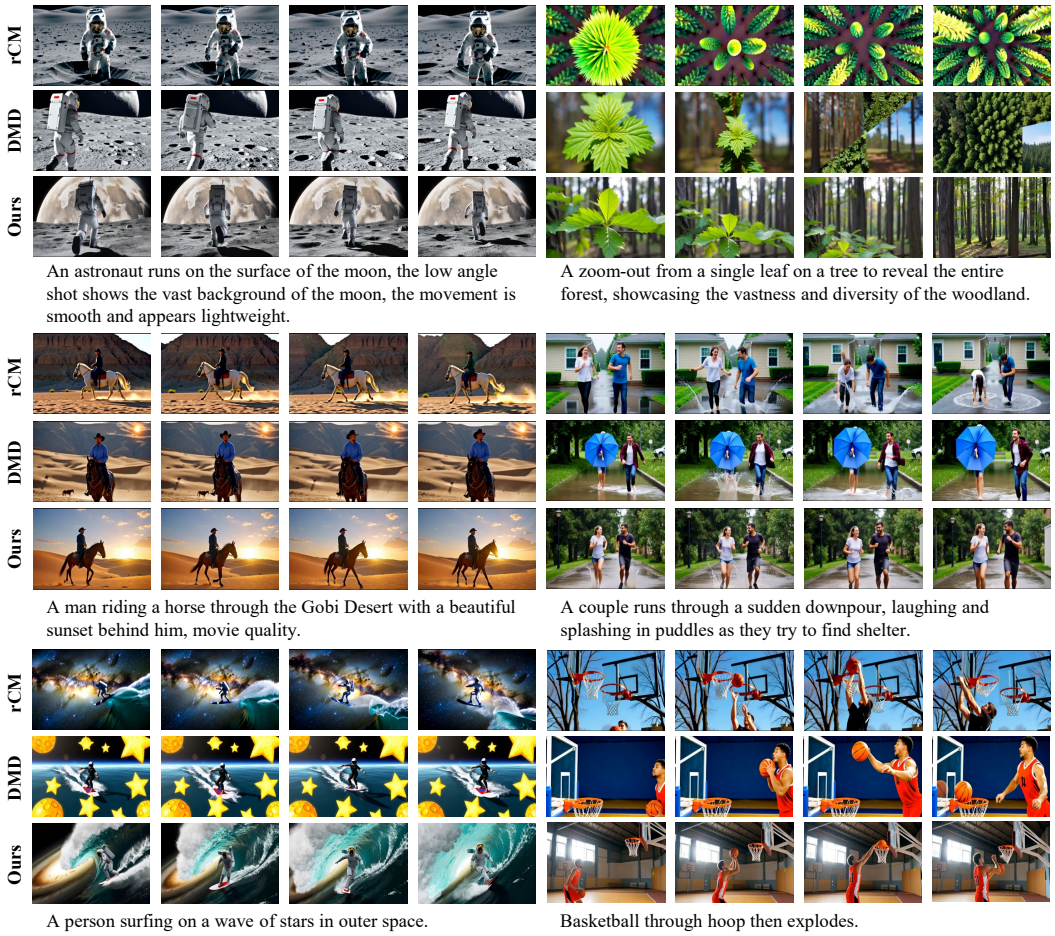


Figure F.1: Additional qualitative comparisons, first set.

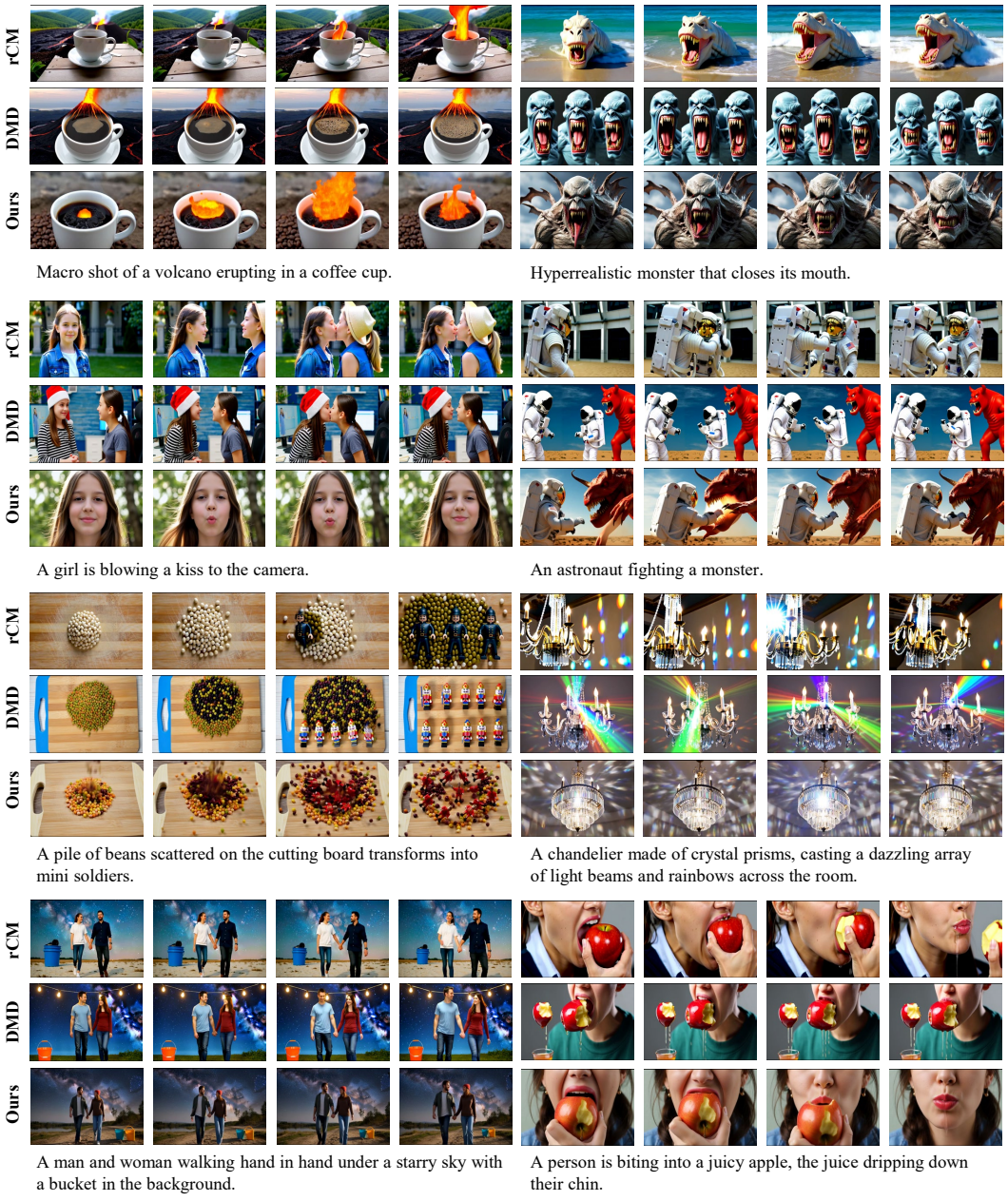


Figure F.2: Additional qualitative comparisons, second set.

G Notation

Table G.1 collects the symbols used in the method and in the derivations that follow in Appendix H.

Table G.1: Notation used throughout the paper.

Symbol	Dimension	Description
Video data & dimensions		
F, C, H, W	$\mathbb{N}$	Number of frames,channels, height, width
$\mathbf{x}_0$	$\mathbb{R}^{F\times C\times H\times W}$	Clean video from the real distribution p_{real}
$\mathbf{z}$	$\mathbb{R}^{d_z}$	Standard Gaussian noise input to the generator
$\mathbf{x}^{(b)}$	$\mathbb{R}^{F\times C\times H\times W}$	b-th generated video, $\mathbf{x}^{(b)} = G_\theta(\mathbf{z}^{(b)})$
$\mathbf{x}_t$	$\mathbb{R}^{F\times C\times H\times W}$	Diffused video at noise level $t \in [0, 1]$
B, B_{g}	$\mathbb{N}$	Per-GPU and global batch size
Diffusion & Score		
t	$\mathbb{R}$	Continuous noise level
α_t, σ_t	$\mathbb{R}$	Noise schedule coefficients; $\alpha_t^2 + \sigma_t^2 = 1$, $\alpha_T \approx 0$
μ_{real}	—	Frozen teacher denoiser (pre-trained, not updated)
μ^{ϕ}_{fake}	—	Online fake denoiser, trained to track p_{fake}
s_{real}	$\mathbb{R}^{F\times C\times H\times W}$	Score of real distribution: $-(\mathbf{x}_t - \alpha_t\mu_{\text{real}})/\sigma_t^2$
s_{fake}	$\mathbb{R}^{F\times C\times H\times W}$	Score of fake distribution: $-(\mathbf{x}_t - \alpha_t\mu^{\phi}_{\text{fake}})/\sigma_t^2$
w_t	$\mathbb{R}$	Noise-level weighting for the KL gradient
DMD Loss		
$\mathbf{\Delta}(\mathbf{x})$	$\mathbb{R}^{F\times C\times H\times W}$	Coordinate-wise score-difference perturbation
$\mathcal{L}_{\text{DMD}}$	$\mathbb{R}$	Score-as-target surrogate: $\frac{1}{2}\|\mathbf{x} - \text{stopgrad}(\mathbf{x} - \mathbf{\Delta})\|_2^2$
$\mathcal{L}^{\phi}_{\text{denoise}}$	$\mathbb{R}$	Denoising loss for fake model: $\|\mu^{\phi}_{\text{fake}}(\mathbf{x}_t, t) - \mathbf{x}\|_2^2$
Copula & CoDMD		
$f^{(i)}, F^{(i)}$	—	Marginal PDF and CDF of the i-th coordinate
$c(\cdot)$	—	Copula density on $[0, 1]^D$
s^{marg}	$\mathbb{R}^D$	Marginal (coordinate-wise) component of the score
s^{cop}	$\mathbb{R}^D$	Copula (dependency) component of the score
Relational Loss		
$\mathbf{S}^{\text{batch}}_{\text{stu/real/fake}}$	$\mathbb{R}^{B_{\text{g}}\times B_{\text{g}}}$	Batch-level cosine similarity matrices from $\mathbf{x}$, $\mu_{\text{real}}(\mathbf{x}_t, t)$, $\mu^{\phi}_{\text{fake}}(\mathbf{x}_t, t)$
$\mathbf{S}^{\text{frame}}_{\text{stu/real/fake}}$	$\mathbb{R}^{F\times F}$	Frame-level cosine similarity matrices from $\mathbf{x}^{(b)}$, $\mu_{\text{real}}(\mathbf{x}_t^{(b)}, t)$, $\mu^{\phi}_{\text{fake}}(\mathbf{x}_t^{(b)}, t)$
$\mathbf{\Delta}_{\mathbf{S}}$	$\mathbb{R}^{N\times N}$	Relational score-difference: $\mathbf{S}_{\text{fake}} - \mathbf{S}_{\text{real}}$
τ	$\mathbb{R}_{>0}$	Temperature for row-wise softmax
$\mathbf{P}_{\text{tgt}}$	$\mathbb{R}^{N\times N}$	Row-wise softmax distribution of target (stop-grad)
$\mathbf{P}_{\text{stu}}$	$\mathbb{R}^{N\times N}$	Row-wise softmax distribution of student
$\mathcal{L}^{\text{batch}}_{\text{rel}}$	$\mathbb{R}$	Batch-level relational loss ($N = B_{\text{g}}$); Eq. (9)
$\mathcal{L}^{\text{frame}}_{\text{rel}}$	$\mathbb{R}$	Frame-level relational loss ($N = F$, averaged over B); Eq. (10)
$\mathcal{L}_{\text{total}}$	$\mathbb{R}$	Generator objective: $\mathcal{L}_{\text{DMD}} + \lambda_{\text{batch}}\mathcal{L}^{\text{batch}}_{\text{rel}} + \lambda_{\text{frame}}\mathcal{L}^{\text{frame}}_{\text{rel}}$; Eq. (11)
$\lambda_{\text{batch}}, \lambda_{\text{frame}}$	$\mathbb{R}_{\geq 0}$	Weights of the batch- and frame-level relational losses

H Theoretical Analysis

This appendix supplies the formal statements and proofs of the three propositions introduced in §3.2 of the main paper. These results identify the structural mechanism behind DMD's fragility in the few-step video regime and motivate the design choices in CoDMD.

Let $X = (X_1, \dots, X_F)$ denote a video with F frames, P the teacher-induced distribution and Q the student-induced distribution, with P_t, Q_t the frame-wise marginals. For a symmetric distance $d(\cdot, \cdot)$ on score-based features, the pairwise relational statistic

$$\mathcal{D}(Q) = \mathbb{E}_{Z,Z' \sim Q}\big[d(Z, Z')\big]$$

captures the relational geometry of a distribution. This statistic motivates relations across samples and frames; it is not identical to the implemented residual-target cosine-KL loss.

H.1 Marginal Matching Does Not Identify Relational Structure

We first formalize the limitation of coordinate-wise or marginal supervision.

Proposition 1 (Relational non-identifiability under marginal matching)**.** *There exist two video distributions P and Q such that*

$$P_t = Q_t, \qquad \forall t \in \{1, \dots, T\},$$

but their temporal relational structures differ. Consequently, any training objective that depends only on the frame-wise marginals $\{Q_t\}_{t=1}^T$ cannot distinguish Q from P.

Proof. It suffices to construct a two-frame example. Let $U \sim \mathcal{N}(0, 1)$ and define

$$P: \quad (X_1, X_2) = (U, U),$$

while

$$Q: \quad (X_1, X_2) = (U, -U).$$

For both distributions, each individual frame marginal is standard Gaussian:

$$P_1 = P_2 = Q_1 = Q_2 = \mathcal{N}(0, 1).$$

Hence any objective depending only on the frame-wise marginals assigns the same value to P and Q.

However, their temporal relations are different. For example,

$$\mathbb{E}_P[X_1 X_2] = 1, \qquad \mathbb{E}_Q[X_1 X_2] = -1.$$

Equivalently, the expected squared frame difference satisfies

$$\mathbb{E}_P\big[(X_1 - X_2)^2\big] = 0, \qquad \mathbb{E}_Q\big[(X_1 - X_2)^2\big] = 4.$$

Thus the two distributions have identical frame-wise marginals but opposite temporal dependence. This proves that marginal matching alone does not identify relational structure. □

Motion, layout consistency and semantic coherence depend on couplings among frames and among generated samples; these couplings can change even when all individual marginals remain matched.

Clarification for exact DMD. For positive differentiable densities on a connected domain, equality $\nabla \log q = \nabla \log p$ implies that $\log(q/p)$ is constant; normalization then gives $q = p$. Gaussian perturbation with $\sigma_t > 0$ yields smooth positive densities on the full space, so exact joint-score equality identifies the perturbed distribution. This statement concerns equality of score fields, not a vanishing parameter gradient after expectation and Jacobian projection.

A different observation applies to marginal-only objectives. Let $U \sim \mathcal{N}(0, 1)$ and consider $P^+ = \mathrm{law}(U, U)$ and $P^- = \mathrm{law}(U, -U)$. Both coordinates have standard Gaussian marginals, but the cross-moments are $+1$ and -1. Marginal-only criteria therefore cannot distinguish them. Exact DMD can: after adding Gaussian noise, rotate coordinates to $u = (x_1 + x_2)/\sqrt{2}$, $v = (x_1 - x_2)/\sqrt{2}$, and set

$$a_t = 2\alpha_t^2 + \sigma_t^2, \qquad b_t = \sigma_t^2, \qquad \gamma_t = \frac{2\alpha_t^2}{a_t b_t} = \frac{1}{b_t} - \frac{1}{a_t}.$$

The covariances are $A_+ = \mathrm{diag}(a_t, b_t)$ and $A_- = \mathrm{diag}(b_t, a_t)$. If P^+ is the teacher and P^- the student, the score residual is

$$s_{P_t^-} - s_{P_t^+} = (A_+^{-1} - A_-^{-1})(u, v)^\top = -\gamma_t(u, -v)^\top,$$

which is nonzero except on a set of measure zero. The example isolates dependence information without asserting its absence from joint scores.

A missing-branch signal under the student measure. Now let the teacher contain both branches and the student only one:

$$P = \tfrac{1}{2}P^+ + \tfrac{1}{2}P^-, \qquad Q = P^+.$$

Writing $w_-(x_t)$ for the posterior responsibility of the negative branch in P_t, direct Gaussian calculation gives

$$w_-(x_t) = \left[1 + \exp\{\tfrac{\gamma_t}{2}(u^2 - v^2)\}\right]^{-1}, \tag{14}$$

$$s_{Q_t}(x_t) - s_{P_t}(x_t) = \gamma_t w_-(x_t)(u, -v)^\top. \tag{15}$$

Indeed, the mixture score is the responsibility-weighted average of its two component scores, and subtracting it from the positive-branch score gives Eq. (15).

Under Q_t, $u \sim \mathcal{N}(0, a_t)$ and $v \sim \mathcal{N}(0, b_t)$ independently. For $0 < \delta < 1$, the Gaussian small-ball bound and tail bound imply

$$\Pr\{|u| < \delta\sqrt{a_t}\} \le \sqrt{2/\pi}\,\delta, \qquad \Pr\{|v| > \delta\sqrt{a_t/2}\} \le 2\exp(-\delta^2 a_t/(4b_t)).$$

On the complementary event, $u^2 - v^2 \ge \delta^2 a_t/2$, hence

$$w_-(x_t) \le \exp\left(-\frac{\delta^2\alpha_t^2}{2b_t}\right).$$

Thus the exact residual distinguishes the distributions, while the missing-branch responsibility can be exponentially small on typical student samples at separated low-noise scales. This is a signal attenuation example, not a claim that the residual vanishes at every noise level or that the proposed regularizer always recovers every missing branch.

H.2 Reverse-KL Selects Dominant Modes Under Restricted Families

We next formalize the mode-seeking behavior of reverse-KL in a restricted family. Unlike an informal statement about "mode collapse", the following result gives an exact calculation.

Proposition 2 (Reverse-KL selection in a disjoint mixture)**.** *Let the teacher distribution be a mixture*

$$P = \sum_{k=1}^{K} \pi_k P_k, \qquad \pi_k > 0, \quad \sum_{k=1}^{K} \pi_k = 1,$$

where the component distributions P_k have disjoint supports A_k. Suppose the student family is restricted to single-component approximants

$$\mathcal{Q}_{\mathrm{single}} = \{P_1, \ldots, P_K\}.$$

Then

$$D_{\mathrm{KL}}(P_j \,\|\, P) = -\log \pi_j.$$

Therefore,

$$\arg\min_{Q \in \mathcal{Q}_{\mathrm{single}}} D_{\mathrm{KL}}(Q \,\|\, P) = P_{j^\star}, \qquad j^\star \in \arg\max_j \pi_j.$$

That is, reverse-KL selects a dominant mixture component rather than covering all teacher-supported components.

Proof. Since the supports are disjoint, for $x \in A_j$ we have

$$p(x) = \pi_j p_j(x).$$

Therefore,

$$\begin{aligned} D_{\mathrm{KL}}(P_j \,\|\, P) &= \int_{A_j} p_j(x) \log \frac{p_j(x)}{p(x)}\, dx \\ &= \int_{A_j} p_j(x) \log \frac{p_j(x)}{\pi_j p_j(x)}\, dx \\ &= \int_{A_j} p_j(x) \log \frac{1}{\pi_j}\, dx \\ &= -\log \pi_j. \end{aligned}$$

Thus minimizing $D_{\mathrm{KL}}(P_j \| P)$ is equivalent to maximizing π_j, proving the claim. □

This result captures the precise sense in which reverse-KL is mode-seeking under a restricted student family. A more general version holds: if Q_j is supported on A_j, then $D_{\mathrm{KL}}(Q_j \| P) = D_{\mathrm{KL}}(Q_j \| P_j) - \log \pi_j$. Reverse-KL therefore trades off per-component fit against mixture weight, but never penalizes the absence of uncovered modes.

H.3 Relational Matching Penalizes Collapsed Solutions

We now show that pairwise relational matching provides a mechanism that standard reverse-KL lacks: it can penalize a student that collapses to one component of a multi-relation teacher distribution.

Let $d(z, z')$ be a nonnegative symmetric distance in feature space, and define

$$\mathcal{D}(Q) = \mathbb{E}_{Z,Z' \sim Q}\big[d(Z, Z')\big].$$

For the teacher mixture $P = \sum_{k=1}^{K} \pi_k P_k$, we have

$$\mathcal{D}(P) = \sum_{a=1}^{K} \sum_{b=1}^{K} \pi_a \pi_b \mathbb{E}_{Z \sim P_a,\, Z' \sim P_b}\big[d(Z, Z')\big].$$

The terms with $a \neq b$ encode cross-mode relational diversity.

Proposition 3 (Relational penalty for single-mode collapse)**.** *Assume the teacher distribution is a mixture $P = \sum_{k=1}^{K} \pi_k P_k$ with disjoint relational modes. Suppose that for some constants $0 \leq \epsilon < \Delta$,*

$$\mathbb{E}_{Z,Z' \sim P_k}\big[d(Z, Z')\big] \leq \epsilon, \qquad \forall k,$$

while for all $a \neq b$,

$$\mathbb{E}_{Z \sim P_a,\, Z' \sim P_b}\big[d(Z, Z')\big] \geq \Delta.$$

Then any single-mode student $Q = P_j$ satisfies

$$\big|\mathcal{D}(Q) - \mathcal{D}(P)\big| \geq \Delta\Big(1 - \sum_{k=1}^{K} \pi_k^2\Big) - \epsilon.$$

The lower bound is positive only when $\Delta(1 - \sum_k \pi_k^2) > \epsilon$. This abstract discrepancy does not establish a lower bound for the implemented cosine-KL loss.

Proof. For the teacher distribution,

$$\begin{aligned} \mathcal{D}(P) &= \sum_{k=1}^{K} \pi_k^2 \mathbb{E}_{Z,Z' \sim P_k}[d(Z, Z')] + \sum_{a \neq b} \pi_a \pi_b \mathbb{E}_{Z \sim P_a, Z' \sim P_b}[d(Z, Z')] \\ &\geq \sum_{a \neq b} \pi_a \pi_b \Delta \\ &= \Delta\Big(1 - \sum_{k=1}^{K} \pi_k^2\Big). \end{aligned}$$

For a single-mode student $Q = P_j$, by assumption,

$$\mathcal{D}(Q) = \mathbb{E}_{Z,Z' \sim P_j}[d(Z, Z')] \leq \epsilon.$$

Therefore,

$$\mathcal{D}(P) - \mathcal{D}(Q) \geq \Delta\Big(1 - \sum_{k=1}^{K} \pi_k^2\Big) - \epsilon,$$

which proves the claim. □

This result formalizes why pairwise relational matching complements DMD. Reverse-KL may select a dominant component under the restricted few-step student family; a relational objective measures the loss of cross-mode geometry through pairwise distances or similarities.

H.4 Implication for CoDMD

The propositions motivate explicit relational supervision under their stated assumptions. The marginal-only example is not a counterexample to exact joint-score matching. The reverse-KL calculation requires a restricted student family, and the pairwise discrepancy is an abstract statistic rather than our implemented loss. CoDMD introduces an explicit relation-space update, whose exact gradient is derived below.

H.5 Gradient Derivation of the Relationship-Distillation Loss

This appendix proves the claim made in §3.3: the gradient of the relationship-distillation loss $\mathcal{L}_{\text{rel}}$ with respect to the student similarity matrix $\mathbf{S}_{\text{stu}}$ is exactly a difference of row-softmax probabilities, and locally a softmax-Jacobian transformation of the relational score difference $\mathbf{\Delta_S} = \mathbf{S}_{\text{fake}} - \mathbf{S}_{\text{real}}$. This makes the analogy with the DMD gradient (1) mathematically precise: CoDMD and DMD are two instances of the *same* score-as-target construction, applied at different granularities (coordinates vs. pairwise interactions).

Let $\mathbf{S}_{\text{stu}}, \mathbf{S}_{\text{real}}, \mathbf{S}_{\text{fake}} \in \mathbb{R}^{N \times N}$ be the cosine similarity matrices of §3.3, and $\mathbf{\Delta_S} = \mathbf{S}_{\text{fake}} - \mathbf{S}_{\text{real}}$. For brevity, write

$$\mathbf{A} \;=\; \tau^{-1}\mathbf{S}_{\text{stu}}, \qquad \mathbf{B} \;=\; \tau^{-1}\big(\mathbf{S}_{\text{stu}} - \mathbf{\Delta_S}\big) \;=\; \mathbf{A} - \tau^{-1}\mathbf{\Delta_S}.$$

Both matrices are passed through a row-wise softmax, and the gradient through the target $\mathbf{P}_{\text{tgt}}$ is detached:

$$\mathbf{P}_{\text{stu}}^{i,:} \;=\; \sigma(\mathbf{A}^{i,:}), \qquad \mathbf{P}_{\text{tgt}}^{i,:} \;=\; \text{stopgrad}\big(\sigma(\mathbf{B}^{i,:})\big),$$

where $\sigma(\mathbf{v})_j \;=\; e^{v_j} / \sum_k e^{v_k}$ denotes the softmax along a row. The relationship-distillation loss (Eq. (8)) is the row-averaged KL,

$$\mathcal{L}_{\text{rel}} \;=\; \frac{1}{N}\sum_{i=1}^{N} \text{KL}\big(\mathbf{P}_{\text{tgt}}^{i,:} \,\big\|\, \mathbf{P}_{\text{stu}}^{i,:}\big) \;=\; \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} P_{\text{tgt},j}^{i}\,\big[\log P_{\text{tgt},j}^{i} - \log P_{\text{stu},j}^{i}\big]. \tag{16}$$

We will compute $\partial\mathcal{L}_{\text{rel}}/\partial\mathbf{S}_{\text{stu}}$ in two steps: (1) the exact closed-form gradient, then (2) its first-order expansion in the small parameter $\tau^{-1}\mathbf{\Delta_S}$.

Step 1: Exact gradient via the softmax Jacobian

Because $\mathbf{P}_{\text{tgt}}$ is detached, only $\log P_{\text{stu},j}^{i}$ in Eq. (16) carries gradient. Fix a row i and recall the standard log-softmax identity

$$\log P_{\text{stu},j}^{i} \;=\; A_{ij} \;-\; \log\sum_{k} e^{A_{ik}}.$$

Differentiating with respect to A_{im} yields

$$\frac{\partial \log P_{\text{stu},j}^{i}}{\partial A_{im}} \;=\; \mathbf{1}[j{=}m] \;-\; P_{\text{stu},m}^{i}. \tag{17}$$

Substituting Eq. (17) into the row-i contribution of Eq. (16) (and using $\sum_j P^i_{\text{tgt},j} = 1$),

$$\frac{\partial}{\partial A_{im}}\Big[-\sum_j P^i_{\text{tgt},j} \log P^i_{\text{stu},j}\Big] = -\sum_j P^i_{\text{tgt},j}\big(\mathbf{1}[j{=}m] - P^i_{\text{stu},m}\big) = P^i_{\text{stu},m} - P^i_{\text{tgt},m}. \tag{18}$$

Chaining through $A_{im} = \tau^{-1}(S_{\text{stu}})_{im}$,

$$\frac{\partial \mathcal{L}_{\text{rel}}}{\partial (S_{\text{stu}})_{im}} \;=\; \frac{1}{N\tau}\big(P^i_{\text{stu},m} - P^i_{\text{tgt},m}\big). \tag{19}$$

This is exact for any $\tau > 0$. The student is updated only where its softmax-normalised relational profile differs from the (detached) target profile, mirroring DMD's principle of "update only where critic and teacher disagree."

Step 2: First-order expansion in τ^{-1}

To make the link with the score difference $\mathbf{\Delta_S}$ explicit, we expand $\mathbf{P}_{\text{tgt}} - \mathbf{P}_{\text{stu}}$ in powers of $\boldsymbol{\delta}^i = \tau^{-1}(\mathbf{\Delta_S})^{i,:}$. The expansion is informative when the per-row entries of $\boldsymbol{\delta}^i$ are moderate; we verify this empirically in the "When does the linearisation hold" paragraph below.

Softmax differential. The Jacobian of the softmax at a point $\mathbf{v} \in \mathbb{R}^N$ is the well-known matrix

$$J(\mathbf{v}) \;=\; \operatorname{diag}(\sigma(\mathbf{v})) - \sigma(\mathbf{v})\,\sigma(\mathbf{v})^\top. \tag{20}$$

Setting $\mathbf{v} = \mathbf{A}^{i,:}$ and noting that $\mathbf{B}^{i,:} = \mathbf{A}^{i,:} - \boldsymbol{\delta}^i$, a first-order Taylor expansion gives

$$\sigma(\mathbf{B}^{i,:}) \;=\; \sigma(\mathbf{A}^{i,:}) \;-\; J(\mathbf{A}^{i,:})\,\boldsymbol{\delta}^i \;+\; O\big(\|\boldsymbol{\delta}^i\|^2\big). \tag{21}$$

Hence

$$\mathbf{P}^{i,:}_{\text{tgt}} - \mathbf{P}^{i,:}_{\text{stu}} \;=\; -J(\mathbf{A}^{i,:})\,\boldsymbol{\delta}^i \;+\; O\big(\|\boldsymbol{\delta}^i\|^2\big). \tag{22}$$

Substituting into the exact gradient. Plugging Eq. (22) into Eq. (19),

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{rel}}}{\partial (S_{\text{stu}})_{im}} &= \frac{1}{N\tau}\big(P^i_{\text{stu},m} - P^i_{\text{tgt},m}\big) \\ &= \frac{1}{N\tau}\big[J(\mathbf{A}^{i,:})\,\boldsymbol{\delta}^i\big]_m \;+\; O(\tau^{-3}) \\ &= \frac{1}{N\tau^2}\big[J(\mathbf{A}^{i,:})\,(\mathbf{\Delta_S})^{i,:}\big]_m \;+\; O(\tau^{-3}),\end{aligned} \tag{23}$$

where we used $\boldsymbol{\delta}^i = \tau^{-1}(\mathbf{\Delta_S})^{i,:}$. In matrix form,

$$\nabla_{\mathbf{S}_{\text{stu}}} \mathcal{L}_{\text{rel}} \;=\; \frac{1}{N\tau^2}\,\mathbf{M}(\mathbf{S}_{\text{stu}};\tau)\;\mathbf{\Delta_S} \;+\; O(\tau^{-3}), \tag{24}$$

where the row-wise positive semidefinite preconditioner $\mathbf{M}(\mathbf{S}_{\text{stu}};\tau)$ collects the row-Jacobians $J(\tau^{-1}\mathbf{S}^{i,:}_{\text{stu}})$.

Interpretation. Eq. (24) establishes the central claim: to first order in τ^{-1}, the gradient of $\mathcal{L}_{\text{rel}}$ on $\mathbf{S}_{\text{stu}}$ is a PSD linear map applied to the relational score difference $\mathbf{\Delta_S}$. Two consequences follow.

(i) Direction. The matrix $\mathbf{M}$ is row-wise PSD (it is the softmax-Jacobian at $\mathbf{A}^{i,:}$), so for each row i the gradient points along the same half-space as $\mathbf{\Delta}^{i,:}_{\mathbf{S}}$ up to a bounded reweighting. This first-order relation holds in similarity space; projection through the generator Jacobian need not preserve the untransformed residual direction.

(ii) Vanishing supervision. Whenever critic and teacher agree relationally, $\mathbf{\Delta_S} \to \mathbf{0}$ and the gradient vanishes; the student is updated only where they disagree. This is the relational analogue of the DMD property that the gradient (1) vanishes whenever $s_{\text{fake}} = s_{\text{real}}$.

Step 3: Mirroring the DMD gradient via the chain rule

To complete the parallel with Eq. (1), chain Eq. (24) through the construction of $\mathbf{S}_{\text{stu}}$ and the generator. Recall that $\mathbf{S}_{\text{stu}} = \mathbf{S}(\bar{\mathbf{x}})$ with $\bar{\mathbf{x}}^{(b)} = \text{pool}\big(G_\theta(\mathbf{z}^{(b)})\big)$, so

$$\frac{\partial \mathcal{L}_{\text{rel}}}{\partial \theta} = \sum_b \underbrace{\frac{\partial \mathcal{L}_{\text{rel}}}{\partial \mathbf{S}_{\text{stu}}}}_{\text{Eq. (24)}} : \underbrace{\frac{\partial \mathbf{S}_{\text{stu}}}{\partial \bar{\mathbf{x}}^{(b)}}}_{\text{cos-sim Jac.}} \cdot \underbrace{\frac{\partial \bar{\mathbf{x}}^{(b)}}{\partial G_\theta(\mathbf{z}^{(b)})}}_{\text{linear pool}} \cdot \frac{dG_\theta(\mathbf{z}^{(b)})}{d\theta}, \tag{25}$$

where ":" denotes the matrix Frobenius pairing. Substituting Eq. (24),

$$\frac{\partial \mathcal{L}_{\text{rel}}}{\partial \theta} \propto \sum_b \boldsymbol{\Delta}_{\mathbf{S}} \text{ (propagated through bilinear pool/cosine) } \cdot \frac{dG_\theta(\mathbf{z}^{(b)})}{d\theta} + O(\tau^{-3}). \tag{26}$$

Compare with the DMD gradient (1),

$$\nabla_\theta D_{\text{KL}} \simeq \mathbb{E}_{\mathbf{z},t}\Big[\underbrace{w_t \alpha_t (s_{\text{fake}} - s_{\text{real}})}_{\boldsymbol{\Delta}(\mathbf{x})} \tfrac{dG_\theta(\mathbf{z})}{d\theta}\Big].$$

Both gradients have the same structure

$$\big(\text{score-difference signal}\big) \times \big(\text{generator Jacobian}\big),$$

with $\boldsymbol{\Delta}(\mathbf{x})$ and $\boldsymbol{\Delta}_{\mathbf{S}}$ playing identical roles at coordinate level and at relational level respectively. This formalises the statement in §3.3 that CoDMD follows the detached-target principle of DMD at a relational granularity, with an additional softmax Jacobian in the local expansion.

When does the linearisation hold? The Taylor expansion requires a sufficiently small row perturbation $\tau^{-1}\boldsymbol{\Delta}_{\mathbf{S}}$. Boundedness alone does not ensure a small approximation error at $\tau = 0.1$. Training uses the exact gradient in Eq. (19) and does not depend on the approximation being accurate.

Why the loss is not just an MSE on $\mathbf{S}_{\text{stu}}$. A simpler choice would be to write $\frac{1}{2}\|\mathbf{S}_{\text{stu}} - \text{stopgrad}(\mathbf{S}_{\text{stu}} - \boldsymbol{\Delta}_{\mathbf{S}})\|_F^2$, mirroring Eq. (3). Its gradient with respect to $\mathbf{S}_{\text{stu}}$ equals $\boldsymbol{\Delta}_{\mathbf{S}}$ directly, while the softmax-KL gradient of Eq. (24) is a Jacobian-transformed version of $\boldsymbol{\Delta}_{\mathbf{S}}$ with the PSD preconditioner $\mathbf{M}$. Both are therefore driven by the same underlying signal, but the softmax-KL form has two practical advantages on similarity matrices: (a) row-wise softmax is invariant to constant row shifts, but not scaling; feature-norm normalization comes from cosine similarity; and (b) the temperature τ controls how peaked the relational target is, which empirically stabilises training when $\boldsymbol{\Delta}_{\mathbf{S}}$ is noisy at early steps. Eq. (19) shows that this stability comes at no cost in fidelity to the score-as-target principle.

Granularity-specific instantiations. The same derivation applies row-wise to both $\mathcal{L}_{\text{rel}}^{\text{batch}}$ ($N = B_{\text{g}}$) and $\mathcal{L}_{\text{rel}}^{\text{frame}}$ ($N = F$), so the conclusion of Eq. (24) carries through unchanged for the two granularities used in our final loss Eq. (11). The only difference is which axis is contracted by the pooling operator $\partial \bar{\mathbf{x}} / \partial G_\theta(\mathbf{z})$ in Eq. (25).

I Limitations and Future Work

Representation granularity. The relational matrices in CoDMD are constructed from globally pooled representations that average over temporal/spatial dimensions while preserving channels. This pooling discards fine-grained spatial detail, such as local texture patterns and per-region layout structure. Replacing the global average with learned or attention-based pooling could retain richer spatial information and strengthen the relational signal, potentially improving performance on spatially demanding metrics such as object consistency and scene composition.

Loss weighting. The two relational weights λ_{batch} and λ_{frame} are currently set as fixed hyperparameters throughout training. An adaptive schedule that adjusts the batch-to-frame balance according to training dynamics, for instance by monitoring the ratio of relational and coordinate-wise gradient norms, may further stabilize training and reduce the need for manual tuning.

Scope of validation. We have validated CoDMD on 4-step text-to-video distillation with the Wan-2.1 diffusion backbone at two parameter scales. Additional evaluations cover low-NFE sampling (Appendix D.4) and causal 10-second generation (Section 4.3); other backbones, image-to-video conditioning, and longer rollouts remain open. These settings introduce additional challenges, such as single-step capacity constraints and long-horizon temporal error accumulation, that may require granularity-specific adaptations of the relational objective.

Broader impact. As with all advances that lower the sampling cost of video generators, faster inference enables beneficial applications in creative tools, and content production, but also lowers the barrier for generating synthetic media at scale. Deployments should pair efficiency gains with provenance tracking, watermarking, and content-safety mechanisms to mitigate potential misuse.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims are stated in the abstract and introduction and are supported by the theoretical analysis in Sec. 3 and the experimental results in Sec. 4 and Appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations and scope, including remaining failure cases and practical constraints, in Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper states the assumptions used in the analysis and provides derivations and proofs in Sec. 3.2 and Appendix H.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the model setting, training setup, baselines, evaluation protocol, and ablations in Sec. 4 and Appendix. A, which provide the information needed to reproduce the main results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We do not release code or training scripts at submission time. The paper and supplementary material provide implementation details, model settings, and evaluation to support reproducibility. We plan to release code and checkpoints upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the training setup, model scales, baselines, metrics, and key hyperparameters in Sec. 4 and Appendix. A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix D.2 reports means, sample standard deviations, and 95% Student-t intervals across six generation-seed configurations at fixed checkpoints. These measure inference variability, not independent training runs. User-study results are reported as aggregate rates for 20 evaluators without inferential statistics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the hardware and runtime used for the main experiments in Appendix A.1. Each training run is conducted on 32 NVIDIA A100 GPUs with 80GB memory and takes approximately 1.5 days.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and believe the research conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses both potential benefits and misuse risks of improved video generation, including possible malicious uses such as deceptive synthetic media, in Appendix I.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We do not release a new public video generation model or dataset as part of this submission, so no additional release safeguards are applicable at this stage.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the original sources of the models, datasets, and benchmarks used in our experiments.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This submission does not release new datasets, code, or model weights as new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The paper includes a blind pairwise user study with 20 evaluators, and Appendix E provides the study protocol, participant instructions, comparison setup, and compensation details.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: The paper includes a low-risk user study for perceptual comparison of generated videos. No formal IRB or equivalent approval was obtained. The study did not collect sensitive personal data, and participants were informed of the task and its minimal risks.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not an important, original, or non-standard component of the proposed method or experiments.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.