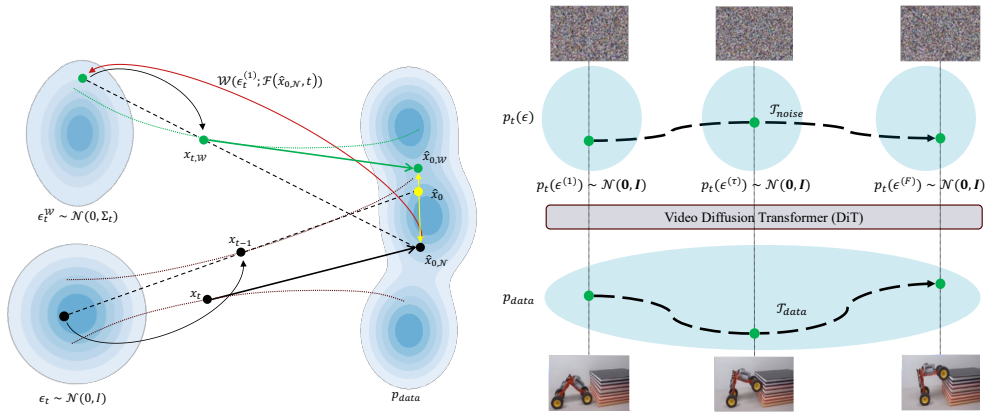
NOISEFLOW: WARP-GUIDED VIDEO GENERATION WITH DISTILLED TEMPORAL PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal coherence in video diffusion models is often imposed using motion correspondences estimated in video space, implicitly assuming that pixel-space motion defines an appropriate transport in latent noise space. We challenge this assumption and introduce NoiseFlow, a framework that learns the temporal noise structure directly from a pretrained video diffusion model. NoiseFlow distills the denoiser-induced local geometry into diffusion-timestep-conditioned autoregressive warp fields that transport Gaussian noise across video frames. We then train a student diffusion model on the resulting family of warped noise priors. During inference, the warp is inferred from the evolving denoising trajectory, and the student provides warped-trajectory guidance to the base model, enabling video generation without a reference video or externally estimated optical flow. Experiments show that NoiseFlow reduces FVD by up to 18%, improves SSIM by up to 43%, and PSNR by up to 19%, and improves several metrics of VBench and VBench2 compared to the unguided base model. These results suggest that denoiser-native temporal priors provide a useful alternative to prescribing video-space correspondences in generative noise space.[1]



(a) (b)

Figure 1: NoiseFlow learns teacher-timestep-dependent temporal noise priors. The warp generator maps the evolving base trajectory to a structured Gaussian noise process, and a student trained on these priors provides warp guidance to the base model at inference.

1 INTRODUCTION

Video diffusion models have advanced rapidly in visual quality, semantic alignment, and motion generation, yet limitations in temporal coherence and controllability remain persistent. A video generator must transform an initially unstructured random field into a sequence whose content remains coherent as objects move, deform, appear, disappear, and interact. Most diffusion and flow-matching

[1]Models and code will be released.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

models nevertheless begin from temporally independent Gaussian noise, leaving the denoising network to construct both the spatial content of individual frames and their cross-frame dependencies during generation. Most video diffusion models nevertheless initialize from temporally independent Gaussian noise, requiring the denoiser to construct cross-frame structure during generation. This motivates encoding part of that structure directly in the noise process.

Previous work has explored this possibility by introducing externally specified temporal correspondences at different stages of the diffusion process. Optical flow estimated from a source video has been used to align intermediate diffusion latents (Bao et al., 2023), restrict temporal attention to corresponding regions (Cong et al., 2024), and propagate edited content across frames (Liang et al., 2024). More directly, recent methods impose temporal structure on the initial diffusion noise itself. Integral-noise methods transport Gaussian samples under prescribed deformation fields while preserving their spatial noise statistics (Chang et al., 2024; Deng et al., 2025), whereas MotionCraft (Montanaro et al., 2024) and Go-with-the-Flow (Burgert et al., 2025) warp noise latents using motion fields obtained from physical simulations or source videos. Work in Daras et al. (2024) and Liu & Vahdat (2025) further demonstrated that training with the corresponding warped videos and noise encourages the denoiser to become equivariant to the imposed transformation. Together, these results establish that the joint temporal structure of the diffusion noise can systematically influence the dynamics generated. However, they also prescribe this structure through optical flow, geometry, or another externally defined correspondence, requiring that the appropriate temporal transport in noise space be known in advance.

Deriving correlations from the optical flow in the data-space imposes a strong assumption that the motion estimated between the image pixels provides an appropriate transport rule for the latent random variables from which a nonlinear denoiser constructs the video. Such approaches implicitly assume that the motion of the image-space provides an appropriate transport in the latent noise space. However, because the denoising map is nonlinear and context-dependent, useful noise-space correspondences do not need to coincide with physical motion during standard training (Liu & Vahdat, 2025; Daras et al., 2024).

We introduce NoiseFlow, a framework that learns this temporal prior directly in the noise space. Given a video during training, a noise generator predicts an autoregressive warp operator $\mathcal{W}$. Starting from a Gaussian noise frame, the operator constructs subsequent noise frames by advecting noise from previously generated frames while introducing fresh Gaussian innovations in regions that require it. Unlike methods that prescribe warps using optical flow, NoiseFlow learns the operator through the downstream diffusion objective. NoiseFlow offers several potential advantages over prescribed optical-flow warping. First, it allows the noise process to encode temporal relationships that correspond to the temporal organization of the data space. Second, it leaves the denoiser to determine which latent transformations are useful, rather than assuming that the geometry of image space should be copied into noise space. Third, NoiseFlow supports video generation without requiring reference video during inference, unlike prior methods. It can generate novel videos from scratch. Finally, because the learned operator acts via denoiser rather than via a prescription derived from optical flow, it can provide temporal organization beyond motion. To our knowledge, NoiseFlow is the first method to distill the warped noise structure from the underlying diffusion model.

Our contributions are threefold. (1) We introduce NoiseFlow, a causal noise-warping framework that learns temporally structured diffusion noise from the local geometry of a pretrained video denoiser, without prescribing an optical-flow or other externally defined correspondence field. (2) We develop a structure-preserving stochastic transport that introduces learned cross-frame dependence while maintaining standard-Gaussian frame marginals, together with a denoiser-pullback objective for learning its spatial warp field. (3) We train a teacher-timestep-conditioned student model on the resulting family of warped diffusion paths and use it to provide orthogonal warped-trajectory guidance during Text-to-Video (T2V) sampling, requiring neither a reference video nor externally estimated motion at inference. Across the evaluated guidance scales, NoiseFlow improves FVD by up to approximately 18% over the unguided CogVideoX-2B baseline while improving several temporal metrics.

In the subsequent sections, we first introduce the preliminaries that are key to understanding our method in Section 2. Then we describe the method in Section 3, followed by a description of

the experiments and their results in Section 4. We conclude our paper with a brief discussion in Section 5.

2 PRELIMINARIES

Diffusion models. In variance-preserving (VP) diffusion models, a clean sample $\boldsymbol{x}_0 \sim p_{\text{data}}$ is progressively corrupted according to

$$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \qquad \alpha_t^2 + \sigma_t^2 = 1, \tag{1}$$

where the signal-to-noise ratio decreases as $t \in [0, 1]$ moves from 0 to 1. Writing $\alpha_t = \cos\theta, \sigma_t = \sin\theta$, we consider the velocity parameterization,

$$\boldsymbol{v}_t^{\text{diff}} = \frac{d\boldsymbol{x}_t}{d\theta} = \alpha_t \boldsymbol{\epsilon} - \sigma_t \boldsymbol{x}_0. \tag{2}$$

A denoising model $\mathcal{G}_\phi(\boldsymbol{x}_t, t, c)$, optionally conditioned on additional information c, is trained at each noise level and may be parameterized to predict clean sample, noise, score, or velocity (Ho et al., 2020; Song et al., 2021). We denote the clean sample estimate implied by any of these parameterizations as $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t)$

Temporally correlated Gaussian noise. For video-shaped latents, independently sampling Gaussian noise for every frame discards temporal correspondence, whereas directly warping discrete white noise by standard image resampling generally alters its statistics. Distribution-preserving noise transport provides constructions for deforming white noise while retaining its Gaussian marginals (Chang et al., 2024; Deng et al., 2025). We denote such a transport on a noise frame $\boldsymbol{\epsilon}_f$ by $\boldsymbol{Q}_f$, with the marginal-preservation property

$$\boldsymbol{\epsilon}_f \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \quad \Longrightarrow \quad \boldsymbol{\epsilon}_{f+1} = \boldsymbol{Q}_f(\boldsymbol{\epsilon}_f) \ \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{3}$$

Applying such transports across frames yields a temporally correlated noise sequence $\boldsymbol{\epsilon}_{\{1:F\}}$ whose individual frame marginals remain Gaussian, while their joint distribution need not factorize over time.

$$p(\boldsymbol{\epsilon}_f) = \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad f = 1, \ldots, F, \qquad p(\boldsymbol{\epsilon}_{\{1:F\}}) \neq \prod_{f=1}^{F} p(\boldsymbol{\epsilon}_f). \tag{4}$$

Thus, distribution-preserving transport can alter the cross-frame covariance of the diffusion noise without changing its per-frame Gaussian marginals, a distinction that is central to our construction.

3 METHOD

This section is organized as follows. We first present the NoiseFlow warping procedure, which is obtained by frame–time integration of a latent noise-advection Stochastic Differential Equation (SDE). We then introduce the objective functions used to learn the resulting warp field. Finally, we define the student model, which is trained to approximate the trajectory mappings from warped noise to data and to capture their variation across diffusion timesteps.

3.1 NOISEFLOW

Our approach constructs temporally correlated Gaussian noise by estimating a per-frame orthogonal operator $\boldsymbol{Q}_f$, obtained through temporal discretization of the underlying noise field (see appendix B for theoretical details). This field is defined as the SDE,

$$\mathrm{d}\boldsymbol{\epsilon}_\tau = \left(\boldsymbol{L}_{\boldsymbol{v}_\tau} - \frac{1}{2} \boldsymbol{B}_\tau \right) \boldsymbol{\epsilon}_\tau \, \mathrm{d}\tau + \sqrt{\boldsymbol{B}_\tau} \, \mathrm{d}W, \qquad \boldsymbol{B}_\tau = \mathrm{diag}(\beta_\tau), \qquad \beta_\tau(\boldsymbol{p}) = [\nabla \cdot \boldsymbol{v}_\tau(\boldsymbol{p})]^+ \tag{5}$$

where $\tau \geq 0$ represents the latent time dimension, $\boldsymbol{\epsilon}_\tau$ is the instantaneous white noise field, $\boldsymbol{B}_\tau$ carries the innovation noise component modeled as the positive local divergence of the spatial flow, $\boldsymbol{L}_{\boldsymbol{v}_\tau} : \mathbb{R}^{HW} \to \mathbb{R}^{HW}$ is a skew-symmetric matrix representing the transformation operator defined by the instantaneous spatial flow field $\boldsymbol{v}_\tau = (\boldsymbol{v}_x, \boldsymbol{v}_y) \in \mathbb{R}^{2HW}$.

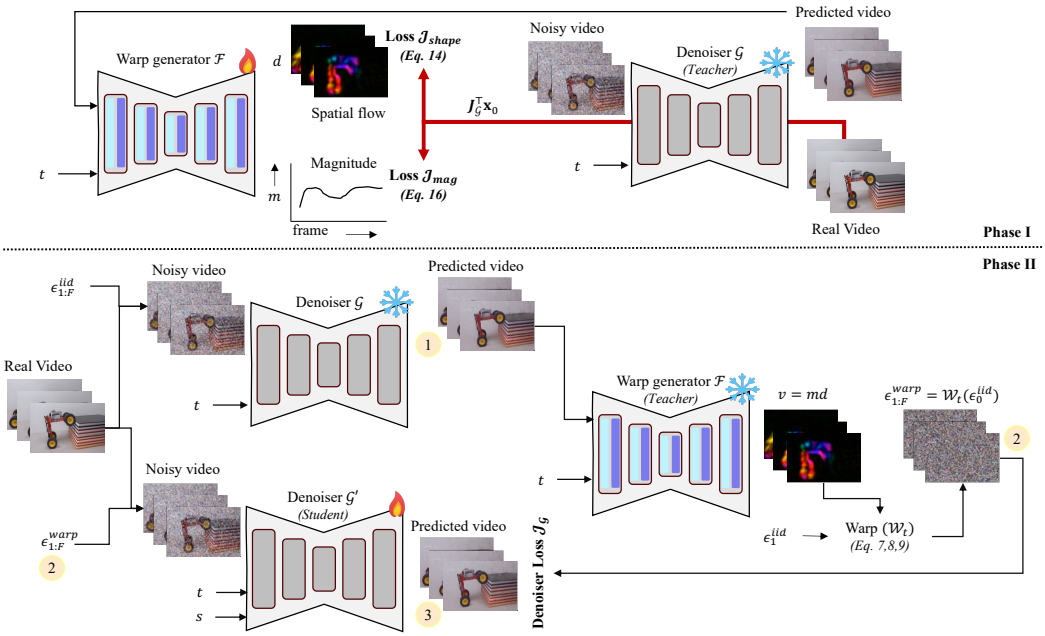


Figure 2: NoiseFlow training. Phase I distills denoiser-sensitive temporal structure into the warp generator F. Phase II freezes F and trains G' on the resulting teacher-indexed warped diffusion paths.

The noise transition from frame f to frame $f+1$ is given by integrating equation 5 over the interval $[\tau_f, \tau_{f+1}]$. Taking $\boldsymbol{v}_f$ to be constant over this interval and writing $\boldsymbol{A}_f = \boldsymbol{L}_{\boldsymbol{v}_f} - \frac{1}{2}\boldsymbol{B}_f$, we have

$$\boldsymbol{\epsilon}_{f+1} = \boldsymbol{\Phi}_f \boldsymbol{\epsilon}_f + \boldsymbol{\eta}_f, \qquad \boldsymbol{\Phi}_f = \exp(\Delta\tau \boldsymbol{A}_f), \qquad \boldsymbol{\eta}_f \sim \mathcal{N}\left(\mathbf{0},\, \boldsymbol{I} - \boldsymbol{\Phi}_f \boldsymbol{\Phi}_f^\top\right). \tag{6}$$

The covariance expression follows from $\boldsymbol{A}_f + \boldsymbol{A}_f^\top = -\boldsymbol{B}_f$ and therefore guarantees that $\boldsymbol{\epsilon}_f \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ implies $\boldsymbol{\epsilon}_{f+1} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Given the initial i.i.d. frame $\boldsymbol{\epsilon}_1$, equation 6 can be used to compute the full video noise prior with innovation noise added along the regions with positive local divergence.

Computing the exact transition in equation 6 is unnecessary. Instead, we approximate the joint dynamics using a Strang splitting between the Ornstein-Uhlenbeck (OU) innovation dynamics and the skew-symmetric advection dynamics. For a substep of length $h = \Delta\tau / K_f$, we define $\boldsymbol{\rho}_{f,h} = \exp\left(-\frac{h}{4}\boldsymbol{B}_f\right)$. A single split substep is

$$\boldsymbol{\epsilon}^{k+\frac{1}{3}} = \boldsymbol{\rho}_{f,h} \boldsymbol{\epsilon}^k + \left(\boldsymbol{I} - \boldsymbol{\rho}_{f,h}^2\right)^{1/2} \xi^{k,1}, \tag{7}$$

$$\epsilon^{k+\frac{2}{3}} = C_h(\boldsymbol{L}_{\boldsymbol{v}_f}) \boldsymbol{\epsilon}^{k+\frac{1}{3}}, \tag{8}$$

$$\boldsymbol{\epsilon}^{k+1} = \boldsymbol{\rho}_{f,h} \boldsymbol{\epsilon}^{k+\frac{2}{3}} + \left(\boldsymbol{I} - \boldsymbol{\rho}_{f,h}^2\right)^{1/2} \xi^{k,2}, \tag{9}$$

where $\xi^{k,1}, \xi^{k,2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and

$$C_h(\boldsymbol{L}) = \left(\boldsymbol{I} - \frac{h}{2}\boldsymbol{L}\right)^{-1} \left(\boldsymbol{I} + \frac{h}{2}\boldsymbol{L}\right) \tag{10}$$

is the Cayley approximation to $\exp(h\boldsymbol{L})$. Each of the three substeps preserves a standard Gaussian marginal, and hence their composition does so exactly. As $K_f \to \infty$, the split transition converges to the continuous SDE. Algorithm 1 describes the full construction.

Given the NoiseFlow construction, the training process is divided into two phases, as described in Figure 2. First, we learn the network parameters of $\mathcal{F}$, responsible for the construction of the warp. In the second phase, we freeze $\mathcal{F}$ and train a student model adapted from the denoiser $\mathcal{G}$ to learn the warp fields as teacher-timestep conditioned trajectory maps. We will describe these phases next.

3.2 LEARNING THE WARP FIELD

Given a pretrained denoiser model, we parameterize $\mathbb{V} : (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_F)$ through a latent video conditioned neural network $\mathcal{F}(\hat{\boldsymbol{x}}_0, t)$, where $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t)$ is the base denoiser prediction using i.i.d. noise seeded latents. We define the objective of this network through the *denoiser pullback*, $(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_F) = \boldsymbol{J}_{\mathcal{G}}^{\top}(\boldsymbol{x}_t, t)\boldsymbol{x}_0$, on clean latent video $\boldsymbol{x}_0$. To identify which noise transports between frames are relevant to the denoiser, we consider the scalar agreement between the transported gradient $\boldsymbol{L}_{\boldsymbol{v}_f}\boldsymbol{g}_f$ and the adjacent frame gradient $\boldsymbol{g}_{f+1}$. We define

$$\left\langle \boldsymbol{g}_{f+1}, \boldsymbol{L}_{\boldsymbol{v}_f}\boldsymbol{g}_f \right\rangle \equiv \langle \boldsymbol{a}_f, \boldsymbol{v}_f \rangle . \tag{11}$$

where $\boldsymbol{a}_f \in \mathbb{R}^{2HW}$ is the pullback of the infinitesimal transport differential to the velocity space (Proposition 10). The field $\boldsymbol{a}_f$ specifies the infinitesimal spatial transport direction that increases the denoiser-sensitive covariance between adjacent frames the most.

Equation 11 identifies a preferred transport direction, but not its scale. Thus, if $\langle \boldsymbol{a}_f, \boldsymbol{v}_f \rangle > 0$, arbitrarily scaling $\boldsymbol{v}_f$ trivially increases the inner product. We therefore separate the velocity into a spatial shape and a global magnitude,

$$\boldsymbol{v}_f = m_f \boldsymbol{d}_f, \tag{12}$$

where $\boldsymbol{d}_f$ retains both the local direction and the relative local magnitude of the velocity field, determining its shape, while m_f represents its global scale. We normalize $\boldsymbol{d}_f$ in the regularized transport metric,

$$\widehat{\boldsymbol{d}}_f = \frac{\boldsymbol{d}_f}{\sqrt{\|\boldsymbol{L}_{\boldsymbol{d}}\|_F^2 + \mu \|\boldsymbol{d}\|^2}}. \tag{13}$$

Similarly, defining $\widehat{\boldsymbol{a}}_f = \boldsymbol{a}_f / \|\boldsymbol{a}_f\|$, the shape objective is

$$\boxed{\mathcal{J}_{\text{shape},f} = -\left\langle \widehat{\boldsymbol{d}}_f, \widehat{\boldsymbol{a}}_f \right\rangle_{HW}}. \tag{14}$$

Equation 14 is of the form $\max_{\boldsymbol{d}} \boldsymbol{a}_f^{\top}\boldsymbol{d} / \sqrt{\boldsymbol{d}^{\top}(\boldsymbol{M} + \mu \boldsymbol{I})\boldsymbol{d}}$, for some positive semi-definite matrix $\boldsymbol{M}$, whose optimum satisfies $\boldsymbol{d}_f^{\star} \propto (\boldsymbol{M} + \mu \boldsymbol{I})^{-1}\boldsymbol{a}_f$ when $\mu > 0$ (Proposition 12).

The scalar m_f is regressed to the characteristic transport scale of $\boldsymbol{g}_f$. Let $\boldsymbol{u}_f = \boldsymbol{L}_{\widehat{\boldsymbol{d}}_f}\boldsymbol{g}_f$. We define the magnitude target $\lambda^{\star}$, and a weight factor w_f

$$\lambda_f^{\star} = \frac{\|\boldsymbol{g}_f\|}{\|\boldsymbol{u}_f\|}, \qquad w_f = \frac{\|\boldsymbol{u}_f\|^2}{\|\boldsymbol{u}_f\|^2 + \|\boldsymbol{g}_f\|^2} |\langle \widehat{\delta \boldsymbol{g}}, \widehat{\boldsymbol{u}}_f \rangle|, \tag{15}$$

where $\widehat{\cdot}$ are the normalized vectors and $\delta \boldsymbol{g} = \boldsymbol{g}_{f+1} - \boldsymbol{g}_f$. Then the magnitude loss is given as

$$\boxed{\mathcal{J}_{\text{mag},f} = w_f \left(m_f - \lambda_f^{\star}\right)^2}, \tag{16}$$

The weight factor w_f discourages the learning of magnitude in the misaligned directions. As $m_f \to \lambda_f^{\star}$, it produces an infinitesimal transport response with magnitude comparable to that of the source pullback.

The complete Phase-I objective for the warp generator is

$$\mathcal{L}_{\mathcal{F}} = \mathbb{E}_{\boldsymbol{x}_0 \sim p_{\text{data}}, t \sim \mathcal{U}(0,1)} \left[\frac{1}{F-1} \sum_{f=1}^{F-1} \left(\mathcal{J}_{\text{shape},f} + \mathcal{J}_{\text{mag},f} \right) \right]. \tag{17}$$

3.3 WARPED DIFFUSION

In Phase II, $\mathcal{F}$ is frozen and we train a student model $\mathcal{G}'$ derived from the denoiser. For a training sample (x_0, c), we first sample a teacher timestep $t \sim \mathcal{U}(0,1)$, and construct an i.i.d.-noise teacher state $\boldsymbol{x}_t^{\mathrm{T}}$ using equation 1. The frozen teacher produces $\widehat{\boldsymbol{x}}_{0,t} = \widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t^{\mathrm{T}}, t, c)$, from which the timestep-dependent warped prior $\boldsymbol{\epsilon}_t^{\mathcal{W}}$ is constructed. Independently, we sample a student timestep $s \sim \mathcal{U}(0,1)$ and define the conditional path as

$$\boldsymbol{x}_{s,t} = \alpha_s \boldsymbol{x}_0 + \sigma_s \boldsymbol{\epsilon}_t^{\mathcal{W}}, \qquad \boldsymbol{v}_{s,t}^{\text{diff}} = \alpha_s \boldsymbol{\epsilon}_t^{\mathcal{W}} - \sigma_s \boldsymbol{x}_0, \qquad \boldsymbol{\epsilon}_t^{\mathcal{W}} = \mathcal{W}\left(\boldsymbol{\epsilon}^{(1)}; \mathcal{F}(\hat{\boldsymbol{x}}_{0,t}, t)\right). \tag{18}$$

The student is optimized using

$$\mathcal{L}_{\text{student}} = \mathbb{E}\left[\left\| \mathcal{G}'(\boldsymbol{x}_{s,t}, s, t, c) - \boldsymbol{v}_{s,t}^{\text{diff}} \right\|_2^2 \right]. \tag{19}$$

Orthogonal warped-trajectory guidance. During inference, the base model is first guided using its standard Classifier-Free Guidance (CFG) (Ho & Salimans, 2021). Let $\widehat{\boldsymbol{x}}_{0,c}$ and $\widehat{\boldsymbol{x}}_{0,\varnothing}$ denote its conditional and unconditional clean predictions, respectively. Then

$$\boldsymbol{d}_t = \widehat{\boldsymbol{x}}_{0,c} - \widehat{\boldsymbol{x}}_{0,\varnothing}, \qquad \widehat{\boldsymbol{x}}_{0,\mathcal{N}} = \widehat{\boldsymbol{x}}_{0,\varnothing} + w\boldsymbol{d}_t, \tag{20}$$

where w is the standard text-CFG scale.

The warped-trajectory model $\mathcal{G}'$ provides a second clean prediction proposal $\widehat{\boldsymbol{x}}_{0,\mathcal{W}}$ at the same timestep $s = t$. Rather than directly composing the two predictions with another CFG style interpolation, which would entangle the proposal correction with the base text-guidance direction $\boldsymbol{d}_t$, we retain only the component of the warped residual that is orthogonal to it. Thus, we compute

$$\boldsymbol{r}_t = \widehat{\boldsymbol{x}}_{0,\mathcal{W}} - \widehat{\boldsymbol{x}}_{0,\mathcal{N}}, \qquad \boldsymbol{r}_t^{\perp} = \left(\boldsymbol{I} - \frac{\boldsymbol{d}_t \boldsymbol{d}_t^{\top}}{\|\boldsymbol{d}_t\|^2}\right) \boldsymbol{r}_t, \tag{21}$$

and use

$$\widehat{\boldsymbol{x}}_0 = \widehat{\boldsymbol{x}}_{0,N} + \gamma \boldsymbol{r}_t^{\perp}. \tag{22}$$

Here γ is the guidance scale.

This construction leaves the component selected by the standard text-CFG unchanged and uses the warped model only in its orthogonal complement. Under a simplifying common-state exact-score idealization assumption, the projection is locally tangent to a level-set of the text-conditioning likelihood. We describe sampling in Algorithm 2 and the idealized motivation in Appendix B.8.

4 Experiments

In this section, we empirically evaluate NoiseFlow. We first describe the experimental setup (datasets and preprocessing, model instantiations and training procedure, evaluation metrics, and comparative baselines). We then present qualitative evidence of the learned latent-noise warp followed by quantitative results on standard benchmarks. Finally, we focus on the warp guidance scale and its trade-off between motion richness and temporal consistency.

4.1 Setup

We train on WISA (Wang et al., 2025b) dataset using CogVideoX-2B (Yang et al., 2025) as the base denoiser. The warp generator is a 25M-parameter U-Net with motion-adapter layers and separate shape and magnitude heads. The student is initialized from the base and augmented with additional teacher-timestep conditioning layers. Full architecture, optimization, and preprocessing details are given in Appendix F.

Evaluation. We evaluate NoiseFlow on VBench (Huang et al., 2024b) metrics *Subject Consistency (SC), Background Consistency (BC), Temporal Flickering (TF), Motion Smoothness (MS)*, and *Dynamic Degree (DD)* and VBench2 (Zheng et al., 2025) metrics *Camera Motion (CM), Dynamic Spatial Relationship (DSR)*, and *Instance Preservation (IP)*. In addition, we report CLIP for text-video alignment, LPIPS, SSIM, and PSNR for paired evaluations, and FVD for distributional comparison against held-out ground truth prompt-video pairs.

Comparative baselines. We evaluated NoiseFlow against similar models of comparable scales. Specifically, we compare with CogvideoX-2B, which is also the implementation base for our method. We also compare this with a finetuned version of this model on the same dataset. Additionally, we evaluate NoiseFlow against the pretrained Wan-1.3B (Wang et al., 2025a) and LTX-Video-2B (HaCohen et al., 2024) baselines. Because these models employ a latent resolution that differs from that of the base model used in NoiseFlow, the comparison is not strictly controlled. Nevertheless, it serves as a useful reference point. The details of the baselines are provided in Table 1. We deliberately exclude comparisons with Video-to-Video (V2V) approaches because such methods require access to a reference video during generation, which introduces additional conditioning information at inference and therefore does not constitute a fair baseline for evaluation.

Table 1: Models used in evaluation and their temporal aspects.

Model	Latent Frames	Output Frames	VAE Temporal Compression	Frame Rate (/s)	Video Length (s)
CogVideoX-2B	13	49	4×	8	6.125
CogVideoX-2B (FT)	13	49	4×	8	6.125
Wan-1.3B	21	81	4×	16	5.063
LTX-Video-2B	21	161	8×	25	6.440
NoiseFlow	13	49	4×	8	6.125

Table 2: Evaluation results. Best results within the same base model architecture and latent space configuration (CogVideoX-2B, marked with '*') are in bold, while best results across all models are underlined.

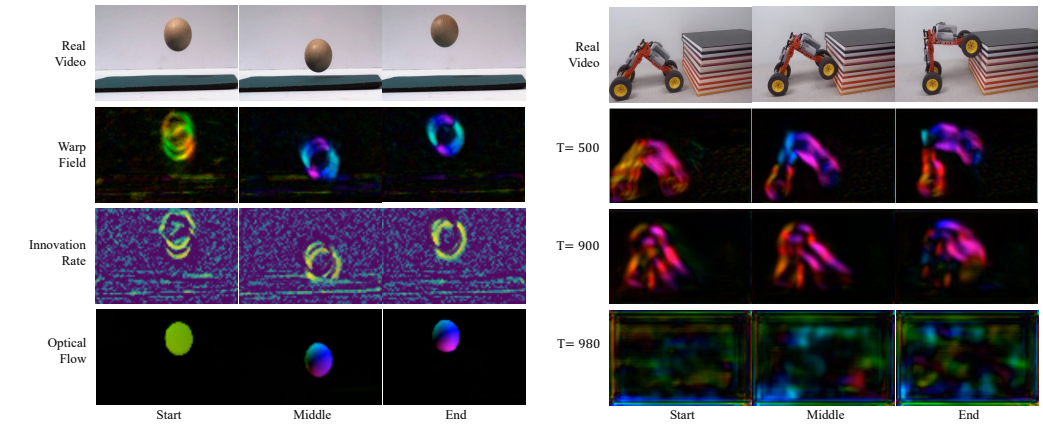
Model						VBench					VBench2		
	CLIP	LPIPS↓	FVD↓	SSIM	PSNR	SC	BC	TF	MS	DD	DSR	CM	IP
$\gamma = 0$ (base)*	0.290	0.670	345.618	0.286	9.903	0.934	0.952	0.974	0.973	0.611	0.184	0.531	**0.825**
NoiseFlow ($\gamma = 0.2$)*	0.299	0.655	311.070	0.315	10.503	0.941	0.954	0.974	0.976	**0.678**	0.213	0.580	0.819
NoiseFlow ($\gamma = 0.5$)*	**0.300**	**0.652**	**283.307**	0.367	11.258	**0.945**	0.957	0.982	0.982	0.606	0.198	**0.593**	0.807
NoiseFlow ($\gamma = 0.8$)*	0.298	0.660	313.816	**0.408**	**11.751**	0.935	**0.959**	**0.988**	**0.987**	0.533	**0.237**	0.583	0.790
$\mathcal{G}' = \mathcal{G}$ ($\gamma = 0.5$)*	0.298	0.669	323.709	0.315	10.313	0.942	0.955	0.975	0.977	0.667	0.198	0.577	0.813
Finetuned*	0.291	0.664	343.947	0.292	9.964	0.932	0.954	0.973	0.974	0.581	0.198	0.519	0.760
Wan-1.3B	0.281	0.703	417.352	0.217	8.625	0.947	0.969	0.992	0.976	0.639	0.271	0.182	0.930
LTX-Video-2B	0.250	0.649	196.736	0.292	10.398	0.907	0.960	0.991	0.982	0.453	0.092	0.176	0.573

4.2 RESULTS

Benchmark results. Table 2 summarizes the quantitative evaluation. Relative to the unguided CogVideoX-2B baseline, NoiseFlow improves several paired, distributional, and temporal metrics across the evaluated guidance scales. The best FVD decreases from 345.6 to 283.3 at $\gamma = 0.5$, corresponding to an improvement of approximately 18%. SSIM and PSNR increase monotonically over the evaluated range, reaching 0.408 and 11.751, respectively, at $\gamma = 0.8$. *Dynamic Degree* initially increases from 0.611 to 0.678 at $\gamma = 0.2$, but decreases at larger guidance scales, indicating a trade-off between stronger warp guidance and the prevalence of dynamic content. *Dynamic Spatial Relationship* (DSR) reaches 0.237 at $\gamma = 0.8$, while Camera Motion reaches its highest NoiseFlow value of 0.593 at $\gamma = 0.5$. *Instance Preservation* decreases as the guidance strength increases. These results indicate that warp guidance provides gains on several temporal metrics, but that the preferred guidance scale depends on the desired trade-off between temporal consistency, dynamics, and instance preservation. Several of these metrics remain better or on a competitive scale when viewed across all configurations, with LTX-Video reporting the best FVD score but notably worse DSR among all configurations considered.

Qualitative analysis. Figure 3a shows an example of the warp field in the latent noise space extracted from a real video and compared against the dense optical flow computed in the pixel space for comparison. It can be observed that the warp directions learned from denoiser feedback appear to be generally consistent with the motion from the pixel space with some residual latent noise motion not observed in the optical flow. It can also be noted that the innovation rate, which we define as the positive local divergence of the warp field, supplies fresh i.i.d. noise to locally divergent regions. Figure 3b shows how the warp changes over time steps. We observe that the warp field varies smoothly as we move toward noise from the data. Near extreme SNR levels, the field changes rapidly from data aligned temporal motion to spatially diffused warp. Collectively, these results demonstrate the existence of a temporally varying warp field that is strongly supported by the gradient of the denoiser with respect to the input inducing a pullback on the clean video latent representation.

Figure 4 shows examples of videos generated by NoiseFlow, showing high quality generation in line with the evaluation results in Table 2. Appendix E extends the qualitative analysis with additional examples and a comparison with the baselines. We also provide a visualization of the warps realized



(a) (b)

Figure 3: (a) Visualization of the learned latent noise warp field on a sample (top row). The next two rows show the learned warp field output and the associated innovation rate, respectively. The bottom row shows a dense optical field in the pixel space extracted using RAFT (Teed & Deng, 2020) for comparison. (b) Comparison of the warp field as it varies across diffusion timesteps. Please see Appendix E for more examples.

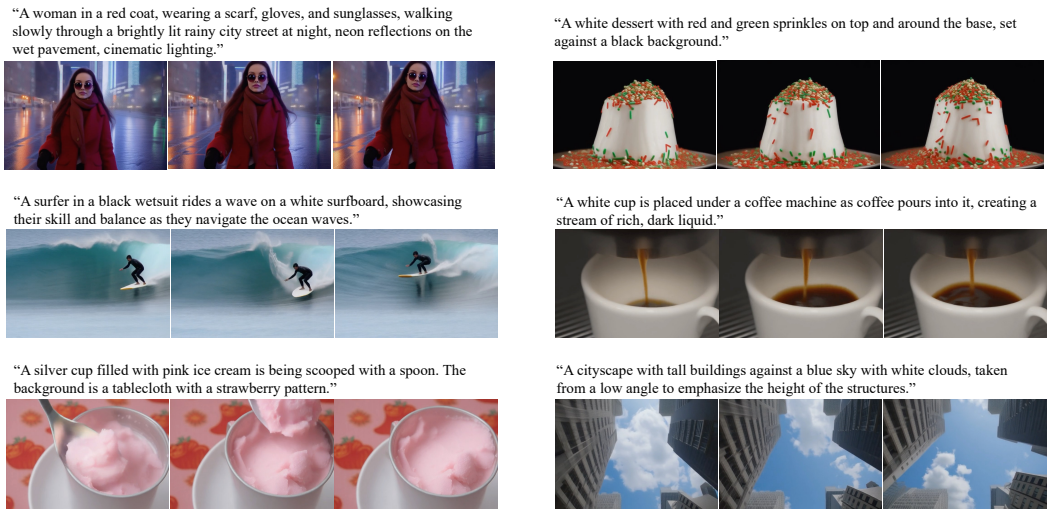
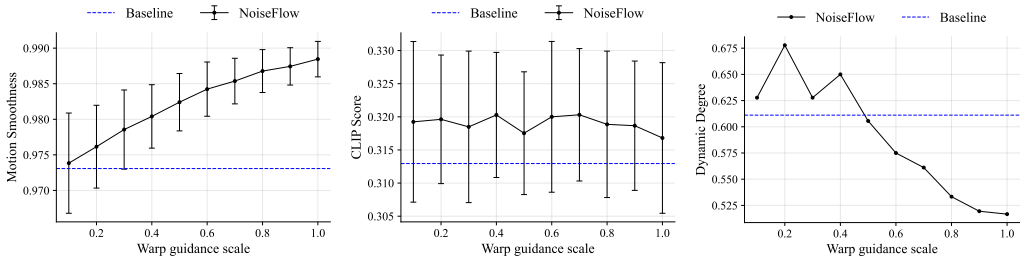


Figure 4: Example of generated videos by NoiseFlow ($\gamma = 0.8$).

during inference. In general, NoiseFlow learns diffusion-time indexed warping transformations as the comparisons in Figures 8 and 9 demonstrate.

Guidance scale ablations. To further examine how warp guidance affects motion quality, we assessed *motion smoothness* and *dynamic degree* across different guidance scales. *Motion smoothness* serves as a direct measure of video motion quality, whereas *dynamic degree* is computed as the average of a binary label indicating the proportion of videos containing dynamic content rather than static scenes. Additionally, we report CLIP scores on the prompt set that span these VBench dimensions. Taken together, these metrics identify whether motion quality and text alignment genuinely improve as a result of warp guidance and not merely an artifact of more static videos being produced by the model. Figure 5 shows that NoiseFlow achieves consistent performance gains on *motion smoothness*, with improvements increasing with guidance strength. Similarly, we observe initial gains in *dynamic degree* followed by a stable decrease after $\gamma = 0.5$ that stabilizes near the end. This indicates a slight bias toward the generation of static content at higher warp-guidance scales, consistent



(a) Motion smoothness (b) CLIP score (c) Dynamic degree

Figure 5: Impact of the warp guidance scale on (a) *motion smoothness*, and (c) *dynamic degree* from the VBench (Huang et al., 2024b) suite. We also evaluate the (b) CLIP scores over this subset. The figures show mean and standard deviation across the 5 seeds used by the VBench standard. *Dynamic degree* is the mean of a binary indicator, demonstrating the fraction of videos with dynamic content.

with the observation in Table 2. The CLIP scores remain stable and above the baseline mean on all measured guidance scales. These results indicate that the application of warp guidance improves motion quality in the generated videos over the baseline with a slight decline in dynamic content at higher warp scales. Figures 11 and 12 visually demonstrate this impact of the warping guidance scale on the video generation quality.

No-warp self-guidance ($\mathcal{F} = \varnothing, \mathcal{G}' = \mathcal{G}$). To understand whether the observed gains can be attributable to the learned priors and not simply due to the use of an additional model invocation beyond the standard CFG, we evaluated the setup where we apply orthogonal self-guidance to the base model using itself as a proxy for $\mathcal{G}'$. We use i.i.d. priors instead of warped priors and fix $\gamma = 0.5$ for this ablation. The results in Table 2 indicate that employing the teacher model to provide orthogonal guidance yields consistent improvements across all evaluated metrics relative to the baseline configuration ($\gamma = 0$). This finding demonstrates the utility of the guidance mechanism when considered independently of the warped prior. However, under an identical guidance scale, this strategy does not outperform NoiseFlow on any metric, except for DD and IP, for which the trade-off associated with NoiseFlow has already been established. Collectively, these results support our hypothesis that an additional invocation of the base model alone is insufficient to explain the performance gains obtained with the warped priors distilled by NoiseFlow.

5 Discussion

Intuitively, $\mathcal{G}$ and $\mathcal{G}'$ can be thought of as defining individual paths from a family of noise distributions to the data distribution, with the warp generator providing a bridge between them, as shown in Figure 1a. Crucially, the proposed approach eliminates the need for a source video at inference time and can effectively internalize the variable warp transformations observed during training (Figures 8 and 9). Consequently, NoiseFlow is not a V2V method or an equivariant diffusion model, as both paradigms typically assume access to a source video at inference. Nevertheless, the method is closely related to, and can be interpreted as a generalization of, these classes of models. Specifically, the warped noise conditioned model $\mathcal{G}'$ defines a family of equivariant models with priors constant in s but changing in t. In fact, $\mathcal{F}$ can be trained as a fixed function of clean data $\boldsymbol{x}_0$ rather than $\hat{\boldsymbol{x}}_{0,t}$, while also keeping it independent of t (e.g., optical flow). Under these simplified conditions, if a source video is available, one may directly sample the student model without the need for alternation. In this specific setup, our method reduces to equivariance training.

In summary, our results suggest that the diffusion prior in video generation need not be treated as a fixed source of independent randomness. Once its marginal statistics are preserved, its joint spatiotemporal structure becomes an additional degree of freedom in modeling. NoiseFlow explores an example of this idea by learning that structure from the denoiser geometry and allowing it to evolve along the diffusion trajectory. This perspective suggests a broader class of generative models in which the prior itself is learned as a structured, trajectory-dependent stochastic representation rather than being specified independently of the generator.

AI USE STATEMENT

In this work, we used generative AI tools to assist in polishing the writing and coding assistance. All AI-assisted outputs were treated as suggestions only and independently reviewed, verified and corrected under strict human supervision. We did not rely on generative AI for scientific content, mathematical development, experimental design, implementation decisions, and conclusions. We take responsibility for the final content of this work, including text, claims, code, and other artifacts.

REFERENCES

Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023.

Yuxiang Bao, Di Qiu, Guoliang Kang, Baochang Zhang, Bo Jin, Kaiye Wang, and Pengfei Yan. Latentwarp: Consistent diffusion latents for zero-shot video-to-video translation. *arXiv preprint arXiv:2311.00353*, 2023.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023a. URL https://doi.org/10.48550/arXiv.2311.15127.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023b.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13–23. IEEE, 2025.

Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23149–23160. IEEE, 2023.

Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius Da Costa De Azevedo. How i warped your noise: a temporally-correlated noise prior for diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 30015–30046, 2024.

Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, et al. Goku: Flow based video generative foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23516–23527, 2025.

Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. In *International Conference on Learning Representations*, volume 2024, pp. 52086–52106, 2024.

Giannis Daras, Weili Nie, Karsten Kreis, Alexandros G Dimakis, Morteza Mardani, Nikola B Kovachki, and Arash Vahdat. Warped diffusion: Solving video inverse problems with image diffusion models. *Advances in Neural Information Processing Systems*, 37:101116–101143, 2024.

Yitong Deng, Winnie Lin, Lingxiao Li, Dmitriy Smirnov, Ryan Burgert, Ning Yu, Vincent Dedun, and Mohammad Taghavi. Infinite-resolution integral noise warping for diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 51390–51408, 2025.

Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22930–22941, 2023.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Fx2SbBgcte.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

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

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

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

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=rB6TpjAuSRy.

Xingchang Huang, Corentin Salaun, Cristina Vasconcelos, Christian Theobalt, Cengiz Oztireli, and Gurprit Singh. Blue noise for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024a.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024b.

Jing Jia, Liyue Shen, and Guanyang Wang. Couple to control: Joint initial noise design in diffusion models. *arXiv preprint arXiv:2605.11311*, 2026.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong MU, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=66NzcRQuOq.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15908–15918, 2023. doi: 10.1109/ICCV51070.2023.01462.

Doyeon Kim, Donggyu Joo, and Junmo Kim. Tivgan: Text to image to video generation with step-by-step evolutionary generator. *IEEE Access*, 8:153113–153122, 2020.

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

Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, et al. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8207–8216. IEEE, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Chao Liu and Arash Vahdat. On equivariance and fast sampling in video diffusion models trained with warped noise. *arXiv preprint arXiv:2504.09789*, 2025.

Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023.

Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025.

Jiafeng Mao, Xueting Wang, and Kiyoharu Aizawa. Guided image synthesis via initial image editing in diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5321–5329, 2023.

Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7038–7048, 2024.

Boming Miao, Chunxiao Li, Xiaoxiao Wang, Andi Zhang, Rui Sun, Zizhe Wang, and Yao Zhu. Noise diffusion for enhancing semantic faithfulness in text-to-image synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23575–23584. IEEE, 2025.

Antonio Montanaro, Luca Savant Aira, Emanuele Aiello, Diego Valsesia, and Enrico Magli. Motioncraft: Physics-based zero-shot video generation. *Advances in Neural Information Processing Systems*, 37:123155–123181, 2024.

Saurabh Pathak, Elahe Arani, Mykola Pechenizkiy, and Bahram Zonooz. Physvid: Physics aware local conditioning for generative video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41847–41858, June 2026.

Can Qin, Congying Xia, Krithika Ramakrishnan, Michael Ryoo, Lifu Tu, Yihao Feng, Manli Shu, Honglu Zhou, Anas Awadalla, Jun Wang, et al. xgen-videosyn-1: High-fidelity text-to-video synthesis with compressed representations. In *European Conference on Computer Vision*, pp. 249–265. Springer, 2024.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. In *International Conference on Learning Representations*, volume 2024, pp. 5260–5274, 2024.

Takahiro Shirakawa and Seiichi Uchida. Noisecollage: A layout-aware text-to-image diffusion model based on noise cropping and merging. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8921–8930. IEEE, 2024.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nJfylDvgzlq.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

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

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vOEXS39nOF.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *CoRR*, abs/2503.20314, March 2025a. URL https://doi.org/10.48550/arXiv.2503.20314.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Jiasong Feng, Zhanjie Zhang, Wanyuan Pang, and Xiaodan Liang. WISA: World simulator assistant for physics-aware text-to-video generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=4jWuS5hye1.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025c.

Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021.

Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European Conference on Computer Vision*, pp. 720–736, 2022.

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. In *European conference on computer vision*, pp. 378–394. Springer, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

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

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

A Related work

In this section, we first briefly situate our work in the broader landscape of generative T2V modeling and discuss the main architectural trends that motivate our approach. We then summarize prior methods most closely related to our contributions, emphasizing how they model spatiotemporal structure and how noise priors are constructed or manipulated. We close each subsection by highlighting the remaining limitations that our method addresses and clarifying the specific points of comparison used in the main paper.

A.1 Generative Text-to-Video Models

Early methods in video generation were largely based on Generative Adversarial Networks (GANs) (Kim et al., 2020; Li et al., 2018) or autoregressive video generation (Hong et al., 2023; Wu et al., 2022; 2021; Villegas et al., 2023). Diffusion-based conditional video generation models have since emerged as a prominent paradigm for video generation operating in continuous pixel or latent spaces. A substantial body of work extends pretrained image generation models to the video domain, leveraging strong image priors. These approaches typically introduce explicit temporal modeling mechanisms into image-based architectures, including temporal convolution or attention layers, structured spatiotemporal latent representations, and cross-frame attention, to promote motion modeling and temporal consistency (Khachatryan et al., 2023; Guo et al., 2024; Singer et al., 2023; Blattmann et al., 2023b;a; An et al., 2023; Wang et al., 2023; Zhou et al., 2022; Wang et al., 2025c; Bar-Tal et al., 2024; Ceylan et al., 2023). Subsequently, a growing line of work has focused on video-native generative architectures that directly model the joint spatiotemporal distribution of videos. Early diffusion-based video models explored joint spatiotemporal denoising in pixel space (Ho et al., 2022b), including large-scale cascaded T2V systems such as Imagen Video (Ho et al., 2022a), with later work shifting to latent space (Qin et al., 2024; Yang et al., 2025; Menapace et al., 2024; HaCohen et al., 2024). Joint spatiotemporal modeling allows spatial appearance and temporal dynamics to be represented within a unified generative process, facilitating the modeling of motion, object persistence, and scene consistency over time.

Beyond diffusion models, recent work has increasingly adopted flow-matching formulations (Lipman et al., 2023) for video generation Jin et al. (2025); Ma et al. (2025); Kong et al. (2024); Wang et al. (2025a); Chen et al. (2025). Rather than learning a diffusion denoising process, these methods learn a time-dependent vector field that transports samples from a simple source distribution toward the video data distribution. This formulation provides an alternative continuous-time generative framework that integrates naturally with modern spatiotemporal video architectures.

Taken together, this line of work has produced increasingly powerful spatiotemporal generators, but in most cases temporal coherence is enforced primarily through architectural choices and training objectives while the generative process is still initialized from temporally independent (or heuristically correlated) noise. This leaves a gap between video-native modeling and principled priors for spatiotemporal correspondence. Existing methods often rely on fixed correlations, external motion cues, or inference-time tricks, rather than learning a temporally structured prior that can represent dynamics across the entire generative trajectory. Our paper addresses this gap by learning a temporally evolving Gaussian noise process that encodes video dynamics directly in the diffusion prior, improving motion consistency while preserving the marginal statistics required by diffusion training.

A.2 Noise warping for video generation

Although diffusion models are conventionally trained using gaussian noise, several works in image generation have shown that the use of spatially structured noise can be advantageous for generation (Mao et al., 2023; Shirakawa & Uchida, 2024; Huang et al., 2024a; Guo et al., 2024; Miao et al., 2025; Jia et al., 2026). Collectively, these results establish that spatial, spectral, and semantic interventions in noise space can systematically influence the generated sample despite leaving the denoising architecture largely unchanged.

This observation becomes especially consequential for video, where the prior must additionally determine temporal correspondence. As early video diffusion primarily arose by extending effective image-diffusion frameworks into the temporal domain, incorporating temporal correlations directly into the diffusion process, instead of forcing a model derived from image generation to infer them

solely from temporally independent noise, provided a straightforward and natural solution. VideoFusion (Luo et al., 2023) and PYoCo (Ge et al., 2023) are examples of this direction, introducing video noise priors with components shared across time alongside frame-specific randomness. Related inference-time methods such as FreeNoise (Qiu et al., 2024) and FreeInit (Wu et al., 2024) exploit the temporal structure of noise without modifying the pretrained video model. FreeNoise reschedules a finite set of noise samples across frames to introduce long-range correlations. More fundamentally, FreeInit identifies an initialization mismatch between training and inference. With only a finite number of forward diffusion steps during training, terminal training latents retain non-negligible low frequency correlations. In contrast, standard inference starts from i.i.d. noise, which lacks these correlations. This suggests that the usual i.i.d. prior at inference may be suboptimal for video generation, and explicitly using temporally correlated priors can yield better performance.

A complementary line of work aims to impose temporal correspondence more directly by defining explicit transformations of the noise itself, typically by advecting noise along motion trajectories. Integral noise-warping techniques treat discrete white noise as the integral of an underlying continuous Gaussian process, which allows the noise to be advected while maintaining its Gaussian statistics at each frame and avoiding the spatial correlations that standard interpolation typically introduces (Chang et al., 2024; Deng et al., 2025). Go-with-the-Flow (Burgert et al., 2025) incorporates warped noise directly into video diffusion training by leveraging optical-flow-based noise trajectories to fine-tune pretrained video diffusion models, enabling motion specified in noise space to dictate motion during generation. The work in Liu & Vahdat (2025) and Daras et al. (2024) provides a theoretical account of this phenomenon, showing that, under corresponding transformations of video frames and input noise, the standard denoising objective admits an equivariant optimal denoiser. Empirically, training video diffusion models with motion warped noise causes motion in the input noise to align with motion in the generated video. Our setting is related to this final perspective, however, it differs substantially with respect to (i) the manner in which temporal structure is incorporated into the noise process, (ii) the procedure by which this structure is inferred or constructed, and (iii) its evolution along the diffusion trajectory.

The key limitation of these warping-based approaches is that they require a source or reference video (or an externally estimated motion field derived from it) at inference to define the noise trajectories, restricting their use in fully unconditional or purely text-conditioned generation. In contrast, we model the diffusion prior as a dynamical random field whose temporal dependence evolves throughout the diffusion trajectory and is learned from the data, without requiring any source video at inference. This turns temporal dependence in the prior into a learned stochastic representation of video dynamics that changes along the denoising path, rather than a hand-crafted correlation scheme or a fixed motion field specified from outside the model.

B Theoretical Details

This section provides the derivations underlying the mathematical objects introduced in Section 3. We first formalize the spatial white-noise field and show that the advection operator used in Eq. equation 5 induces an orthogonal transport that preserves standard Gaussian marginals. We then derive the stochastic frame transition and the structure-preserving Cayley discretization used to apply this transport without explicitly constructing the corresponding matrix. Next, we analyze the denoiser-pullback objective used to supervise the velocity field, including its reduction to a local velocity-space objective and the resulting shape and magnitude losses. Finally, we characterize the teacher-indexed family of warped diffusion paths introduced in Section 3.3. Collectively, these results justify the fact that our construction preserves the required Gaussian marginals while providing a principled mechanism for temporally coherent noise evolution and a well-defined learning signal for the velocity field.

B.1 Spatial white noise

We start by introducing the isonormal Gaussian process in the following definition.

Definition 1 (Isonormal Gaussian process)**.** *Let $L^2(\Omega)$ denote the Hilbert space of square-integrable real-valued functions on $\Omega \subset \mathbb{R}^2$, equipped with*

$$\langle \phi, \psi \rangle_{L^2(\Omega)} = \int_\Omega \phi(\boldsymbol{x})\psi(\boldsymbol{x})\, d\boldsymbol{x}. \tag{23}$$

An isonormal Gaussian process ξ over $L^2(\Omega)$ is a centered, jointly Gaussian family satisfying

$$\mathbb{E}[\xi(\phi)\xi(\psi)] = \langle \phi, \psi \rangle_{L^2(\Omega)} \qquad \forall \phi, \psi \in L^2(\Omega). \tag{24}$$

In particular,

$$\xi(\phi) \sim \mathcal{N}\left(0, \|\phi\|^2_{L^2(\Omega)}\right). \tag{25}$$

The following proposition then recovers the finite-dimensional spatial Gaussian noise used by the generative models as a specific discrete realization of this process.

Proposition 1 (Discrete realization of spatial white noise)**.** *Let $\{\mathbb{A}_i\}_{i=1}^N$ be a partition of Ω into disjoint cells of positive area with variance $|\mathbb{A}_i|$ and define*

$$\phi_i(\boldsymbol{x}) = \frac{\mathbf{1}_{\mathbb{A}_i}(\boldsymbol{x})}{\sqrt{|\mathbb{A}_i|}}. \tag{26}$$

Then the collection of random variables

$$\epsilon_i = \xi(\phi_i), \qquad i = 1, \dots, N, \tag{27}$$

satisfies

$$(\epsilon_1, \dots, \epsilon_N) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N). \tag{28}$$

Proof. Since the cells are disjoint,

$$\langle \phi_i, \phi_j \rangle_{L^2(\Omega)} = \delta_{ij}. \tag{29}$$

By Definition 1,

$$\mathbb{E}[\boldsymbol{\epsilon}_i] = 0, \qquad \mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}. \tag{30}$$

Because the collection is jointly Gaussian, the resulting vector is $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_N)$. □

Proposition 1 identifies the spatially i.i.d. Gaussian latent used by the diffusion model with a finite-dimensional representation of continuous spatial white noise. This allows the transport properties below to be stated first in continuous space and then transferred to the discrete latent grid.

B.2 Spatial noise transport

We can now present the spatial noise transport. We begin by constructing the continuous advection dynamics using a skew-symmetric spatial operator. We then discretize the spatial domain to derive the corresponding discrete advection operator.

Remark 1 (Channelwise spatial transport)**.** *In the following discussion, we assume that the same spatial transport operator is applied independently to each latent channel. Consequently, all spatial operators below act on the $H \times W$ domain and are shared across channels. This is done purely for notational simplicity. However, one may similarly define independent per-channel transports without violating any of the ideas presented here.*

B.2.1 Continuous spatial transport

Definition 2 (Continuous transport dynamics)**.** *Let $\boldsymbol{v} : \Omega \to \mathbb{R}^2$. The transported field ϕ_τ evolves according to*

$$\partial_\tau \phi_\tau = \mathcal{L}_{\boldsymbol{v}} \phi_\tau, \qquad \phi_0 = \phi, \tag{31}$$

where $\mathcal{L}_{\boldsymbol{v}}$ is the advection operator.

$$\mathcal{L}_{\boldsymbol{v}} \phi = -\frac{1}{2} \left(\boldsymbol{v} \cdot \nabla \phi + \nabla \cdot (\boldsymbol{v} \phi) \right). \tag{32}$$

We now show that the advection operator $\mathcal{L}_{\boldsymbol{v}}$ defined above is skew-adjoint under the following assumption.

Assumption 1 (Periodic boundary)**.** *The spatial domain $\Omega \subset \mathbb{R}^2$ is periodic on $\partial\Omega$, such that boundary terms vanish under integration by parts.*

Lemma 1 (Skew-adjointness of the advection generator)**.** *Under Assumption 1,*

$$\mathcal{L}_{\boldsymbol{v}}^{*} = -\mathcal{L}_{\boldsymbol{v}}. \tag{33}$$

Proof. Let the position variable be $\boldsymbol{p} \in \Omega$. For $\phi, \psi \in L^2(\Omega)$,

$$\langle \psi, \mathcal{L}_{\boldsymbol{v}} \phi \rangle = -\frac{1}{2} \int_{\Omega} \psi \left[\boldsymbol{v} \cdot \nabla \phi + \nabla \cdot (\boldsymbol{v} \phi) \right] d\boldsymbol{p}. \tag{34}$$

Integration by parts, together with Assumption 1, gives

$$\langle \psi, \mathcal{L}_{\boldsymbol{v}} \phi \rangle = \frac{1}{2} \int_{\Omega} \phi \left[\boldsymbol{v} \cdot \nabla \psi + \nabla \cdot (\boldsymbol{v} \psi) \right] d\boldsymbol{p} \tag{35}$$

$$= -\langle \mathcal{L}_{\boldsymbol{v}} \psi, \phi \rangle. \tag{36}$$

Hence $\mathcal{L}_{\boldsymbol{v}}^{*} = -\mathcal{L}_{\boldsymbol{v}}$. □

Next, we describe the continuous transport induced by the dynamics in Definition 2 under the simplifying assumption below.

Assumption 2 (Framewise constant dynamics)**.** *For the transition from frame f to frame $f+1$, the velocity $\boldsymbol{v}_f$ and innovation rate $\boldsymbol{B}_f$ are held constant over the corresponding latent-time interval $[\tau_f, \tau_{f+1}]$ of length $\Delta\tau = \tau_{f+1} - \tau_f$.*

Proposition 2 (Continuous transport induced by the advection dynamics)**.** *Under Assumption 2, the solution of equation 32 over the interval $\Delta\tau$ is*

$$\phi_{\Delta\tau} = U_{\boldsymbol{v}_f}(\Delta\tau)\phi_f, \qquad U_{\boldsymbol{v}_f}(\Delta\tau) = \exp(\Delta\tau \mathcal{L}_{\boldsymbol{v}_f}). \tag{37}$$

Under Assumption 1, $U_{\boldsymbol{v}_f}(\Delta\tau)$ is unitary:

$$U_{\boldsymbol{v}_f}(\Delta\tau)^{*} U_{\boldsymbol{v}_f}(\Delta\tau) = I. \tag{38}$$

Proof. Equation equation 31 is a linear autonomous evolution equation, so its solution operator over an interval $\Delta\tau$ is

$$U_{\boldsymbol{v}_f}(\Delta\tau) = \exp(\Delta\tau \mathcal{L}_{\boldsymbol{v}_f}). \tag{39}$$

By Lemma 1

$$U_{\boldsymbol{v}_f}(\Delta\tau)^{*} U_{\boldsymbol{v}_f}(\Delta\tau) = \exp(\Delta\tau \mathcal{L}_{\boldsymbol{v}_f}^{*}) \exp(\Delta\tau \mathcal{L}_{\boldsymbol{v}_f}) \tag{40}$$

$$= \exp(-\Delta\tau \mathcal{L}_{\boldsymbol{v}_f}) \exp(\Delta\tau \mathcal{L}_{\boldsymbol{v}_f}) \tag{41}$$

$$= I. \tag{42}$$

□

B.2.2 Discrete spatial transport

Next, we define advection in discrete space by discretizing the continuous spatial advection operator $\mathcal{L}$. To do so, we first define the spatial derivative operators $\boldsymbol{D}_x$ and $\boldsymbol{D}_y$ and prove an associated Lemma.

Definition 3 (Discrete spatial derivative operators)**.** *Let the spatial domain be discretized on an $H \times W$ grid with spacings h_x and h_y. After vectorizing a discrete field $\phi \in \mathbb{R}^{H \times W}$ as $\phi \in \mathbb{R}^{HW}$, we define the centered finite-difference operators $\boldsymbol{D}_x, \boldsymbol{D}_y \in \mathbb{R}^{HW \times HW}$ by*

$$(\boldsymbol{D}_x \phi)_{i,j} = \frac{\phi_{i,j+1} - \phi_{i,j-1}}{2h_x}, \tag{43}$$

$$(\boldsymbol{D}_y \phi)_{i,j} = \frac{\phi_{i+1,j} - \phi_{i-1,j}}{2h_y}, \tag{44}$$

with positional indices i, j.

Lemma 2 (Skew-symmetry of the discrete derivatives)**.** *Under Assumption 1, and Definition 3*

$$\boldsymbol{D}_x^{\top} = -\boldsymbol{D}_x, \qquad \boldsymbol{D}_y^{\top} = -\boldsymbol{D}_y. \tag{45}$$

Proof. Consider $\boldsymbol{D}_x$; the argument for $\boldsymbol{D}_y$ is identical. For discrete fields $\phi, \psi \in \mathbb{R}^{H\times W}$,

$$\langle \psi, \boldsymbol{D}_x \phi \rangle = \sum_{i,j} \psi_{i,j} \frac{\phi_{i,j+1} - \phi_{i,j-1}}{2h_x}. \tag{46}$$

Reindexing the two sums and using the boundary convention of Assumption 1,

$$\langle \psi, \boldsymbol{D}_x \phi \rangle = \frac{1}{2h_x} \sum_{i,j} \left(\psi_{i,j-1} - \psi_{i,j+1}\right) \phi_{i,j} \tag{47}$$

$$= -\sum_{i,j} \frac{\psi_{i,j+1} - \psi_{i,j-1}}{2h_x} \phi_{i,j} \tag{48}$$

$$= -\langle \boldsymbol{D}_x \psi, \phi \rangle. \tag{49}$$

Hence $\boldsymbol{D}_x^\top = -\boldsymbol{D}_x$. Similarly, $\boldsymbol{D}_y^\top = -\boldsymbol{D}_y$. □

Now we are prepared to define the discrete advection operator $\boldsymbol{L}_{\boldsymbol{v}}$.

Definition 4 (Discrete advection generator)**.** *Let*

$$\boldsymbol{V}_x = \mathrm{diag}(v_x), \qquad \boldsymbol{V}_y = \mathrm{diag}(v_y) \tag{50}$$

denote the diagonal matrices that contain the two components v_x, v_y of the discrete velocity field $\boldsymbol{v}$. We discretize the continuous generator $\mathcal{L}_{\boldsymbol{v}}$ as

$$\boldsymbol{L}_{\boldsymbol{v}} = -\frac{1}{2} \left(\boldsymbol{V}_x \boldsymbol{D}_x + \boldsymbol{D}_x \boldsymbol{V}_x + \boldsymbol{V}_y \boldsymbol{D}_y + \boldsymbol{D}_y \boldsymbol{V}_y\right). \tag{51}$$

Proposition 3 (Skew-symmetry of the discrete generator)**.** $\boldsymbol{L}_{\boldsymbol{v}}$ *is skew-symmetric:*

$$\boldsymbol{L}_{\boldsymbol{v}}^\top = -\boldsymbol{L}_{\boldsymbol{v}}. \tag{52}$$

Proof. Since $\boldsymbol{V}_x^\top = \boldsymbol{V}_x$ and $\boldsymbol{V}_y^\top = \boldsymbol{V}_y$, while $\boldsymbol{D}_x^\top = -\boldsymbol{D}_x$ and $\boldsymbol{D}_y^\top = -\boldsymbol{D}_y$,

$$\boldsymbol{L}_{\boldsymbol{v}}^\top = -\frac{1}{2} \left(\boldsymbol{D}_x^\top \boldsymbol{V}_x + \boldsymbol{V}_x \boldsymbol{D}_x^\top + \boldsymbol{D}_y^\top \boldsymbol{V}_y + \boldsymbol{V}_y \boldsymbol{D}_y^\top\right) \tag{53}$$

$$= \frac{1}{2} \left(\boldsymbol{D}_x \boldsymbol{V}_x + \boldsymbol{V}_x \boldsymbol{D}_x + \boldsymbol{D}_y \boldsymbol{V}_y + \boldsymbol{V}_y \boldsymbol{D}_y\right) \tag{54}$$

$$= -\boldsymbol{L}_{\boldsymbol{v}}. \tag{55}$$

□

Similarly to the continuous domain, $\boldsymbol{L}_{\boldsymbol{v}}$ can be used to define an advection in a discrete space.

Proposition 4 (Discrete transport induced by the latent advection dynamics)**.** *After spatial discretization, the deterministic latent transport dynamics are*

$$d\boldsymbol{\epsilon}_\tau = \boldsymbol{L}_{\boldsymbol{v}} \boldsymbol{\epsilon}_\tau d\tau, \qquad \boldsymbol{\epsilon}_0 = \boldsymbol{\epsilon}. \tag{56}$$

Under Assumption 2, the velocity is fixed over the frame interval $[\tau_f, \tau_{f+1}]$. Therefore, the exact transition from τ_f to τ_{f+1} is

$$\boldsymbol{\epsilon}_{f+1} = \boldsymbol{Q}_{\boldsymbol{v}}(\Delta\tau) \boldsymbol{\epsilon}_f, \qquad \boldsymbol{Q}_{\boldsymbol{v}}(\Delta\tau) = \exp(\Delta\tau \boldsymbol{L}_{\boldsymbol{v}}). \tag{57}$$

This transition is orthogonal:

$$\boldsymbol{Q}_{\boldsymbol{v}}(\Delta\tau)^\top \boldsymbol{Q}_{\boldsymbol{v}}(\Delta\tau) = \boldsymbol{I}. \tag{58}$$

Proof. By Proposition 3, $\boldsymbol{L}_{\boldsymbol{v}}^\top = -\boldsymbol{L}_{\boldsymbol{v}}$. Therefore,

$$\boldsymbol{Q}_{\boldsymbol{v}}^\top \boldsymbol{Q}_{\boldsymbol{v}} = \exp(\Delta\tau \boldsymbol{L}_{\boldsymbol{v}}^\top) \exp(\Delta\tau \boldsymbol{L}_{\boldsymbol{v}}) \tag{59}$$

$$= \exp(-\Delta\tau \boldsymbol{L}_{\boldsymbol{v}}) \exp(\Delta\tau \boldsymbol{L}_{\boldsymbol{v}}) \tag{60}$$

$$= \boldsymbol{I}. \tag{61}$$

□

Corollary 1 (Preservation of Gaussian marginals)**.** *If $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, then*

$$\boldsymbol{Q}_{\boldsymbol{v}} \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{62}$$

Having established a Gaussian measure-preserving deterministic transport, we subsequently extend this construction to derive the SDE introduced in Section 3.

B.3 STOCHASTIC INNOVATION AND THE FRAME TRANSITION

Pure transport gives

$$\boldsymbol{\epsilon}_{f+1} = \boldsymbol{Q}_f \boldsymbol{\epsilon}_f. \tag{63}$$

Once the initial noise frame is sampled, this construction introduces no new stochastic degrees of freedom along the temporal dimension. Therefore, we augment orthogonal transport with a local OU innovation process. To do so, we observe the finite transport of the OU process.

Proposition 5 (Gaussian preservation of the OU innovation)**.** *Consider the local OU dynamics*

$$d\boldsymbol{\epsilon}_\tau = -\frac{1}{2}\boldsymbol{B}_f \boldsymbol{\epsilon}_\tau \, d\tau + \sqrt{\boldsymbol{B}_f}\, dW_\tau. \tag{64}$$

With $\boldsymbol{B}_f$ fixed over an interval of length $\Delta\tau$, its finite transition is

$$\boldsymbol{\epsilon}' = \boldsymbol{\rho}_f \boldsymbol{\epsilon} + \sqrt{\boldsymbol{I} - \boldsymbol{\rho}_f^2} \cdot \boldsymbol{\eta}, \qquad \boldsymbol{\rho}_f = \exp\left(-\frac{1}{2}\Delta\tau \boldsymbol{B}_f\right), \tag{65}$$

where $\eta \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is independent of $\boldsymbol{\epsilon}$. If $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, then

$$\boldsymbol{\epsilon}' \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{66}$$

Proof. The mean remains zero. Since $\boldsymbol{B}_f$ and hence $\boldsymbol{\rho}_f$ are diagonal,

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{\epsilon}') &= \boldsymbol{\rho}_f \mathrm{Cov}(\boldsymbol{\epsilon}) \boldsymbol{\rho}_f^\top + (\boldsymbol{I} - \boldsymbol{\rho}_f^2)\mathrm{Cov}(\boldsymbol{\eta}) && (67)\\ &= \boldsymbol{\rho}_f^2 + (\boldsymbol{I} - \boldsymbol{\rho}_f^2) && (68)\\ &= \boldsymbol{I}. && (69) \end{aligned}$$

The update is linear in independent Gaussian variables and is therefore Gaussian with covariance $\boldsymbol{I}$. □

The SDE used in the paper can be derived by combining the deterministic transport in Proposition 4 and the OU innovation process in Proposition 5. We have

$$\boxed{d\boldsymbol{\epsilon}_\tau = \left(\boldsymbol{L}_{\boldsymbol{v}_\tau} - \frac{1}{2}\boldsymbol{B}_\tau\right)\boldsymbol{\epsilon}_\tau \, d\tau + \sqrt{\boldsymbol{B}_\tau}\, dW_\tau.} \tag{70}$$

Remark 2 (Choice of innovation rate)**.** *We choose*

$$\boldsymbol{B}_\tau(\boldsymbol{p}) = [\nabla \cdot \boldsymbol{v}_\tau(\boldsymbol{p})]^+. \tag{71}$$

This choice determines where *fresh stochastic innovation enters. Regions with positive local divergence receive new noise. Gaussian marginal preservation does not require this functional form. It follows from the matched damping and diffusion terms in equation 70. The positive-divergence rule is therefore a modeling choice rather than a consequence of Gaussian-measure preservation.*

Next, we show that integrating equation 70, yields a stochastic transport.

Proposition 6 (Gaussian preservation of the split NoiseFlow transition)**.** *Under Assumption 2, consider the Strang-split update in equation 9. If $\boldsymbol{\epsilon}^k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, then*

$$\boldsymbol{\epsilon}^{k+1} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}).$$

Consequently, after any finite number K_f of substeps, the framewise marginal remains standard Gaussian.

Proof. The first and final OU half-steps preserve $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ by Proposition 5. Since $\boldsymbol{L}_{\boldsymbol{v}_f}$ is skew-symmetric, its Cayley transform $C_h(L_{v_f})$ is orthogonal and therefore also preserves $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The composition of these marginal-preserving maps thus preserves the standard Gaussian distribution. □

Corollary 2 (Temporal covariance induced by the warp)**.** *Conditioned on the predicted operators, let*

$$\boldsymbol{T}_{f,h} = \boldsymbol{\rho}_{f,h} C_h(\boldsymbol{L}_{\boldsymbol{v}_f}) \boldsymbol{\rho}_{f,h}. \tag{72}$$

For constant dynamics over the frame interval, the deterministic component of the K_f-step transition is

$$\boldsymbol{T}_f^{(K_f)} = \boldsymbol{T}_{f,h}^{K_f}. \tag{73}$$

Hence

$$\mathrm{Cov}(\boldsymbol{\epsilon}_{f+1}, \boldsymbol{\epsilon}_f) = \boldsymbol{T}_f^{(K_f)}, \tag{74}$$

while $\mathrm{Cov}(\boldsymbol{\epsilon}_f) = \boldsymbol{I}$ for every frame.

B.4 Structure-preserving discretization

Proposition 7 (Convergence to the exponential transport)**.** *Consider equation 56 and let $h = \Delta\tau / K_f$. Then for $k, k+1 \in K_f$, the implicit-midpoint rule satisfies*

$$\boldsymbol{\epsilon}^{k+1} = C_h(\boldsymbol{L}_{\boldsymbol{v}_f}) \boldsymbol{\epsilon}^k, \tag{75}$$

where

$$C_h(\boldsymbol{L}) = \left(\boldsymbol{I} - \frac{h}{2}\boldsymbol{L}\right)^{-1} \left(\boldsymbol{I} + \frac{h}{2}\boldsymbol{L}\right) \tag{76}$$

is the Cayley transform. Then, $\boldsymbol{Q}_f$ can be approximated as

$$\boldsymbol{Q}_f \approx \left[C_h(\boldsymbol{L}_{\boldsymbol{v}_f})\right]^{K_f}, \tag{77}$$

and

$$\lim_{K_f \to \infty} C_h(\boldsymbol{L}_{\boldsymbol{v}_f})^{K_f} = \exp(\Delta\tau \boldsymbol{L}_{\boldsymbol{v}_f}). \tag{78}$$

Proof. Let $\boldsymbol{L} = \boldsymbol{L}_{\boldsymbol{v}_f}, K = K_f$ and $\boldsymbol{X} = \frac{h}{2}\boldsymbol{L}$. An implicit midpoint step satisfies

$$\boldsymbol{\epsilon}^{k+1} - \boldsymbol{\epsilon}^k = \boldsymbol{X}\left(\boldsymbol{\epsilon}^{k+1} + \boldsymbol{\epsilon}^k\right). \tag{79}$$

Rearranging, we have

$$(\boldsymbol{I} - \boldsymbol{X})\,\boldsymbol{\epsilon}^{k+1} = (\boldsymbol{I} + \boldsymbol{X})\,\boldsymbol{\epsilon}^k. \tag{80}$$

Since $\boldsymbol{I} - \boldsymbol{X}$ is invertible due to Proposition 3, we get equation 75.

For sufficiently large K, $\|X\| < 1$, therefore, the matrix logarithms below are well defined by their convergent power series. Since $\boldsymbol{I} + \boldsymbol{X}$ and $\boldsymbol{I} - \boldsymbol{X}$ are both polynomials in $\boldsymbol{L}$, they commute, and hence

$$\log C_h(\boldsymbol{L}) = \log\left[(\boldsymbol{I} - \boldsymbol{X})^{-1}(\boldsymbol{I} + \boldsymbol{X})\right] \tag{81}$$

$$= \log(\boldsymbol{I} + \boldsymbol{X}) - \log(\boldsymbol{I} - \boldsymbol{X}). \tag{82}$$

Using

$$\log(\boldsymbol{I} + \boldsymbol{X}) = \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j} \boldsymbol{X}^j, \qquad \log(\boldsymbol{I} - \boldsymbol{X}) = -\sum_{j=1}^{\infty} \frac{1}{j} \boldsymbol{X}^j, \tag{83}$$

all even-order terms cancel, giving

$$\log C_h(\boldsymbol{L}) = 2\left(\boldsymbol{X} + \frac{\boldsymbol{X}^3}{3} + \frac{\boldsymbol{X}^5}{5} + \cdots\right) \tag{84}$$

$$= h\boldsymbol{L} + \frac{h^3}{12}\boldsymbol{L}^3 + \mathcal{O}(h^5). \tag{85}$$

Therefore,

$$K \log C_h(\boldsymbol{L}) = \Delta\tau \boldsymbol{L} + \frac{\Delta\tau^3}{12K^2}\boldsymbol{L}^3 + \mathcal{O}(K^{-4}). \tag{86}$$

Since

$$C_h(\boldsymbol{L})^K = \exp\left(K \log C_h(\boldsymbol{L})\right), \tag{87}$$

Eq. equation 86 implies

$$C_h(\boldsymbol{L})^K = \exp\left(\Delta\tau\, \boldsymbol{L} + \mathcal{O}(K^{-2})\right). \tag{88}$$

Taking $K \to \infty$ and using continuity of the matrix exponential gives

$$\lim_{K\to\infty} C_h(\boldsymbol{L})^K = \exp(\Delta\tau\, \boldsymbol{L}). \tag{89}$$

□

Remark 3. *In particular, the expansion shows that the accumulated approximation error over a fixed interval decays quadratically in the number of substeps, $\mathcal{O}(K^{-2})$.*

B.5 Denoiser-induced transport objective

In this subsection, we discuss the noisy state inter-frame covariance maximization objective which is derived from the sensitivity of the denoiser to clean latent directions measured in input latent space.

Definition 5 (Denoiser pullback)**.** *Let*

$$G(\cdot, t) : \mathbb{R}^{FCHW} \to \mathbb{R}^{FCHW} \tag{90}$$

denote the frozen denoiser, and let $\boldsymbol{J}_G(\boldsymbol{x}_t, t)$ be its Jacobian with respect to $\boldsymbol{x}_t$. For a clean latent video x_0, define

$$\boldsymbol{g} = \boldsymbol{J}_G(\boldsymbol{x}_t, t)^\top \boldsymbol{x}_0. \tag{91}$$

Let $\boldsymbol{E}_f : \mathbb{R}^{CHW} \to \mathbb{R}^{FCHW}$ inject a perturbation into frame f. The framewise component of the pullback is then

$$\boldsymbol{g}_f = \boldsymbol{E}_f^\top \boldsymbol{g} = \boldsymbol{E}_f^\top \boldsymbol{J}_G(\boldsymbol{x}_t, t)^\top \boldsymbol{x}_0. \tag{92}$$

The vector $\boldsymbol{g}$ maps the output-space direction specified by $\boldsymbol{x}_0$ through the denoiser Jacobian and therefore identifies the corresponding first-order sensitivity direction in the noisy latent representation. To formally interpret the pullback, we consider a perturbation $\boldsymbol{\epsilon}$ of the noisy latent.

Proposition 8 (Denoiser-sensitive covariance under transport)**.** *Let*

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{93}$$

and couple two adjacent frame perturbations through

$$\boldsymbol{\epsilon}_f = \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon}_{f+1} = \boldsymbol{Q}_f \boldsymbol{\epsilon}. \tag{94}$$

For each frame, define the projected first-order denoiser response

$$z_f = \langle \boldsymbol{g}_f, \boldsymbol{\epsilon}_f \rangle, \qquad z_{f+1} = \langle \boldsymbol{g}_{f+1}, \boldsymbol{\epsilon}_{f+1} \rangle. \tag{95}$$

Then

$$\mathrm{Cov}(z_{f+1}, z_f) = \langle \boldsymbol{g}_{f+1}, \boldsymbol{Q}_f \boldsymbol{g}_f \rangle. \tag{96}$$

Proof. To interpret $\boldsymbol{g}_f$, consider a perturbation $\eta_f \in \mathbb{R}^{CHW}$ applied only to frame f, and define

$$\phi_f(\delta) = \langle \boldsymbol{x}_0,\, \mathcal{G}(\boldsymbol{x}_t + \delta \boldsymbol{E}_f \boldsymbol{\eta}_f, t) \rangle\,. \tag{97}$$

Differentiating at $\delta = 0$ gives

$$\phi_f'(0) = \langle \boldsymbol{x}_0,\, \boldsymbol{J}_{\mathcal{G}}(\boldsymbol{x}_t, t) \boldsymbol{E}_f \boldsymbol{\eta}_f \rangle \tag{98}$$

$$= \left\langle \boldsymbol{E}_f^\top \boldsymbol{J}_{\mathcal{G}}(\boldsymbol{x}_t, t)^\top \boldsymbol{x}_0,\, \boldsymbol{\eta}_f \right\rangle \tag{99}$$

$$= \langle \boldsymbol{g}_f, \boldsymbol{\eta}_f \rangle. \tag{100}$$

Thus, $\boldsymbol{g}_f$ measures the first-order sensitivity of the full denoiser output, projected along x_0, to a perturbation of input frame f. Taking $\boldsymbol{\eta}_f = \boldsymbol{\epsilon}_f$ in equation 100 gives $z_f = \phi_f'(0)$, and the same argument applies to frame $f + 1$. Thus, z_f and z_{f+1} are precisely the projected first-order denoiser responses induced by the coupled input perturbations.

Since $\boldsymbol{\epsilon}$ is zero-mean, both responses are centered:

$$\mathbb{E}[z_f] = \mathbb{E}[z_{f+1}] = 0. \tag{101}$$

Using $\boldsymbol{\epsilon}_f = \boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}_{f+1} = \boldsymbol{Q}_f\boldsymbol{\epsilon}$, we therefore obtain

$$\begin{aligned}\mathrm{Cov}(z_{f+1}, z_f) &= \mathbb{E}\left[\langle \boldsymbol{g}_{f+1}, \boldsymbol{Q}_f\boldsymbol{\epsilon}\rangle \langle \boldsymbol{g}_f, \boldsymbol{\epsilon}\rangle\right] &(102)\\ &= \boldsymbol{g}_{f+1}^\top \boldsymbol{Q}_f \mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right] \boldsymbol{g}_f &(103)\\ &= \boldsymbol{g}_{f+1}^\top \boldsymbol{Q}_f \boldsymbol{g}_f &(104)\\ &= \langle \boldsymbol{g}_{f+1}, \boldsymbol{Q}_f\boldsymbol{g}_f\rangle, &(105)\end{aligned}$$

where the third equality uses $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \boldsymbol{I}$. □

Corollary 3 (Finite transport discrepancy)**.** *For fixed $\boldsymbol{g}_f$ and $\boldsymbol{g}_{f+1}$,*

$$\arg\max_{\boldsymbol{Q}_f} \langle \boldsymbol{g}_{f+1}, \boldsymbol{Q}_f\boldsymbol{g}_f\rangle = \arg\min_{\boldsymbol{Q}_f} \|\boldsymbol{g}_{f+1} - \boldsymbol{Q}_f\boldsymbol{g}_f\|_2^2 \tag{106}$$

over orthogonal $\boldsymbol{Q}_f$.

Proof. Using $\boldsymbol{Q}_f^\top \boldsymbol{Q}_f = \boldsymbol{I}$,

$$\begin{aligned}\|\boldsymbol{g}_{f+1} - \boldsymbol{Q}_f\boldsymbol{g}_f\|_2^2 &= \|\boldsymbol{g}_{f+1}\|_2^2 + \|\boldsymbol{Q}_f\boldsymbol{g}_f\|_2^2 - 2\langle \boldsymbol{g}_{f+1}, \boldsymbol{Q}_f\boldsymbol{g}_f\rangle &(107)\\ &= \|\boldsymbol{g}_{f+1}\|_2^2 + \|\boldsymbol{g}_f\|_2^2 - 2\langle \boldsymbol{g}_{f+1}, \boldsymbol{Q}_f\boldsymbol{g}_f\rangle. &(108)\end{aligned}$$

The first two terms are independent of $\boldsymbol{Q}_f$. □

The finite criterion in Corollary 3 constrains the action of the transport $\boldsymbol{Q}_f$ on $\boldsymbol{g}_f$ but does not uniquely specify the velocity field from which $\boldsymbol{Q}_f$ is generated. NoiseFlow therefore learns instantaneous velocity, and obtains finite frame-to-frame transport through the structure-preserving integration in Section B.4. The following Proposition formalizes it.

Proposition 9 (Infinitesimal denoiser-sensitive transport)**.** *Let*

$$\boldsymbol{Q}_f(\Delta\tau) = \exp(\Delta\tau \boldsymbol{L}_{\boldsymbol{v}_f}). \tag{109}$$

Then

$$\frac{d}{d(\Delta\tau)}\mathrm{Cov}(z_{f+1}, z_f)\Big|_{\Delta\tau=0} = \langle \boldsymbol{g}_{f+1}, \boldsymbol{L}_{\boldsymbol{v}_f}\boldsymbol{g}_f\rangle. \tag{110}$$

Proof. Using the first-order expansion

$$\boldsymbol{Q}_f(\Delta\tau) = \boldsymbol{I} + \Delta\tau \boldsymbol{L}_{\boldsymbol{v}_f} + \mathcal{O}(\Delta\tau^2), \tag{111}$$

Proposition 8 gives

$$\mathrm{Cov}(z_{f+1}, z_f) = \langle \boldsymbol{g}_{f+1}, \boldsymbol{g}_f\rangle + \Delta\tau\langle \boldsymbol{g}_{f+1}, \boldsymbol{L}_{\boldsymbol{v}_f}\boldsymbol{g}_f\rangle + \mathcal{O}(\Delta\tau^2). \tag{112}$$

Taking the derivative at $\Delta\tau = 0$ yields the result. □

Next, we derive the velocity-space pullback signal using the infinitesimal transport defined in Proposition 9.

Proposition 10 (Velocity-space form of the transport differential)**.** *For the discrete generator $\boldsymbol{L}_{\boldsymbol{v}_f}$ (equation 51),*

$$\langle \boldsymbol{g}_{f+1}, \boldsymbol{L}_{\boldsymbol{v}_f}\boldsymbol{g}_f\rangle = \langle \boldsymbol{a}_f, \boldsymbol{v}_f\rangle, \tag{113}$$

where $\boldsymbol{a}_f = (\boldsymbol{a}_{x,f}, \boldsymbol{a}_{y,f}) \in \mathbb{R}^{2\times H\times W}$ with

$$\boldsymbol{a}_{x,f} = -\frac{1}{2}\sum_c \left(\boldsymbol{g}_{f+1,c} \odot \boldsymbol{D}_x\boldsymbol{g}_{f,c} - \boldsymbol{g}_{f,c} \odot \boldsymbol{D}_x\boldsymbol{g}_{f+1,c}\right), \tag{114}$$

$$\boldsymbol{a}_{y,f} = -\frac{1}{2}\sum_c \left(\boldsymbol{g}_{f+1,c} \odot \boldsymbol{D}_y\boldsymbol{g}_{f,c} - \boldsymbol{g}_{f,c} \odot \boldsymbol{D}_y\boldsymbol{g}_{f+1,c}\right). \tag{115}$$

Here, $\odot$ denotes pointwise multiplication, and c are the latent channel dimensions.

Proof. Consider first the horizontal component:

$$\boldsymbol{L}_{v_x} = -\frac{1}{2}\left(\boldsymbol{V}_x\boldsymbol{D}_x + \boldsymbol{D}_x\boldsymbol{V}_x\right). \tag{116}$$

Summing over latent channels,

$$\langle \boldsymbol{g}_{f+1}, \boldsymbol{L}_{v_x}\boldsymbol{g}_f\rangle = -\frac{1}{2}\sum_c \left[\boldsymbol{g}_{f+1,c}^\top \boldsymbol{V}_x\boldsymbol{D}_x\boldsymbol{g}_{f,c} + \boldsymbol{g}_{f+1,c}^\top \boldsymbol{D}_x\boldsymbol{V}_x\boldsymbol{g}_{f,c}\right]. \tag{117}$$

Since

$$\boldsymbol{g}_{f+1,c}^\top \boldsymbol{D}_x\boldsymbol{V}_x\boldsymbol{g}_{f,c} = -(\boldsymbol{D}_x\boldsymbol{g}_{f+1,c})^\top \boldsymbol{V}_x\boldsymbol{g}_{f,c}. \tag{118}$$

Hence

$$\langle \boldsymbol{g}_{f+1}, \boldsymbol{L}_{\boldsymbol{v}_x}\boldsymbol{g}_f\rangle = \left\langle -\frac{1}{2}\sum_c \left(\boldsymbol{g}_{f+1,c}\odot \boldsymbol{D}_x\boldsymbol{g}_{f,c} - \boldsymbol{g}_{f,c}\odot \boldsymbol{D}_x\boldsymbol{g}_{f+1,c}\right), v_x\right\rangle. \tag{119}$$

An analogous argument applied to the vertical component yields equation 115. Combining the two component-wise relations then establishes equation 113. □

Corollary 4 (First-order velocity ascent direction)**.** *The field $\boldsymbol{a}_f$ is the Euclidean gradient of the infinitesimal denoiser-sensitive covariance with respect to the velocity field:*

$$\nabla_{\boldsymbol{v}_f}\left\langle \boldsymbol{g}_{f+1}, \boldsymbol{L}_{\boldsymbol{v}_f}\boldsymbol{g}_f\right\rangle = \boldsymbol{a}_f. \tag{120}$$

Thus, $\boldsymbol{a}_f$ specifies the local spatial transport direction that most rapidly increases the covariance of perturbations measured along adjacent-frame denoiser pullbacks.

B.6 Shape and magnitude of the velocity field

In this subsection, we analyze the shape and magnitude losses introduced in Section 3.2. First, we identify the motivation behind the second loss for frame-wise global magnitude in Proposition 11. Then, we show the existence of an optimal shape under the shape loss objective of 14.

Proposition 11 (Scale degeneracy)**.** *The objective*

$$\langle \boldsymbol{a}_f, \boldsymbol{v}_f\rangle \tag{121}$$

does not determine the scale of $\boldsymbol{v}_f$. In particular, if $\langle \boldsymbol{a}_f, \boldsymbol{v}_f\rangle > 0$, then for any $\kappa > 1$,

$$\langle \boldsymbol{a}_f, \kappa\boldsymbol{v}_f\rangle = \kappa\langle \boldsymbol{a}_f, \boldsymbol{v}_f\rangle > \langle \boldsymbol{a}_f, \boldsymbol{v}_f\rangle. \tag{122}$$

Proof. This follows immediately from linearity in $\boldsymbol{v}_f$. □

We therefore parameterize

$$\boldsymbol{v}_f = m_f\boldsymbol{d}_f, \tag{123}$$

where $\boldsymbol{d}_f$ specifies the spatial shape, including relative local magnitudes, while m_f determines the global transport scale. For shape loss, the maximizer is derived as follows.

Proposition 12 (Optimal regularized transport shape)**.** *Because $\boldsymbol{L}_{\boldsymbol{d}}$ is linear in $\boldsymbol{d}$, there exists a positive-semidefinite matrix $\boldsymbol{M}$ such that for $\mu > 0$, the maximizer of*

$$\frac{\boldsymbol{a}_f^\top \boldsymbol{d}}{\sqrt{\boldsymbol{d}^\top(\boldsymbol{M} + \mu\boldsymbol{I})\boldsymbol{d}}} \tag{124}$$

satisfies

$$\boldsymbol{d}_f^\star \propto (\boldsymbol{M} + \mu\boldsymbol{I})^{-1}\boldsymbol{a}_f. \tag{125}$$

Proof. Let

$$y = (\boldsymbol{M} + \mu\boldsymbol{I})^{1/2}\boldsymbol{d}. \tag{126}$$

Then

$$\frac{\boldsymbol{a}_f^\top \boldsymbol{d}}{\sqrt{\boldsymbol{d}^\top (\boldsymbol{M}+\mu\boldsymbol{I})\boldsymbol{d}}} = \frac{\boldsymbol{a}_f^\top (\boldsymbol{M}+\mu\boldsymbol{I})^{-1/2}\boldsymbol{y}}{\|\boldsymbol{y}\|_2} \tag{127}$$

$$= \frac{\left[(\boldsymbol{M}+\mu\boldsymbol{I})^{-1/2}\boldsymbol{a}_f\right]^\top \boldsymbol{y}}{\|\boldsymbol{y}\|_2}. \tag{128}$$

By Cauchy–Schwarz, this is maximized when

$$\boldsymbol{y} \propto (\boldsymbol{M}+\mu\boldsymbol{I})^{-1/2}\boldsymbol{a}_f. \tag{129}$$

Substituting back gives

$$\boldsymbol{d} \propto (\boldsymbol{M}+\mu\boldsymbol{I})^{-1}\boldsymbol{a}_f. \tag{130}$$

□

Proposition 13 (Characteristic magnitude)**.** *Given the normalized shape $\widehat{d}_f$, define*

$$\boldsymbol{u}_f = \boldsymbol{L}_{\widehat{\boldsymbol{d}}_f}\boldsymbol{g}_f, \qquad v_f = m_f \widehat{d}_f. \tag{131}$$

Then

$$\boldsymbol{L}_{\boldsymbol{v}_f}\boldsymbol{g}_f = m_f \boldsymbol{u}_f. \tag{132}$$

The magnitude that matches the norm of the first-order transport response to the source pullback,

$$\|\boldsymbol{L}_{\boldsymbol{v}_f}\boldsymbol{g}_f\|_2 = \|\boldsymbol{g}_f\|_2, \tag{133}$$

is

$$\lambda_f^\star = \frac{\|\boldsymbol{g}_f\|_2}{\|\boldsymbol{u}_f\|_2}. \tag{134}$$

Proof. By linearity of the generator in its velocity field,

$$\boldsymbol{L}_{m_f\widehat{\boldsymbol{d}}_f}\boldsymbol{g}_f = m_f \boldsymbol{L}_{\widehat{\boldsymbol{d}}_f}\boldsymbol{g}_f = m_f\boldsymbol{u}_f. \tag{135}$$

Therefore, equation 133 gives

$$m_f\|\boldsymbol{u}_f\|_2 = \|\boldsymbol{g}_f\|_2, \tag{136}$$

which yields equation 134. □

B.7 Warped diffusion as a family of conditional probability paths

The second training stage introduces two diffusion coordinates. The teacher timestep t determines the temporal noise prior inferred from the current base-model trajectory, while the student timestep s indexes a diffusion path under that fixed prior.

Definition 6 (Teacher-indexed warped noise prior)**.** *Let*

$$\widehat{\boldsymbol{x}}_{0,t} = \widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t) \tag{137}$$

be the clean estimate produced by the frozen base denoiser at teacher timestep t, and let

$$\mathbb{V}_t = \mathcal{F}(\widehat{\boldsymbol{x}}_{0,t}, t) \tag{138}$$

be the corresponding predicted warp field. Starting from

$$\boldsymbol{\epsilon}^{(1)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{139}$$

the autoregressive NoiseFlow construction defines

$$\boldsymbol{\epsilon}_t = \mathcal{W}\left(\boldsymbol{\epsilon}^{(1)}; \mathbb{V}_t\right). \tag{140}$$

Conditioned on the inferred warp, we write

$$\boldsymbol{\epsilon}_t \sim \mathcal{N}\left(\boldsymbol{0}, \Sigma_t(\widehat{\boldsymbol{x}}_{0,t})\right). \tag{141}$$

Proposition 14 (Framewise marginals of the warped prior)**.** *Each frame of the warped noise prior in Definition 6 has standard Gaussian marginal,*

$$p(\boldsymbol{\epsilon}_{t,f}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \qquad \forall f, \tag{142}$$

while its joint spatiotemporal covariance $\Sigma_t(\widehat{\boldsymbol{x}}_{0,t})$ may contain nonzero cross-frame blocks.

Proof. Each autoregressive frame transition is a composition of Gaussian-preserving transport and innovation steps by Proposition 6. However, temporal transitions introduce cross-frame covariance according to Corollary 2. □

Proposition 15 (Warped diffusion marginal)**.** *Let*

$$x_t^{\mathrm{T}} \sim q_t(\,\cdot\,|x_0)$$

denote the teacher noisy state and $\hat{x}_{0,t} = \hat{x}_0(x_t^{\mathrm{T}}, t)$. The unconditional marginal is then

$$\begin{aligned} p_{s,t}(x) = \iint \mathcal{N}\left(x; \alpha_s x_0,\, \sigma_s^2 \Sigma_t\big(\hat{x}_0(x_t^{\mathrm{T}}, t)\big)\right) \\ \times\, q_t(x_t^{\mathrm{T}} \mid x_0) p_{\mathrm{data}}(x_0)\, dx_t^{\mathrm{T}}\, dx_0. \end{aligned} \tag{143}$$

Proof. For fixed $\boldsymbol{x}_0$ and $\widehat{\boldsymbol{x}}_{0,t}$, Definition 6 gives

$$\epsilon_t \sim \mathcal{N}\left(\mathbf{0}, \Sigma_t(\widehat{\boldsymbol{x}}_{0,t})\right). \tag{144}$$

Marginalizing over $\boldsymbol{x}_0, \widehat{\boldsymbol{x}}_{0,t}$ gives equation 143. □

Corollary 5 (Fixed warped prior as a special case)**.** *If the warp generator is independent of the teacher diffusion state and instead is conditioned directly on the clean latents, so that*

$$\mathcal{F}(\widehat{\boldsymbol{x}}_{0,t}, t) = \mathcal{F}(\boldsymbol{x}_0), \tag{145}$$

then Σ_t is independent of t and the construction reduces in capacity to a single fixed warped-noise diffusion family.

B.8 Interpretation of orthogonal warped-trajectory guidance

Proposition 16 (First-order preservation of text guidance)**.** *Let*

$$\boldsymbol{d}_c = (\widehat{\boldsymbol{x}}_{0,c} - \widehat{\boldsymbol{x}}_{0,\varnothing}) \propto \nabla_{\boldsymbol{x}_t} \log p_t(c \mid \boldsymbol{x}_t), \qquad \boldsymbol{P}_c^{\perp} = \boldsymbol{I} - \frac{\boldsymbol{d}_c \boldsymbol{d}_c^{\top}}{\|\boldsymbol{d}_c\|^2}, \tag{146}$$

and let $\boldsymbol{r}$ be any guidance residual represented in the same score coordinates. Then the projected residual

$$\boldsymbol{r}^{\perp} = \boldsymbol{P}_c^{\perp} \boldsymbol{r} \tag{147}$$

satisfies

$$\frac{d}{d\eta} \log p_t\left(c \mid \boldsymbol{x}_t + \eta \boldsymbol{r}^{\perp}\right)\Big|_{\eta=0} = 0. \tag{148}$$

Proof. Since

$$\frac{d}{d\eta} \log p_t\left(c \mid \boldsymbol{x}_t + \eta \boldsymbol{r}^{\perp}\right)\Big|_{\eta=0} \propto \boldsymbol{d}_c^{\top} \boldsymbol{r}^{\perp}, \qquad \boldsymbol{d}_c^{\top} \boldsymbol{r}^{\perp} = 0, \tag{149}$$

the proof follows. □

Proposition 17 (Constrained warped guidance)**.** *Assume additionally that the warped and base guided predictions define the true scores $\boldsymbol{s}_{\mathcal{W}}$ and $\boldsymbol{s}_{\mathcal{N}}$ at a common noisy state, and let*

$$\boldsymbol{r} = \boldsymbol{s}_{\mathcal{W}} - \boldsymbol{s}_{\mathcal{N}} = \nabla_{\boldsymbol{x}_t} \log \frac{p_{\mathcal{W}}(\boldsymbol{x}_t)}{p_{\mathcal{N}}(\boldsymbol{x}_t)}. \tag{150}$$

Among infinitesimal perturbations of fixed norm that preserve the text-conditioning likelihood to first order, the direction that maximizes the first-order increase in $\log p_{\mathcal{W}} - \log p_{\mathcal{N}}$ is

$$\boldsymbol{r}^{\perp} = \boldsymbol{P}_c^{\perp} \boldsymbol{r}. \tag{151}$$

Proof. Consider

$$\max_{\delta \boldsymbol{x}} \quad \boldsymbol{r}^\top \delta \boldsymbol{x} \tag{152}$$

subject to

$$\boldsymbol{g}_c^\top \delta \boldsymbol{x} = 0, \qquad \|\delta \boldsymbol{x}\|_2 = 1. \tag{153}$$

The feasible vectors lie in the orthogonal complement of $\boldsymbol{g}_c$. The component of $\boldsymbol{r}$ in this subspace is $\boldsymbol{P}_c^\perp \boldsymbol{r}$. By Cauchy–Schwarz, its normalized direction maximizes the inner product over the feasible set. Therefore

$$\delta \boldsymbol{x}^\star \propto \boldsymbol{P}_c^\perp \boldsymbol{r}. \tag{154}$$

□

Remark 4 (Scope of the probabilistic interpretation)**.** *The constrained-density interpretation in Proposition 17 assumes that the base and warped scores are represented in a common state coordinate. In NoiseFlow, however, the warped student is evaluated after transferring the current base trajectory to the warped-noise path. Consequently, $\widehat{\boldsymbol{x}}_{0,W} - \widehat{\boldsymbol{x}}_{0,N}$ should not be interpreted as an exact score difference between two densities evaluated at an identical noisy state.*

We therefore use Proposition 17 as the local geometric motivation for the guidance rule rather than as a claim that the implemented vector field is the score of a globally normalized target distribution. The implemented projection has the exact algebraic property

$$\left\langle \widehat{\boldsymbol{x}}_{0,c} - \widehat{\boldsymbol{x}}_{0,\varnothing}, \boldsymbol{r}_t^\perp \right\rangle = 0, \tag{155}$$

and, under the common-state exact-score idealization, this property corresponds to preserving the base text-conditioning likelihood to first order.

C Implementation Details

In this section, we describe the implementation choice to choose the number of Cayley iterations K, and the two algorithms that implement the key aspects of our method, namely the warping and the sampling process.

Algorithm 1 describes the realization of the warping method in pseudocode. A key design choice here is choosing the upper bound for the number of Cayley iteration steps K_f introduced in Section 3.1 and formalized in Appendix B.4. First, we need to determine the Cayley step using an inner optimization loop by iteratively solving equation 80 for each $k \in K_f$. Here we use a fixed desired tolerance δ_{min} to converge within an orthogonal solution with a hard maximum number of iterations $N_{\max}$. To identify the number of outer steps K_f, the following heuristic is used:

$$K_f = \max\left(1, \lceil \|\boldsymbol{v}_x\|_\infty + \|\boldsymbol{v}_y\|_\infty \rceil\right) \tag{156}$$

Thus, we scale K_f with warp-velocity magnitude. We set $\delta_{\min} = 10^{-3}$ and $N_{\max} = 20$ in our experiments. In the worst case, the number of steps involved in the transition per frame is $K_f \times N_{\max}$. However, this limit is never empirically realized in our experiments, where N remains around 3 iterations, while K stays close to 30 ± 5 on average.

We provide the sampling procedure in Algorithm 2 to complement the discussion in Section 3.3 and Figure 1a. Herein, we use $\widehat{\boldsymbol{x}}_{(0,\cdot)}$ to mean the implied clean video prediction derived from the respective denoiser's native prediction quantity.

D Limitations and future work

Several limitations remain. First, the learned warp is identified only through local denoiser geometry. The pullback objective determines transport directions that increase adjacent-frame denoiser-sensitive covariance, but it does not establish that the resulting field is unique or that it corresponds to a physical motion field. The shape regularization and magnitude construction are, therefore, part of the inductive bias required to select a useful transport from an underdetermined family. Second, our stochastic innovation rate is tied to positive local divergence of the learned field. This provides a simple coupling between transport and renewal of randomness, but it is a modeling choice rather

Algorithm 1 NoiseFlow warping

Require: Warp generator $\mathcal{F}$, Predicted clean state $\widehat{\boldsymbol{x}}_0$, Timestep t, Cayley-step-finding error tolerance $\delta_{\min}$, max-iterations $N_{\max}$
Ensure: Warped noise $\boldsymbol{\epsilon}_t$

procedure CAYLEY_STEP($\boldsymbol{v}, \boldsymbol{\epsilon}, K_f$)
  $a \leftarrow 1/(2K_f)$
  $b \leftarrow \boldsymbol{\epsilon} + a\boldsymbol{L_v}\boldsymbol{\epsilon}$ ▷ RHS, equation 80
  $\boldsymbol{\epsilon}' \leftarrow \boldsymbol{\epsilon}$
  for $i = 1, \ldots, N_{\max}$ **do**
    $\tilde{\boldsymbol{\epsilon}} \leftarrow b + a\boldsymbol{L_v}\boldsymbol{\epsilon}'$ ▷ LHS, equation 80
    $\delta \leftarrow \|\tilde{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}'\|/\|b\|$ ▷ Calculate residual
    $\boldsymbol{\epsilon}' \leftarrow \tilde{\boldsymbol{\epsilon}}$
    if $\delta \leq \delta_{\min}$ **then** ▷ Converge if true
      break
    end if
  end for
  return $\boldsymbol{\epsilon}'$
end procedure

procedure WARPING($\mathcal{F}, \widehat{\boldsymbol{x}}_0, t$)
  $\mathbb{V} \leftarrow \mathcal{F}(\widehat{\boldsymbol{x}}_0, t)$ ▷ Predict $\mathbb{V}$ with $\mathcal{F}$
  $\boldsymbol{\epsilon}_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ ▷ Sample first i.i.d. Gaussian frame
  $\mathcal{E} \leftarrow \{\boldsymbol{\epsilon}_1\}$
  for all $f \in \{1, \ldots, F-1\}$ **do**
    $\boldsymbol{v}_f \leftarrow \mathrm{Vec}(\mathbb{V}_f)$
    $\boldsymbol{B} \leftarrow \max(0, \nabla \cdot \boldsymbol{v}_f)$ ▷ Positive local divergence
    $K_f \leftarrow \max(1, \lceil \|\boldsymbol{v}_x\|_\infty + \|\boldsymbol{v}_y\|_\infty \rceil)$
    $\boldsymbol{\rho} \leftarrow \exp(-\frac{1}{4K_f}\boldsymbol{B}), \quad \boldsymbol{\sigma} \leftarrow \sqrt{1-\boldsymbol{\rho}^2}$ ▷ OU half-step
    for $k = 1, \ldots, K_f$ **do**
      $\boldsymbol{\epsilon}_f \leftarrow \boldsymbol{\rho}\,\boldsymbol{\epsilon}_f + \boldsymbol{\sigma}\,\xi, \quad \xi \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
      $\boldsymbol{\epsilon}_f \leftarrow$ CAYLEY_STEP($\boldsymbol{v}_f, \boldsymbol{\epsilon}_f, K_f$)
      $\boldsymbol{\epsilon}_f \leftarrow \boldsymbol{\rho}\,\boldsymbol{\epsilon}_f + \boldsymbol{\sigma}\,\xi, \quad \xi \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
    end for
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{\boldsymbol{\epsilon}_f\}$
  end for
  $\boldsymbol{\epsilon}_t \leftarrow \mathcal{E}$
  return $\boldsymbol{\epsilon}_t$
end procedure

Algorithm 2 Sampling from NoiseFlow

Require: Warp generator $\mathcal{F}$, Frozen base denoiser $\mathcal{G}$, NoiseFlow warp trained denoiser $\mathcal{G}'$, Noisy state $\boldsymbol{x}_t$, Conditioning c, Warp-guidance scale γ, CFG-scale w
Ensure: Noisy state $\boldsymbol{x}_{t-1}$

procedure SAMPLE($\boldsymbol{x}_t, t, c, \mathcal{G}, \mathcal{G}', \mathcal{F}, w, \gamma$)
  $\widehat{\boldsymbol{x}}_{0,\mathcal{N}} \leftarrow \mathrm{CFG}(\mathcal{G}, \boldsymbol{x}_t, c, w)$ ▷ Predict $\widehat{\boldsymbol{x}}_{0,\mathcal{N}}$ with standard CFG applied to $\mathcal{G}$
  $\boldsymbol{\epsilon}_t \leftarrow$ WARPING($\mathcal{F}, \widehat{\boldsymbol{x}}_{0,\mathcal{N}}, t$) ▷ Compute warped noise at t
  $\boldsymbol{x}_{t,\mathcal{W}} \leftarrow$ ADD_NOISE($\widehat{\boldsymbol{x}}_{0,\mathcal{N}}, t, \boldsymbol{\epsilon}_t$)
  $\widehat{\boldsymbol{x}}_{0,\mathcal{W}} \leftarrow \mathcal{G}'(\boldsymbol{x}_{t,\mathcal{W}}, t, t, c)$ ▷ Predict $\widehat{\boldsymbol{x}}_{0,\mathcal{W}}$ with $\mathcal{G}'$
  $\boldsymbol{r}_t \leftarrow \widehat{\boldsymbol{x}}_{0,\mathcal{W}} - \widehat{\boldsymbol{x}}_{0,\mathcal{N}}$
  $\boldsymbol{r}_t^{\perp} \leftarrow \boldsymbol{P}_c^{\perp}\boldsymbol{r}_t$ ▷ Compute orthogonal component
  $\widehat{\boldsymbol{x}}_0 \leftarrow \widehat{\boldsymbol{x}}_{0,\mathcal{N}} + \gamma\boldsymbol{r}_t^{\perp}$ ▷ Compute guided prediction
  $\boldsymbol{x}_{t-1} \leftarrow$ SCHEDULER_STEP($\widehat{\boldsymbol{x}}_0, \boldsymbol{x}_t, t$)
  return $\boldsymbol{x}_{t-1}$
end procedure

than a consequence of Gaussian preservation, and more expressive models of innovation are possible. Third, the current construction is autoregressive across frames, which is consistent with the existing literature. However, more general spatiotemporal, channel-dependent, or nonlocal operators could represent dependencies that cannot be expressed by a two-dimensional velocity field. From an implementation perspective, the principal limitation is the requirement for an additional model evaluation to compute the warp-guidance term. Nevertheless, we posit that the associated performance improvements outweigh the incurred computational cost.

Because the warp generator is trained from the local geometry of a particular pretrained model, the learned structure is model specific. Therefore, the learned warp would inherit the imperfections resulting from the linearized pullback around the difficult or undertrained regions of an otherwise nonlinear denoiser-geometry. This dependence is also what distinguishes NoiseFlow from externally specified motion priors. Studying whether such priors transfer between model families, whether they can be learned jointly across multiple denoisers, and whether richer stochastic transports yield further gains are natural directions for future work.

E Additional qualitative results

In this section, we provide additional visual examples to supplement Figures 3 and 4. First, we provide visualization of warp and innovation rate maps for additional real videos from the dataset in Figures 6 and 7. Then, we provide visualizations of these fields for the videos generated by NoiseFlow during inference in Figures 8 and 9. Finally, we provide a visual comparison of different models under the same noise seed in Figures 11 and 12.

F Setup details

In this section, we cover the data preprocessing details and enumerate the hyperparameter settings that were used in the reported results.

Dataset and preprocessing. We use the WISA (Wang et al., 2025b) dataset in all our experiments. WISA contains more than 80 thousand videos that represent rich temporal dynamics relevant to understanding the physical phenomena of the real world, including motion and beyond. The denoiser $\mathcal{G}$ is instantiated as CogvideoX-2B (Yang et al., 2025) in our experiments. To process the dataset to match the format expected by the model, we take an approach similar to Pathak et al. (2026). Here we follow the model configuration in CogvideoX that generates 6 second clips at resolution 720×480 and 8fps. We remove videos less than 6 seconds long and divide the remaining videos randomly into a training set and a hold-out test set using a 90–10 split. The 6 second clips are sampled at 720×480 at 8 frames per second. This results in approximately 39 thousand video clips.

Model and training. For $\mathcal{F}$, we use a small U-Net with transformer-based motion adapter layers in an architecture largely inspired by Guo et al. (2024) to generate the velocity field. Our resultant model is small, with only ≈ 25M parameters, compared to the denoiser scale. We use two heads with one modeling the spatially varying field $\boldsymbol{d}_f$, while another head outputs a per-frame magnitude scalar m_f. The student copy of the base denoiser model is augmented with new dedicated timestep embedding layers to incorporate an additional conditioning signal for the teacher diffusion timestep. The student $\mathcal{G}'$ is fully fine-tuned, while the pretrained teacher $\mathcal{G}$ is used without any finetuning.

Hyperparameters and inference times Table 3 reports the hyperparameters used for each model and the corresponding mean inference times under non-batched inference conditions. For inference, we adopt the default hyperparameter configurations specified for each baseline (e.g., guidance scale). The inference times reported are for a single NVidia B200 GPU.

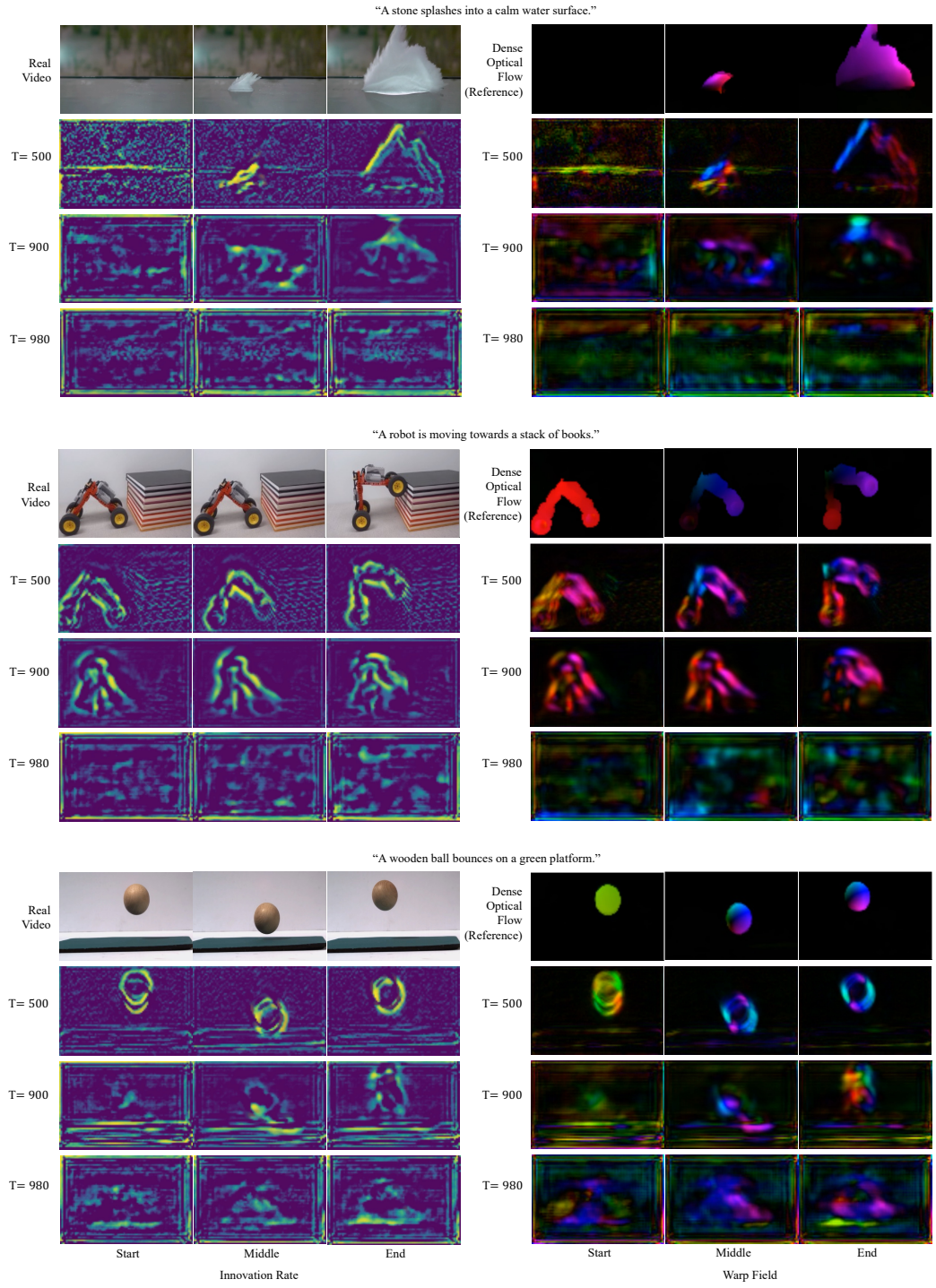


Figure 6: Additional examples of warp field visualization on real videos.

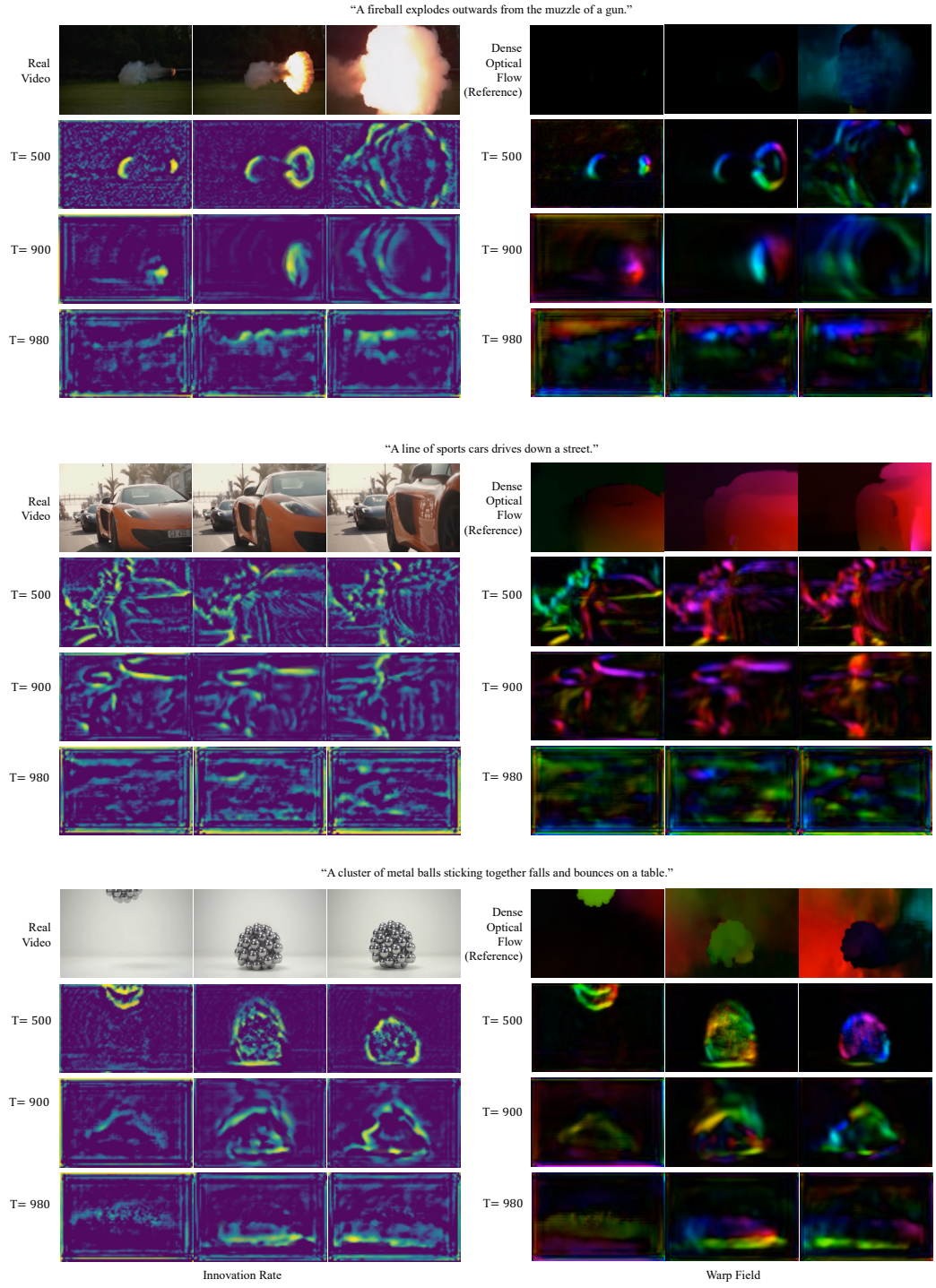


Figure 7: Additional examples of warp field visualization on real videos.

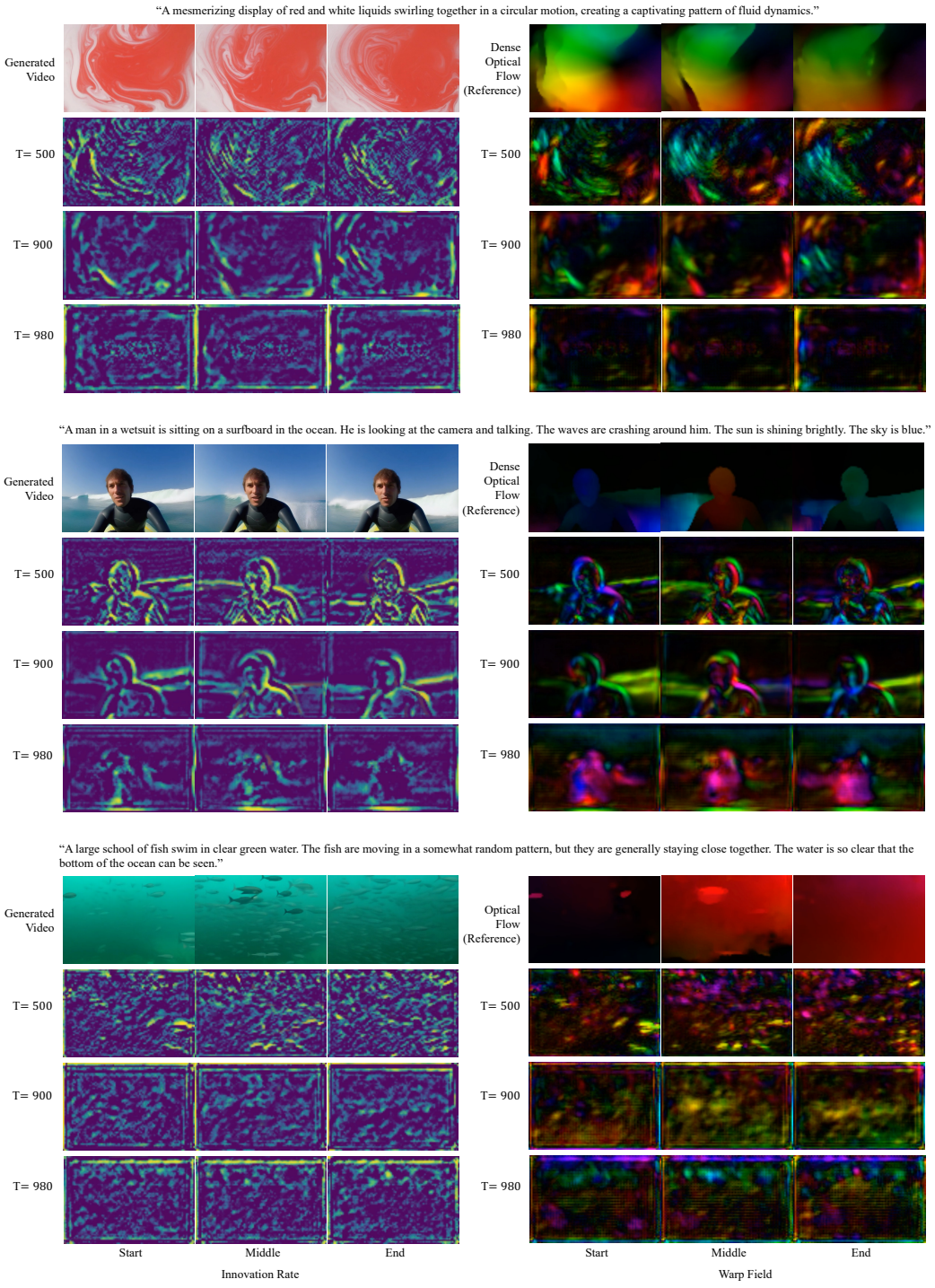


Figure 8: Warp field realization on videos generated by NoiseFlow during inference.

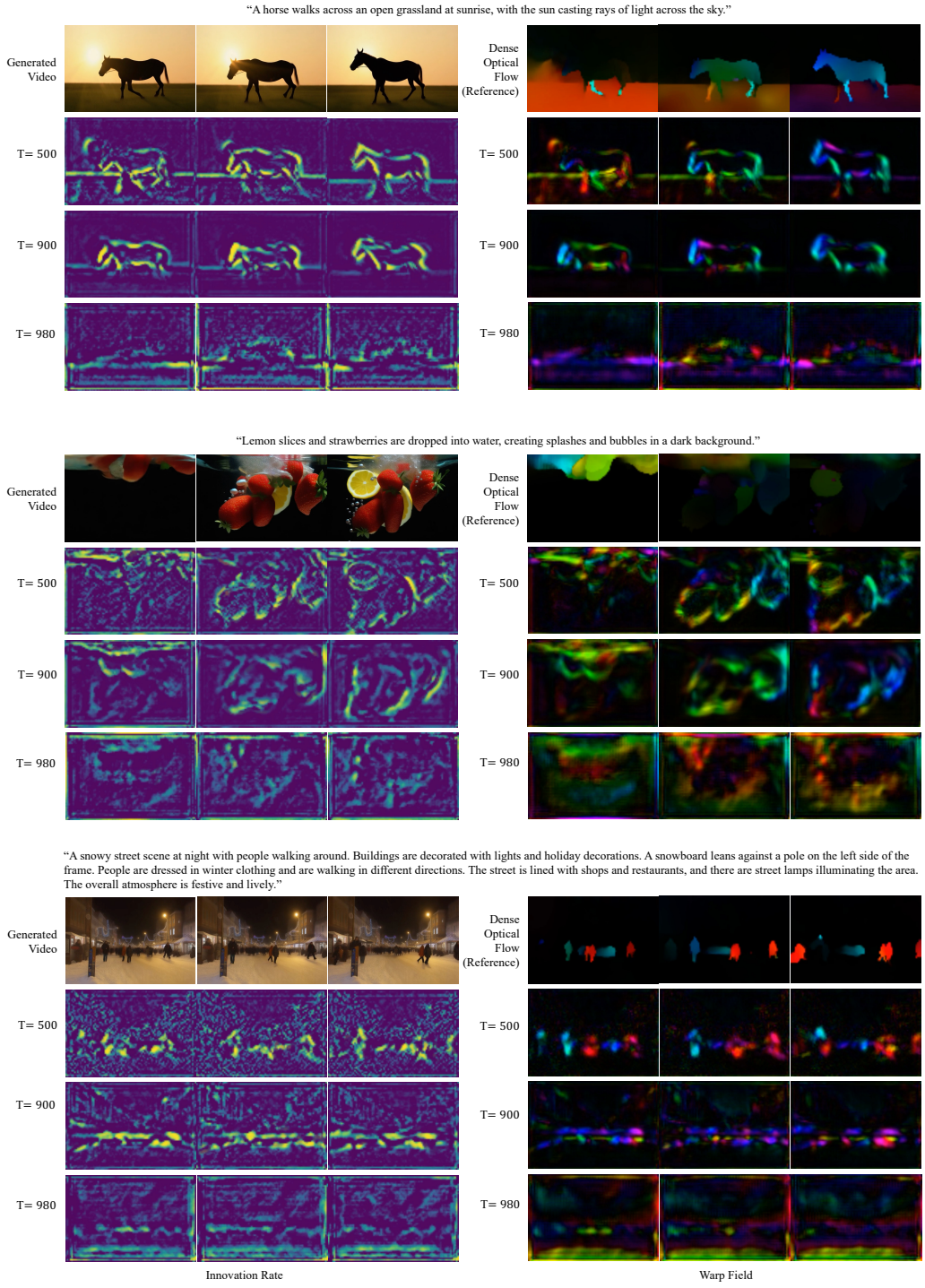


Figure 9: Warp field realization on videos generated by NoiseFlow during inference.

"A woman in a business suit takes a sip from a blue cup while gazing out of a window, with a plant in the background."



"A group of people are performing on stage at night. The singer is in the foreground, holding a microphone and singing. There are two other musicians in the background playing guitars. The stage is lit with colorful lights, and there is a tent in the background."



"Two boxers spar in a dimly lit room with spotlights, surrounded by smoke, wearing protective gear and boxing gloves."



"A group of people are sitting around a table, eating and drinking. There are plates with food, bowls, glasses, and bottles on the table. Some people are holding glasses and pouring drinks."



"Red chili peppers falling into water, creating a splash. The peppers are shown in mid-air before hitting the water, causing a dynamic and colorful splash."



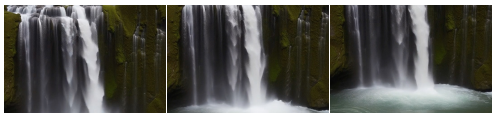
"A large white airplane with a colorful tail takes off from the runway at an airport. Other airplanes are visible on the ground in the background."



"A light fixture on a brick wall. The camera slowly zooms in on the light, which is the main focus of the video. The light is bright and casts a warm glow on the wall. The video ends with a close-up of the light bulb."



"A waterfall cascades down a rocky cliff covered in moss, crashing into a pool below and creating a misty spray."



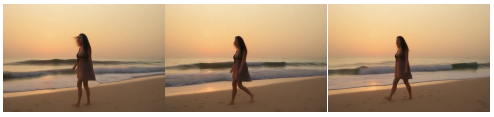
"A man tends to a fire in a circular arrangement of cement blocks, adding logs to make it burn brighter."



"People walking down a busy city street at night with neon lights and stores. A man and woman walk together, while another woman walks alone. A man sits on a bench, and a woman takes a selfie. The scene is illuminated by streetlights and neon signs."



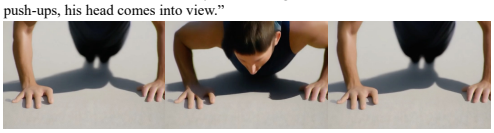
"A woman walks along the beach at sunset, with the sun setting in the background. The waves gently lap at the shore as she strolls along the sand."



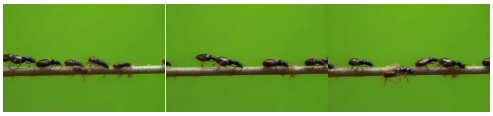
"A person kneads dough on a floured table, their hands covered in flour as they press and fold the dough."



"A man is doing push-ups on a concrete surface. His hands and arms are in focus, with the shadow of his body cast on the ground. As he continues to do push-ups, his head comes into view."



"A group of ants are crawling on a wire against a green background. The ants are moving in different directions, and some of them are carrying food. The wire is covered in moss, and there is a small amount of water on the wire. The ants are black and red, and they are about 1/4 inch long."



"A drop of blue ink falls into water, creating a mesmerizing pattern as it disperses and swirls, forming various shapes and patterns in slow motion."



"Two hikers with backpacks walking along a mountain trail at sunrise or sunset. The sun is shining brightly behind them, creating a silhouette effect. The mountains in the background are covered in trees and the sky is clear with some clouds."



"A swimmer in a yellow swim cap and black swim trunks swims freestyle in a pool, viewed from below the water surface."



"A close-up shot of a person's hand playing a guitar, strumming the strings and pressing down on the frets."



Figure 10: Additional examples of videos generated with NoiseFlow ($\gamma = 0.8$).



Figure 11: Comparison of all models under the same seed. For models with same base latent structure, sharing the seed also implies identical starting noise. Wan and LTX-Video use different latent shapes hence they start sampling from non-identical states under shared seed configuration.



Figure 12: Comparison of all models under the same seed. For models with same base latent structure, sharing the seed also implies identical starting noise. Wan and LTX-Video use different latent shapes hence they start sampling from non-identical states under shared seed configuration.

Table 3: Additional configurations and inference latency

Training	
Base architecture	CogVideoX-2B
Warp generator $\mathcal{F}$ parameters (M)	25
Additional time conditioning parameters in $\mathcal{G}'$ (M)	37
Effective batch size	64
Number of Phase-I training steps	3000
Number of Phase-II training steps	3000
Number of data epochs	5
Maximum Learning rate (Phase I: Warp generator)	1×10^{-4}
Minimum Learning rate (Phase I: Warp generator)	1×10^{-5}
Maximum Learning rate (Phase II: Denoiser)	1×10^{-5}
Minimum Learning rate (Phase II: Denoiser)	2×10^{-6}
LR Schedule	Cosine
Loss	V-Prediction
Optimizer	AdamW
Inference	
Number of Denoising Steps	50
Guidance Scale (Wan, CogvideoX and NoiseFlow)	6
Guidance Scale (LTX-Video)	3
CogVideoX Latency per Video (s)	32
Wan Latency per Video (s)	66
LTX Latency per Video (s)	11
NoiseFlow Latency per Video (s)	65