

ART-VSR: Adaptive Rectified Trajectories for One-Step Video Super-Resolution

Jianhui Zhang¹, Chen Fang², Hao Wang³, Chaoyu Feng³, Lei Lei³, Jue Wang²,
and Shuaicheng Liu^{1*}

¹ University of Electronic Science and Technology of China, Chengdu, China

² Dzine AI, SeeKoo LLC, United States

³ Independent Researcher, China

Abstract. Generative priors have advanced VSR, yet their multi-step inference remains costly. Recent trajectory-based one-step VSR methods accelerate generation from low-resolution (LR) inputs, but can struggle under complex real-world degradations. We analyze this limitation as *trajectory drift*, caused by **initial state mismatch** and a **global timestep constraint**. We propose **ART-VSR**, a **state-time adaptation** framework that estimates a continuous token-wise timestep map with an Adaptive Timestep Estimator (ATE) and uses it to guide a Latent Trajectory Rectifier (LTR) toward a compatible starting state. Conditioned on the rectified latent and spatial timestep map, ART-VSR achieves highly competitive one-step VSR performance, producing crisp, artifact-reduced details while preserving temporal consistency. Code: https://github.com/Roveer/ART_VSR.

Keywords: Video Super-Resolution · One-Step VSR · Flow Matching · Trajectory Rectification

1 Introduction

Video Super-Resolution (VSR) is critical for high-fidelity streaming and mobile communication [2, 3]. While traditional regression-based methods [2, 3, 13, 34] offer efficiency, they often yield over-smoothed textures due to the regression-to-the-mean effect. Recently, generative models [10, 22, 29, 41] have pushed the perceptual boundaries of video restoration [39, 40]. However, their heavy multi-step sampling process imposes a prohibitive computational bottleneck, making real-world deployment largely impractical [8, 38, 47]. Consequently, One-Step VSR has emerged as a promising frontier to bridge high-quality generation and inference efficiency [5, 27, 36, 49].

Within generative VSR, two strategies prevail. Condition-based approaches [1, 8, 15, 38, 44, 49] synthesize content from pure noise using the low-resolution (LR) input as an external constraint. Although the LR is utilized as a guide, this external injection mechanism often fails to fully preserve deterministic structural

* Corresponding author.

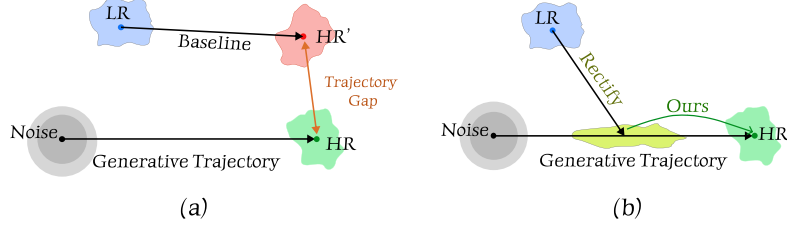


Fig. 1: Conceptual illustration of trajectory drift. (a) The Problem: When uncalibrated LR inputs are directly used for fine-tuning, the learned generation path (Baseline [5, 36]) can deviate from the reference generative trajectory. This creates an observable trajectory gap and may terminate at a degraded output (HR'). **(b) Our Solution:** ART-VSR introduces a rectification operation to pull the off-path LR features closer to the reference trajectory, enabling the corrected path (Ours) to better approach the HR manifold.

priors, leading to structural hallucinations and temporal flickering. Alternatively, Trajectory Adjustment methods [5, 12, 36, 43] transition directly from the LR input to the high-resolution (HR) target. To leverage powerful pre-trained priors, they fine-tune the base model to bridge the degraded LR and the HR target. Ideally, this task-specific trajectory should smoothly inherit the rich geometry of the pre-trained prior’s authentic generative trajectory.

Despite their appeal, current trajectory-based methods [5, 36, 43] can struggle in complex, in-the-wild scenarios. We analyze this behavior as **trajectory drift**, which is associated with two compounding limitations in existing architectures:

- 1) **Initial state mismatch:** By injecting uncalibrated LR inputs directly during fine-tuning, existing approaches implicitly assume the network can adapt its pre-trained prior to these out-of-distribution states. In practice, the geometric mismatch between real-world degradations and expected noise distributions can distort the learned vector field. To visualize this, we reproduce the representative trajectory-adjustment paradigm [5] on the Wan2.1 [29] base model, fine-tuned on our exact dataset. As illustrated in Fig. 1a, generation from uncalibrated LR inputs does not always integrate smoothly into the reference generative trajectory, leaving a trajectory gap and potentially terminating at a degraded state (HR').
- 2) **Global timestep constraint:** Existing methods typically apply a globally uniform starting timestep across the entire frame. As demonstrated in Fig. 2, varying this global timestep introduces a trade-off: small timesteps leave textures over-smoothed, while large timesteps may corrupt clean backgrounds with noise artifacts. Since degradation severity can vary within a single frame, a fixed global integration path may be insufficient for all regions.

To address these challenges, we propose **ART-VSR** (Adaptive Rectified Trajectories for VSR). Instead of forcing a direct transition from uncalibrated inputs, ART-VSR anchors the generation process via a *state-time adaptation* mechanism. First, an Adaptive Timestep Estimator (ATE) evaluates local frequency variations to generate a continuous t -map, assigning an independent integration

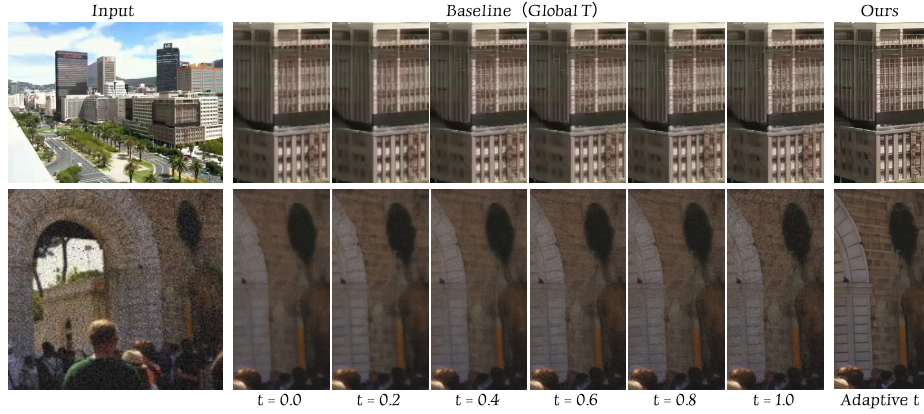


Fig. 2: The dilemma of global timestep constraint. Direct generation from out-of-distribution inputs under a globally fixed timestep introduces an observable trade-off: small timesteps ($t \leq 0.4$) yield over-smoothed outputs, while large timesteps ($t \geq 0.6$) may hallucinate noise artifacts. By employing state-time adaptation to align the initial state and dynamically allocate local integration lengths, ART-VSR mitigates this dilemma and restores crisp, artifact-reduced details.

length to each spatial token. Second, a Latent Trajectory Rectifier (LTR) uses this t-map to move off-path LR features toward a compatible rectified starting state. The DiT is then conditioned on both the rectified state and the t-map to synthesize high-fidelity HR frames in a single integration step.

The main contributions of this work are summarized as follows:

- We analyze the **trajectory drift** phenomenon in flow-based one-step VSR, highlighting **initial state mismatch** and **global timestep constraint** as important bottlenecks.
- We propose **ART-VSR** with a state-time adaptation mechanism, coupling an **Adaptive Timestep Estimator (ATE)** for local integration scheduling and a $\mathbf{T}_{\text{map}}$ -guided **Latent Trajectory Rectifier (LTR)** for rectified initialization.
- Extensive experiments show that ART-VSR achieves highly competitive one-step VSR performance, reducing structural hallucinations and producing high-fidelity restoration under complex real-world degradations.

2 Related Work

2.1 Video Super-Resolution

Video Super-Resolution (VSR) aims to restore high-frequency details from low-resolution video sequences. Traditional VSR predominantly formulates restoration as a deterministic regression task. Early approaches utilized Convolutional Neural Networks (CNNs) [7, 17, 28, 34] and Recurrent Neural Networks (RNNs)

[2–4] to align and aggregate temporal features. More recently, Vision Transformers [13, 14, 48] have further pushed the limits of spatio-temporal modeling. While these regression-based architectures benefit from highly efficient inference pipelines, their reliance on pixel-wise optimization criteria (*e.g.*, L1 or MSE loss) intrinsically leads to statistical averaging. To break this perceptual bottleneck and synthesize finer textures, Generative Adversarial Networks (GANs) [6, 19, 39] were introduced, though they frequently struggle with unstable training dynamics and temporal flickering.

2.2 Generative Models for Video Enhancement

Recently, Denoising Diffusion Probabilistic Models (DDPMs) [9, 26] and Flow Matching [16] have established strong benchmarks in generative modeling. By leveraging powerful visual priors distilled from large-scale text-to-video pre-training [10, 22, 29, 41], generative enhancement models [1, 8, 38, 44] can synthesize highly realistic textures even under severe degradation. However, this generation capability relies on dozens of sequential Neural Function Evaluations (NFEs). Compounded by the heavy 3D nature of video features (*e.g.*, temporal attention), this multi-step iterative sampling imposes a prohibitive computational burden, severely limiting their feasibility for highly efficient, latency-sensitive deployment [40].

2.3 One-Step Generative VSR

To bypass the prohibitive latency of multi-step sampling, foundational efforts have introduced consistency models [20, 25] and adversarial distillation [18, 23] for single-step generation. When adapting these efficient generative priors to VSR, models must be carefully formulated to incorporate degradation conditions. Conditioning-based approaches [27, 31, 32, 49] encode LR sequences as external control signals while generating content from pure Gaussian noise. However, as noted in Sec. 1, forcing the network to synthesize high-frequency details from scratch exacerbates the learning difficulty, frequently inducing severe hallucinations and temporal instability [15].

To circumvent this, trajectory-based frameworks [5, 12, 36, 43] directly map the LR input to the HR target, effectively setting the degraded video as the starting point of the generative trajectory. While this formulation is architecturally advantageous for restoration, current one-step implementations employ a rigid, static generation path. As discussed in Sec. 1, feeding uncalibrated real-world inputs into a globally fixed integration process can suffer from initial state mismatch and limited spatial adaptability, leading to trajectory drift. ART-VSR advances this paradigm by coupling token-wise spatial modulation (ATE) with explicit geometric projection (LTR), dynamically rectifying off-path states and assigning localized integration lengths to mitigate these bottlenecks.

3 Method

3.1 Preliminaries and Spatial-Modulation Formulation

Standard Flow Matching [16, 29] formulates generative modeling as learning a continuous vector field $\mathbf{v}_\theta$ that transports a simple prior distribution (*e.g.*, standard Gaussian noise $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$) to a complex target data distribution (*e.g.*, HR latent representations $\mathbf{z}_0 \sim p_{data}$). The generative process is governed by an Ordinary Differential Equation (ODE) using a global scalar time variable $t \in [1, 0]$:

$$d\mathbf{z}_t = \mathbf{v}_\theta(\mathbf{z}_t, t)dt. \quad (1)$$

In standard One-Step VSR, recent trajectory adjustment methods [5] leverage the LR input to shorten the path. Assuming the LR latent $\mathbf{z}_{LR}$ acts as an intermediate noisy state at a specific global timestep $T_L \in (0, 1)$, the ODE is truncated and approximated via a single Euler step:

$$\mathbf{z}_0 \approx \mathbf{z}_{LR} - \mathbf{v}_\theta(\mathbf{z}_{LR}, T_L) \cdot T_L. \quad (2)$$

However, adhering to a globally fixed scalar T_L fails to account for spatial degradation heterogeneity, leading to the initial state mismatch and the resulting trajectory drift discussed in Sec. 1. To address this limitation, we transition from a global scalar ODE to a *spatially-modulated local Euler approximation*. For a given spatial token index i , the one-step integration is decoupled as:

$$\mathbf{z}_0^i \approx \mathbf{z}_{LR}^i - \tilde{\mathbf{v}}_\theta(\mathbf{z}_{LR}, \mathbf{T}_{map})^i \cdot \mathbf{T}_{map}^i. \quad (3)$$

Here, the token-wise $\mathbf{T}_{map}$ serves as a learned spatial condition for the jointly fine-tuned vector field. The DiT is fine-tuned jointly with ATE/LTR under the full $\mathbf{T}_{map}$, rather than queried as independent token-wise marginals from a frozen global-time model. Eq. (3) extends the global- T_L update in Eq. (2) to a trained spatially-conditioned local Euler approximation, where $\mathbf{T}_{map}^i$ acts as the local integration-length scale.

3.2 Explorations

To motivate our *state-time adaptation* paradigm, we conduct a comprehensive comparative analysis to geometrically and empirically examine the trajectory drift. We establish a baseline adopting the standard trajectory adjustment strategy [5], trained with the same architecture and parameter count as ART-VSR, to isolate the algorithmic paradigm as the primary variable.

Initial State Mismatch. To visualize generative dynamics without single-sample bias, we project the inference trajectories of 100 diverse video samples from the YouHQ dataset [47] onto a normalized 2D latent plane via Gram-Schmidt orthogonalization. By constructing a sample-specific local coordinate system anchored at its respective $\mathbf{z}_{HR}$, we abstract away individual video semantics and isolate the pure geometric progression. For each sample, the X-axis

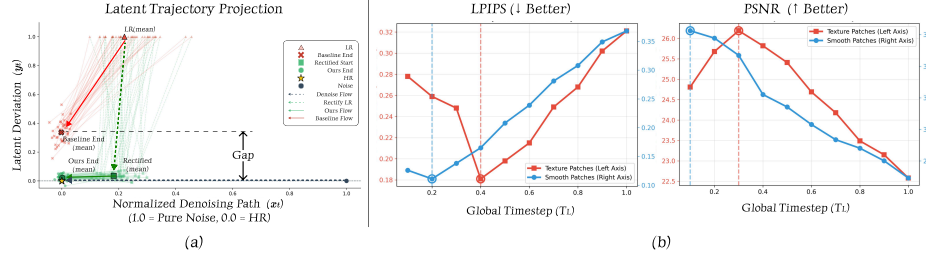


Fig. 3: Comparative analysis of limitations in standard one-step VSR. **(a)** Latent trajectory projection onto a normalized 2D plane via Gram-Schmidt orthogonalization. The baseline flow (red) exhibits **initial state mismatch**, leaving a geometric trajectory gap (y_t). Our rectified flow (green) reduces this orthogonal deviation, pulling the initial state closer to the reference generative path to facilitate a more linear flow to the target. **(b)** Performance across different global timesteps. A globally fixed timestep creates a trade-off between smooth regions (peaking at small timesteps, left axis) and complex textures (peaking at large timesteps, right axis), illustrating the **global timestep constraint**.

basis ($\mathbf{v}_x$) is the ideal linear denoising direction from pure noise $\mathbf{z}_{\text{noise}}$ to the clean HR latent $\mathbf{z}_{\text{HR}}$. The Y-axis basis ($\mathbf{v}_y$) captures the orthogonal residual degradation in the LR input $\mathbf{z}_{\text{LR}}$:

$$\mathbf{v}_x = \frac{\mathbf{z}_{\text{noise}} - \mathbf{z}_{\text{HR}}}{\|\mathbf{z}_{\text{noise}} - \mathbf{z}_{\text{HR}}\|_2}, \quad \mathbf{v}_y = \frac{\mathbf{z}_{\text{ortho}}}{\|\mathbf{z}_{\text{ortho}}\|_2}, \quad (4)$$

where $\mathbf{z}_{\text{ortho}} = (\mathbf{z}_{\text{LR}} - \mathbf{z}_{\text{HR}}) - \langle \mathbf{z}_{\text{LR}} - \mathbf{z}_{\text{HR}}, \mathbf{v}_x \rangle \mathbf{v}_x$, and $\langle \cdot, \cdot \rangle$ denotes the inner product in the flattened high-dimensional latent space. For any intermediate generative state $\mathbf{z}_t$, its 2D coordinates on this plane are calculated via scalar projections onto this basis: $(x_t, y_t) = (\langle \mathbf{z}_t - \mathbf{z}_{\text{HR}}, \mathbf{v}_x \rangle, \langle \mathbf{z}_t - \mathbf{z}_{\text{HR}}, \mathbf{v}_y \rangle)$. This projection is used as a VAE-latent directional analysis and regularization tool, not as an assumption that raw VAE-latent channels are isotropic or physically orthogonal.

Under this canonical mapping, the reference generative trajectory is expected to align with the X-axis ($y_t = 0$). However, a direct transition from off-path real-world LR inputs can cause generation paths to diverge, creating an orthogonal trajectory gap (Fig. 3a). Our rectified states reduce this deviation, anchoring the initial state closer to the reference trajectory to facilitate a more linear generative flow.

Global Timestep Constraint. Real-world videos exhibit varying local degradation complexity. To empirically evaluate this, we sample 100 diverse videos from the YouHQ dataset, extracting 50 smooth and 50 complex-textured 17-frame test patches to assess one-step VSR performance across varying global timesteps T_L . As shown in Fig. 3b, the two groups prefer different operating points: smooth regions achieve higher PSNR at an early timestep to avoid noise corruption, whereas complex textures benefit from later timesteps to reduce LPIPS and recover missing frequencies. This trade-off shows that a globally fixed timestep can limit generation quality.

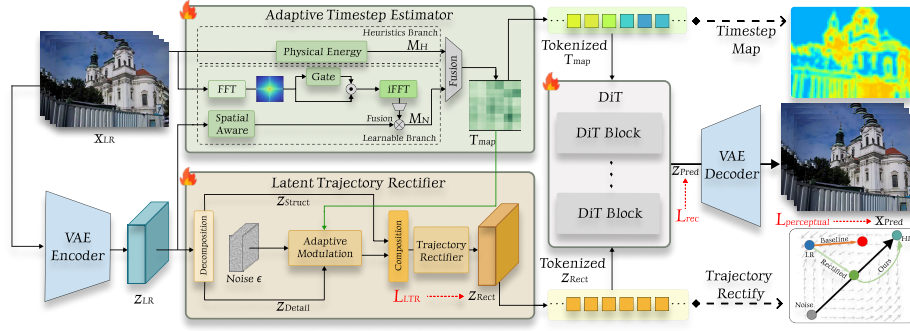


Fig. 4: Overview of the proposed ART-VSR framework. Operating under a **state-time adaptation** paradigm, the architecture consists of two core modules. **Top:** The Adaptive Timestep Estimator (ATE) addresses the global timestep constraint by fusing physical heuristics with deep spectro-spatial features to predict a token-wise timestep map $\mathbf{T}_{\text{map}}$. **Bottom:** The Latent Trajectory Rectifier (LTR) aligns the off-path latent $\mathbf{z}_{\text{LR}}$ to mitigate initial state mismatch. By decoupling structure and detail, and adaptively injecting noise scaled by $\mathbf{T}_{\text{map}}$, the LTR moves the initial state closer to the reference **generative path** before feeding it into the DiT blocks.

Motivation for ART-VSR. These findings motivate ART-VSR (Fig. 4). To resolve the global timestep constraint, we introduce the Adaptive Timestep Estimator (ATE) to estimate a token-wise t -map. To bridge the geometric trajectory gap caused by initial state mismatch, we propose the Latent Trajectory Rectifier (LTR) to align off-path LR latents, narrowing the gap and facilitating a more linear one-step flow.

3.3 Adaptive Timestep Estimator (ATE)

To estimate spatial degradation severity, the ATE generates a continuous, token-wise timestep map $\mathbf{T}_{\text{map}} \in [0, 1]$, spatially aligned with the resolution of the DiT latent tokens. To ensure robust estimation while avoiding cold-start instability, the ATE formulates the final prediction as a weighted fusion of a deterministic heuristic branch ($\mathbf{M}_{\text{H}}$) and a learnable neural branch ($\mathbf{M}_{\text{N}}$).

Heuristic Physical Prior. The heuristic branch serves as a parameter-free physical anchor. Operating on $\mathbf{x}_{\text{LR}}$, it computes a composite physical energy map $\mathbf{E}$ by linearly aggregating local texture variance, spatial structural gradients, and high-pass residuals. This energy is spatially smoothed and normalized:

$$\mathbf{M}_{\text{H}} = \tanh(\text{Smooth}(\mathbf{E}) \cdot \tau), \quad (5)$$

where $\tau = 4.0$ is an empirical scaling factor. In implementation, $\mathbf{E} = 0.8\mathbf{E}_{\text{std}} + 0.2\mathbf{E}_{\text{Sobel}} + 0.6\mathbf{E}_{\text{FFT}}$, where $\mathbf{E}_{\text{std}}$ uses a 7×7 local window and $\mathbf{E}_{\text{FFT}}$ uses the radial high-pass weighting $\sigma((r - 0.35) \cdot 40)$. We apply 9×9 average smoothing before the $\tanh(4\mathbf{E})$ normalization. This helps distinguish heavily degraded textures from smooth backgrounds.

Learnable Neural Diagnosis. Concurrently, the neural branch leverages a dual-path architecture. The latent path extracts high-level structural semantics $\mathbf{F}_{\text{latent}}$. The pixel-frequency path operates on $\mathbf{x}_{\text{LR}}$ via a 2D spatial Real-FFT, yielding the complex spectrum $\mathbf{X} = \mathcal{F}(\mathbf{x}_{\text{LR}})$. To effectively process structured phase patterns, we decouple $\mathbf{X}$ into real and imaginary components, processing them through a gating network Φ_{gate} to generate a real-valued mask $\mathbf{G}$:

$$\mathbf{G} = \Phi_{\text{gate}}([\mathbf{X}_{\text{real}}, \mathbf{X}_{\text{imag}}]). \quad (6)$$

We then modulate the complex spectrum and project it back to the spatial domain:

$$\mathbf{x}_{\text{gated}} = \mathcal{F}^{-1}(\mathbf{G} \odot \mathbf{X}). \quad (7)$$

This real-valued modulation dynamically scales frequency amplitudes while preserving the complex phase. After spatiotemporal alignment, the resulting feature $\mathbf{F}_{\text{rgb}}$ is fused with $\mathbf{F}_{\text{latent}}$ to generate the learnable map:

$$\mathbf{M}_{\text{N}} = \sigma\left(\Phi_{\text{fuse}}([\mathbf{F}_{\text{latent}}, \mathbf{F}_{\text{rgb}}])\right), \quad (8)$$

where σ denotes the Sigmoid activation.

Stable Integration. The ultimate token-wise timestep map is computed as a convex combination:

$$\mathbf{T}_{\text{map}} = \lambda \mathbf{M}_{\text{H}} + (1 - \lambda) \mathbf{M}_{\text{N}}. \quad (9)$$

We set $\lambda = 0.25$, which serves as a regularization term to stabilize trajectories during early training phases, while allowing the neural network to guide fine-grained allocation upon convergence.

3.4 Latent Trajectory Rectifier (LTR)

Guided by $\mathbf{T}_{\text{map}}$, the Latent Trajectory Rectifier (LTR) realigns the generative trajectory based on a frequency-aware decomposition within the frozen VAE latent space.

Given the initial LR latent $\mathbf{z}_{\text{LR}}$, a decomposition block explicitly decouples it into a low-frequency structural component $\mathbf{z}_{\text{struct}}$ and a high-frequency detail component $\mathbf{z}_{\text{detail}}$. The structural component predominantly retains correct spatial semantics, thus bypassing stochastic modulation. Conversely, the detail component $\mathbf{z}_{\text{detail}}$ is dynamically calibrated with standard Gaussian noise ϵ via an Adaptive Modulation module Φ_{mod} conditioned on $\mathbf{T}_{\text{map}}$:

$$\hat{\mathbf{z}}_{\text{detail}} = \Phi_{\text{mod}}(\mathbf{z}_{\text{detail}}, \epsilon, \mathbf{T}_{\text{map}}). \quad (10)$$

The preserved structure and modulated detail are reassembled via a Composition block and passed through a non-linear Trajectory Rectifier Φ_{rect} . A global residual connection from $\mathbf{z}_{\text{LR}}$ encourages identity mapping in artifact-free regions:

$$\mathbf{z}_{\text{rect}} = \mathbf{z}_{\text{LR}} + \Phi_{\text{rect}}(\mathbf{z}_{\text{struct}} + \hat{\mathbf{z}}_{\text{detail}}). \quad (11)$$

The rectified latent $\mathbf{z}_{\text{rect}}$, aligned closer to the reference trajectory, is tokenized and fed alongside $\mathbf{T}_{\text{map}}$ into the DiT blocks, enabling a more flow-compatible, artifact-reduced one-step generation.

Geometric Projection Constraint. To encourage the geometric rectification discussed in Sec. 3.2, we introduce a Geometric Projection Loss ($\mathcal{L}_{\text{LTR}}$). The goal is two-fold: minimize the orthogonal deviation of the LR latent and better align its longitudinal progression with the estimated temporal state $\mathbf{T}_{\text{map}}$. The projection is used as a VAE-latent directional regularizer rather than a claim that latent channels are physically orthogonal.

Formally, let $\mathbf{u} = (\epsilon - \mathbf{z}_{\text{HR}}) / \|\epsilon - \mathbf{z}_{\text{HR}}\|_2$ denote the unit direction vector representing the ideal generative path from the target HR latent to pure Gaussian noise ϵ . The overall path vector is $\mathbf{v}_{\text{path}} = \epsilon - \mathbf{z}_{\text{HR}}$. For any latent state $\mathbf{z}$, its longitudinal projection along this ideal path, relative to the HR anchor, is defined as $\pi(\mathbf{z}) = \langle \mathbf{z} - \mathbf{z}_{\text{HR}}, \mathbf{u} \rangle \mathbf{u}$. The geometric constraint comprises two orthogonal components:

$$\mathcal{L}_{\text{LTR}} = \underbrace{\|(\mathbf{z}_{\text{rect}} - \mathbf{z}_{\text{HR}}) - \pi(\mathbf{z}_{\text{rect}})\|_2^2}_{\text{Trajectory Deviation Loss}} + \lambda_{\text{align}} \underbrace{\|\pi(\mathbf{z}_{\text{rect}}) - (\mathbf{T}_{\text{map}} \odot \mathbf{v}_{\text{path}})\|_2^2}_{\text{Longitudinal Alignment}}. \quad (12)$$

The first term penalizes the orthogonal distance from the ideal trajectory, pulling $\mathbf{z}_{\text{rect}}$ towards the reference path (Y-axis correction). The second term encourages *Longitudinal Alignment* (X-axis correction). Because the local Euler approximation integrates from $\mathbf{T}_{\text{map}}$, the initial state should reside near the corresponding proportional distance along the generative vector. This term aligns the progression state of $\mathbf{z}_{\text{rect}}$ with the spatial diagnostic map predicted by the ATE, supporting the trained spatially-conditioned local Euler approximation and reducing cyclic dependency during joint training.

Remark on Inference: Note that $\mathcal{L}_{\text{LTR}}$ and the requirement for $\mathbf{z}_{\text{HR}}$ are confined to the training phase to supervise the rectifier. During inference, the converged non-linear network Φ_{rect} directly predicts the aligned latent from the unknown real-world LR input without requiring any HR reference or computing any loss, introducing no additional inference overhead.

Total Objective. The overall training objective combines this geometric constraint with standard signal reconstruction and perceptual enhancements:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \alpha \mathcal{L}_{\text{perceptual}} + \beta \mathcal{L}_{\text{LTR}}, \quad (13)$$

where $\mathcal{L}_{\text{rec}}$ is the Mean Squared Error (MSE) loss that encourages spatiotemporal fidelity, $\mathcal{L}_{\text{perceptual}}$ leverages LPIPS to promote high-frequency perceptual quality, and α, β are weighting hyperparameters. This objective encourages geometric trajectory alignment alongside high-quality visual generation.

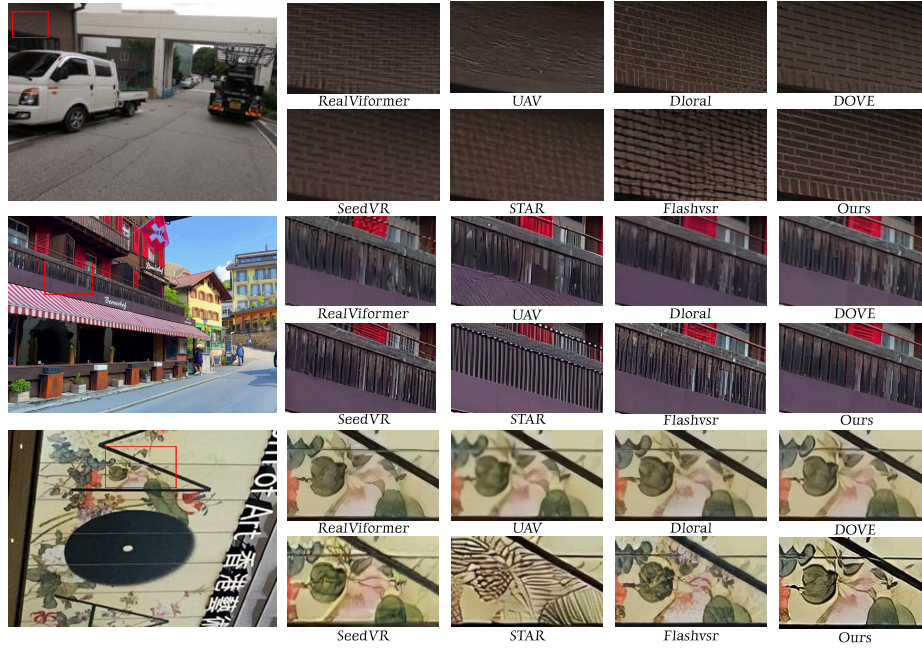


Fig. 5: Visual comparison of ART-VSR against state-of-the-art methods. By mitigating the initial state mismatch, our method restores high-frequency textures while reducing the structural hallucinations observed in uncalibrated generative baselines.

4 Experiments

4.1 Experimental Setup

Datasets. To comprehensively evaluate our method against complex real-world degradations, we conduct experiments on five widely adopted benchmarks: REDS [24], YouHQ40 [47], UDM10 [42], VideoLQ [4], and MVSR4x [33]. These datasets encompass a diverse range of spatial resolutions, motion dynamics, and in-the-wild degradation types (*e.g.*, sensor noise, motion blur, and compression artifacts).

Evaluation Metrics. Traditional pixel-wise metrics (PSNR and SSIM) inherently favor over-smoothed, statistical-average outputs and often fail to reflect human visual perception in generative enhancement tasks. Therefore, following recent literature [39], we primarily focus on perceptual evaluations. We employ LPIPS [45] and NIQE [21] to assess structural distortion, and reference-free metrics including MUSIQ [11], CLIPQA [30], and DOVER [35] to evaluate authentic perceptual quality. Furthermore, to evaluate temporal consistency and penalize flickering artifacts, we conduct temporal profile (X-t slice) analyses [37, 42].

Implementation Details. ART-VSR is initialized from the official Wan2.1 DiT checkpoint [29] and jointly fine-tuned with ATE/LTR on NVIDIA A800 GPUs.

Table 1: Quantitative comparison with State-of-the-Art methods across multiple benchmarks. Red and blue indicate the best and second-best performance, respectively. The $\uparrow$ denotes higher is better, while $\downarrow$ denotes lower is better. ART-VSR achieves highly competitive perceptual quality among one-step generative methods while maintaining competitive fidelity.

Dataset	Metric	UAV (CVPR'24)	RealVifformer (ECCV'24)	SeedVR (CVPR'25)	STAR (ICCV'25)	DloraL (NIPS'25)	DOVE (NIPS'25)	FlashVSR (CVPR'26)	ART-VSR (Ours)
UDM10	PSNR $\uparrow$	22.778	25.431	24.442	24.592	24.575	26.944	24.012	<u>26.302</u>
	SSIM $\uparrow$	0.608	0.710	0.738	0.709	0.709	0.771	0.685	<u>0.764</u>
	LPIPS $\downarrow$	0.411	0.348	0.282	0.401	0.331	<u>0.270</u>	0.294	0.242
	MUSIQ $\uparrow$	60.666	58.296	46.222	35.490	60.267	61.251	67.618	<u>62.741</u>
	CLIPQA $\uparrow$	0.457	0.412	0.288	0.239	0.577	0.450	<u>0.592</u>	0.623
	DOVER $\uparrow$	0.435	0.432	0.310	0.286	0.456	0.507	<u>0.540</u>	0.546
	NIQE $\downarrow$	4.155	3.921	5.981	6.198	4.309	5.331	3.715	<u>3.824</u>
MVSR4x	PSNR $\uparrow$	20.979	22.800	22.061	22.087	22.291	<u>22.518</u>	22.171	22.310
	SSIM $\uparrow$	0.715	0.731	<u>0.747</u>	0.726	0.744	0.753	0.734	0.742
	LPIPS $\downarrow$	0.469	0.478	0.379	0.446	0.363	0.347	<u>0.337</u>	0.325
	MUSIQ $\uparrow$	40.215	42.496	34.515	42.008	59.672	62.467	<u>65.776</u>	66.563
	CLIPQA $\uparrow$	0.266	0.457	0.270	0.314	<u>0.639</u>	0.515	0.595	0.711
	DOVER $\uparrow$	0.241	0.198	0.195	0.203	0.347	0.465	0.472	<u>0.465</u>
	NIQE $\downarrow$	8.032	7.661	7.100	7.440	<u>5.555</u>	7.073	4.888	5.812
YouHQ40	PSNR $\uparrow$	23.385	<u>26.304</u>	22.008	20.104	25.952	27.824	23.421	25.780
	SSIM $\uparrow$	0.644	<u>0.787</u>	0.639	0.557	0.723	0.813	0.650	0.775
	LPIPS $\downarrow$	0.285	0.204	0.223	0.388	0.226	0.165	0.246	<u>0.180</u>
	MUSIQ $\uparrow$	<u>70.849</u>	65.490	65.810	67.782	68.686	64.105	69.737	70.917
	CLIPQA $\uparrow$	0.630	0.451	0.483	0.557	0.645	0.456	0.590	<u>0.640</u>
	DOVER $\uparrow$	0.653	0.668	0.682	0.572	0.646	0.686	<u>0.708</u>	0.738
	NIQE $\downarrow$	3.004	3.735	3.517	4.484	3.881	4.237	3.138	<u>3.070</u>
REDS	PSNR $\uparrow$	15.592	26.682	23.673	22.522	24.767	<u>25.841</u>	21.422	23.948
	SSIM $\uparrow$	0.410	0.749	0.679	0.608	0.649	<u>0.719</u>	0.541	0.679
	LPIPS $\downarrow$	0.491	0.248	0.258	0.342	<u>0.230</u>	0.252	0.273	0.197
	MUSIQ $\uparrow$	<u>63.935</u>	61.681	60.072	54.947	62.993	53.463	63.051	64.965
	CLIPQA $\uparrow$	<u>0.501</u>	0.337	0.300	0.300	0.498	0.270	0.382	0.560
	DOVER $\uparrow$	0.357	0.361	0.377	<u>0.431</u>	0.314	0.353	0.373	0.442
	NIQE $\downarrow$	2.929	2.850	3.518	4.158	2.955	3.666	2.711	<u>2.806</u>
VideoLQ	MUSIQ $\uparrow$	52.292	51.776	39.058	37.332	60.530	44.793	54.759	<u>55.272</u>
	CLIPQA $\uparrow$	0.366	0.341	0.229	0.290	0.583	0.291	0.423	<u>0.501</u>
	DOVER $\uparrow$	0.412	0.384	0.365	0.407	0.461	0.465	<u>0.476</u>	0.489
	NIQE $\downarrow$	4.317	4.077	5.429	5.822	4.065	5.336	<u>3.985</u>	3.934

Training uses HQ-VSR/YouHQ videos and Flickr2K/DIV2K images, following the mixed image/video training and degradation setup used by DOVE/FlashVSR [5, 49]. To balance trajectory alignment and texture generation, the training pipeline is decoupled into two progressive stages.

Stage 1 (Latent Fine-Tuning): Optimization is confined to the latent space to establish a stable vector field and structural mapping. We train on 33-frame sequences at 640×960 resolution (batch size 32, learning rate 1×10^{-5}). The reconstruction loss $\mathcal{L}_{rec}$ measures latent MSE, LPIPS is disabled ($\alpha = 0$), and $\mathcal{L}_{LTR}$ is applied so the model focuses on geometric correction before RGB perceptual refinement.

Stage 2 (RGB Optimization): The generated latents are decoded by the frozen VAE decoder and optimized in RGB space, where $\mathcal{L}_{rec}$ computes pixel-wise MSE. Training shifts to 17-frame clips at 480×640 (batch size 4, learning

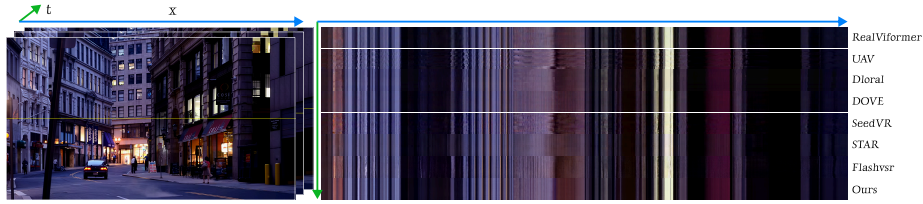


Fig. 6: Evaluation of temporal consistency via X-t temporal profiles. The X-t slices are extracted from a specific horizontal scanline over time. Compared to generative baselines that exhibit jagged and discontinuous trajectories (indicating noticeable flickering), ART-VSR produces continuous and sharp temporal edges, suggesting competitive cross-frame stability.

rate 6×10^{-5}). LPIPS is enabled ($\alpha = 1.0$) to encourage realistic high-frequency textures while preserving the trajectory alignment learned in Stage 1.

Throughout both stages, the overall Latent Trajectory Rectifier loss weight β is set to 0.5, and the longitudinal alignment weight λ_{align} within $\mathcal{L}_{LTR}$ is set to 0.1. For ATE, the heuristic scaling factor and neural-heuristic fusion weight are set to $\tau = 4.0$ and $\lambda = 0.25$, respectively; the heuristic map follows the texture/Sobel/FFT energy construction described in Sec. 3.3, including 9×9 smoothing and $\tanh(\tau E)$ normalization.

Protocol Notes and Sensitivity: All ART-VSR ablations use the same two-stage recipe and optimizer schedule; only the stated ATE/LTR component is changed. The constants τ , λ , β , and λ_{align} are selected by one-factor validation around the default setting. Removing the neural contribution to ATE, using only the longitudinal LTR term, or over-weighting LTR degrades the fidelity-consistency balance in our validation, so the same default configuration is used for all reported benchmarks.

4.2 Comparison with State-of-the-Art Methods

We compare ART-VSR against a comprehensive suite of recent state-of-the-art methods, including traditional regression networks (RealViformer [46]), and recent generative VSR models (UAV [47], DLoRAL [27], STAR [38], DOVE [5], FlashVSR [49], SeedVR [31, 32]). For fairness, all baselines in Tab. 1 are re-tested from their official released checkpoints using official inference settings under our unified evaluation protocol. The same test splits, video sequences, frame indices, and LR degradations are used across all methods. Compared with original reports, the re-tested baseline numbers differ in both directions, which is expected under a unified pipeline and does not indicate method-specific tuning.

Quantitative Results. As reported in Tab. 1, ART-VSR achieves highly competitive performance across multiple benchmarks. While regression-based methods like RealViformer occasionally score high on PSNR/SSIM, they lag behind in perceptual fidelity. Compared to recent trajectory-based one-step generative methods (*e.g.*, DOVE and FlashVSR), our approach yields strong LPIPS,

MUSIQ, and CLIPQA scores. These results support that the **state-time adaptation** paradigm mitigates trajectory drift caused by initial state mismatch (Sec. 1), enabling strong detail synthesis without sequential sampling.

Qualitative Results. Spatial visual comparisons are presented in Fig. 5. In severely degraded in-the-wild scenarios, baseline one-step generative models often produce structurally distorted hallucinations or fail to reconstruct fine-grained textures. This behavior is consistent with generation from an off-path initial state, which yields a distorted vector field (the baseline flow in Fig. 3a). In contrast, by pulling the latent representation closer to the reference generative path, ART-VSR restores crisp details (*e.g.*, sharp text contours and complex building facades) while maintaining structural fidelity to the original video context.

Temporal Consistency. Beyond spatial fidelity, maintaining temporal coherence is a challenge for generative VSR. To investigate this, we extract temporal profiles (X-t slices) by sampling a fixed spatial scanline across the temporal axis, as visualized in Fig. 6. Baseline generative models can exhibit temporal flickering, which manifests as jagged or disjointed trajectories in the X-t domain. Regression models (*e.g.*, RealVifformer) maintain structural continuity but yield overly blurred temporal edges. In contrast, ART-VSR achieves smooth, continuous, and sharp temporal edges, suggesting a favorable fidelity-consistency trade-off.

Degradation Heterogeneity. The trajectory drift discussed in Sec. 1 can vary with degradation type and severity, but the key difficulty for one-step VSR is often local: textured and flat regions, moving and stationary areas, or compressed and clean patches can coexist in the same frame. ART-VSR therefore does not require a manually defined degradation taxonomy. Instead, ATE estimates token-wise integration cues, and LTR aligns the initial latent state conditioned on those cues, allowing the one-step update to adapt locally while retaining a unified inference pipeline.

4.3 Ablation Studies

To validate the core components of our proposed framework, we conduct controlled ablation studies on five benchmarks. Table 1 compares complete external systems, whereas Tables 2 and 3 isolate components inside ART-VSR. All ablation variants share the same data, degradation, frame setting, optimizer, and schedule.

Effectiveness of Adaptive Timestep Estimator (ATE). To address the global timestep constraint, ATE generates token-wise continuous t -maps ($\mathbf{T}_{map}$) rather than relying on a globally fixed temporal anchor. In Tab. 2, the scalar- T_L rows replace only ATE’s adaptive $\mathbf{T}_{map}$ with a broadcast global T_L inside ART-VSR, while keeping the remaining pipeline, including LTR and $\mathcal{L}_{LTR}$. Evaluating this controlled variant across global timesteps ($T_L \in [0.0, 1.0]$) highlights the trade-off in Fig. 3b: small timesteps (*e.g.*, $T_L = 0.2$) preserve smooth background structures but under-enhance textures, whereas large timesteps (*e.g.*, $T_L = 0.8$) may hallucinate details but introduce noise artifacts in clean regions. By estimating spatial heterogeneity, ATE dynamically allocates local integration lengths.

Table 2: Ablation study on the Adaptive Timestep Estimator (ATE). We evaluate controlled ART-VSR variants using varying globally fixed timesteps ($T_L \in [0.0, 1.0]$) versus token-wise dynamic timesteps. The fusion of heuristic ($\mathbf{M}_H$) and neural ($\mathbf{M}_N$) branches yields a strong balance between detail synthesis and structural preservation.

Dataset	Metric	Globally Fixed Timestep (T_L)						Adaptive Timestep (ATE)		
		0.0	0.2	0.4	0.6	0.8	1.0	Only $\mathbf{M}_H$	Only $\mathbf{M}_N$	Full (Ours)
REDS	LPIPS ↓	0.253	0.251	0.231	0.264	0.273	0.261	0.205	0.201	0.197
	MUSIQ ↑	41.532	43.058	42.595	41.391	41.971	40.081	61.684	62.911	64.965
	NIQE ↓	5.039	4.823	5.154	5.297	5.351	5.093	3.259	2.886	2.806
YouHQ40	LPIPS ↓	0.202	0.196	0.185	0.191	0.184	0.231	0.185	0.182	0.180
	MUSIQ ↑	48.173	49.236	49.674	50.128	48.516	47.363	66.487	65.640	70.917
	NIQE ↓	5.288	5.251	5.407	5.523	5.651	5.385	3.644	4.117	3.070
UDM10	LPIPS ↓	0.353	0.315	0.302	0.385	0.389	0.357	0.268	0.284	0.242
	MUSIQ ↑	27.486	28.131	28.237	29.111	28.563	30.944	54.534	43.987	62.741
	NIQE ↓	6.207	6.211	6.016	5.989	6.094	5.904	4.632	5.252	3.824
MVSR4x	LPIPS ↓	0.341	0.335	0.348	0.358	0.355	0.358	0.351	0.343	0.325
	MUSIQ ↑	28.163	29.236	28.214	28.837	30.132	29.861	53.407	45.555	66.563
	NIQE ↓	7.615	7.551	7.843	7.920	7.646	7.610	5.914	6.448	5.812
VideoLQ	CLIPQA ↑	0.291	0.277	0.286	0.291	0.298	0.313	0.402	0.401	0.501
	MUSIQ ↑	29.582	30.577	30.872	27.624	31.063	32.266	53.984	51.084	55.272
	NIQE ↓	7.615	7.064	6.143	6.973	5.805	5.562	3.966	4.394	3.934

Table 3: Ablation study on the Latent Trajectory Rectifier (LTR) and Geometric Projection Loss ($\mathcal{L}_{LTR}$). Discarding the LTR weakens trajectory alignment, while removing $\mathcal{L}_{LTR}$ can shift the rectified state along the temporal axis and conflict with the local integration cue.

Method Setup	REDS			YouHQ40			UDM10			MVSR4x			VideoLQ	
	LPIPS↓	MUSIQ↑	NIQE↓	LPIPS↓	MUSIQ↑	NIQE↓	LPIPS↓	MUSIQ↑	NIQE↓	LPIPS↓	MUSIQ↑	NIQE↓	MUSIQ↑	NIQE↓
w/o LTR	0.264	46.437	5.856	0.235	42.548	5.706	0.371	27.047	6.809	0.364	34.176	7.874	23.249	7.243
w/ LTR (w/o $\mathcal{L}_{LTR}$)	0.212	63.497	3.972	0.208	63.467	4.637	0.356	59.527	6.313	0.346	48.227	5.836	43.397	4.114
w/ LTR (Ours)	0.197	64.965	2.806	0.180	70.917	3.070	0.242	62.741	3.824	0.325	66.563	5.812	55.272	3.934

The heuristic branch ($\mathbf{M}_H$, scaled by $\tau = 4.0$) provides a texture-variance anchor, while the neural branch ($\mathbf{M}_N$) captures complex semantic degradation. The full ATE, which fuses both branches ($\lambda = 0.25$), provides the strongest balance in our study, improving LPIPS (0.180 on YouHQ40) and MUSIQ (70.917) compared with the globally fixed timestep variants.

Effectiveness of Latent Trajectory Rectifier (LTR). LTR is designed to bridge the geometric trajectory gap by aligning the initial LR latent. Table 3 keeps ATE’s adaptive $\mathbf{T}_{map}$ and ablates LTR in two ways: removing the LTR branch, or keeping the LTR structure but removing $\mathcal{L}_{LTR}$. Entirely removing LTR (“w/o LTR”) leaves the model to transition directly from off-path inputs, causing clear drops across metrics (*e.g.*, 18.528 MUSIQ on REDS). This supports the initial state mismatch analysis visualized in Fig. 3a. Furthermore, without the Geometric Projection Loss (“w/ LTR (w/o $\mathcal{L}_{LTR}$)”), LTR becomes

an unconstrained latent filter, whose longitudinal shift can conflict with the local integration cue estimated by ATE ($\mathbf{T}_{map}$). The full model couples ATE, LTR, and DiT: ATE estimates the local integration cue, LTR uses it to obtain $\mathbf{z}_{rect}$, and DiT uses both for one-step generation. Constraining LTR with $\mathcal{L}_{LTR}$ ($\beta = 0.5, \lambda_{align} = 0.1$) encourages orthogonal trajectory correction and longitudinal alignment, yielding the strongest controlled variant.

5 Conclusion

In this paper, we analyzed *trajectory drift* as an important bottleneck in current flow-based one-step VSR methods. To address **initial state mismatch** and the **global timestep constraint** under complex real-world degradations, we proposed **ART-VSR** under a **state-time adaptation** paradigm. By introducing an Adaptive Timestep Estimator (ATE) for fine-grained spatial modulation and a Latent Trajectory Rectifier (LTR) with geometric constraints, our method moves off-path LR inputs closer to a reference generative trajectory while assigning locally adaptive integration lengths. Extensive experiments show that ART-VSR effectively mitigates trajectory drift and achieves competitive perceptual quality while maintaining one-step efficiency. We hope this state-time perspective provides a practical direction for robust, high-fidelity video enhancement in complex real-world scenarios.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China under grant No.62372091 and in part by Hainan Province Key R&D Program under grant No.ZDYF2024(LALH)001.

References

1. Bai, H., Chen, X., Yang, C., He, Z., Deng, S., Chen, Y.: Vivid-VR: Distilling concepts from text-to-video diffusion transformer for photorealistic video restoration. In: ICLR (2026)
2. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: BasicVSR: The search for essential components in video super-resolution and beyond. In: CVPR. pp. 4947–4956 (2021)
3. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In: CVPR. pp. 5972–5981 (2022)
4. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Investigating tradeoffs in real-world video super-resolution. In: CVPR. pp. 5962–5971 (2022)
5. Chen, Z., Zou, Z., Zhang, K., Su, X., Yuan, X., Guo, Y., Zhang, Y.: DOVE: Efficient one-step diffusion model for real-world video super-resolution. In: NeurIPS (2025)
6. Chu, M., Xie, Y., Mayer, J., Leal-Taixé, L., Thurey, N.: Learning temporal coherence via self-supervision for GAN-based video generation. ACM TOG **39**(4), 75–1 (2020)

7. Haris, M., Shakhnarovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: CVPR. pp. 3897–3906 (2019)
8. He, J., Xue, T., Liu, D., Lin, X., Gao, P., Lin, D., Qiao, Y., Ouyang, W., Liu, Z.: VEnhancer: Generative space-time enhancement for video generation. arXiv preprint arXiv:2407.07667 (2024)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. vol. 33, pp. 6840–6851 (2020)
10. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: CogVideo: Large-scale pretraining for text-to-video generation via transformers. In: ICLR (2023)
11. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: MusiQ: Multi-scale image quality transformer. In: ICCV. pp. 5148–5157 (2021)
12. Li, J., Cao, J., Guo, Y., Li, W., Zhang, Y.: One diffusion step to real-world super-resolution via flow trajectory distillation. In: ICML. pp. 34044–34053 (2025)
13. Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L.: VRT: A video restoration transformer. IEEE TIP **33**, 2171–2182 (2024)
14. Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Gool, L.V.: Recurrent video restoration transformer with guided deformable attention. In: NeurIPS. vol. 35, pp. 378–393 (2022)
15. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: DiffBIR: Toward blind image restoration with generative diffusion prior. In: ECCV. pp. 430–448. Springer (2024)
16. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
17. Liu, D., Wen, B., Fan, Y., Loy, C.C., Huang, T.S.: Non-local recurrent network for image restoration. In: NeurIPS. vol. 31, pp. 1673–1682 (2018)
18. Liu, X., Zhang, X., Ma, J., Peng, J., et al.: InstaFlow: One step is enough for high-quality diffusion-based text-to-image generation. In: ICLR (2024)
19. Lucas, A., Lopez-Tapia, S., Molina, R., Katsaggelos, A.K.: Generative adversarial networks and perceptual losses for video super-resolution. IEEE TIP **28**(7), 3312–3327 (2019)
20. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
21. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. IEEE Sign. Process. Letters **20**(3), 209–212 (2012)
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
23. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: ECCV. pp. 87–103. Springer (2024)
24. Son, S., Lee, S., Nah, S., Timofte, R., Lee, K.M.: NTIRE 2021 challenge on video super-resolution. In: CVPRW. pp. 166–181 (2021)
25. Song, Y., Dhariwal, P.: Improved techniques for training consistency models. In: ICLR (2024)
26. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: NeurIPS. vol. 32, pp. 11895–11907 (2019)
27. Sun, Y., Sun, L., Liu, S., Wu, R., Zhang, Z., Zhang, L.: One-step diffusion for detail-rich and temporally consistent video super-resolution. In: NeurIPS (2025)
28. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: TDAN: Temporally-deformable alignment network for video super-resolution. In: CVPR. pp. 3360–3369 (2020)
29. Wan Team, Wang, A., Ai, B., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)

30. Wang, J., Chan, K.C., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: AAAI. vol. 37, pp. 2555–2563 (2023)
31. Wang, J., Lin, S., Lin, Z., Ren, Y., Wei, M., Yue, Z., Zhou, S., Chen, H., Zhao, Y., Yang, C., Xiao, X., Loy, C.C., Jiang, L.: SeedVR2: One-step video restoration via diffusion adversarial post-training. In: ICLR (2026)
32. Wang, J., Lin, Z., Wei, M., Zhao, Y., Yang, C., Loy, C.C., Jiang, L.: SeedVR: Seeding infinity in diffusion transformer towards generic video restoration. In: CVPR (2025)
33. Wang, R., Liu, X., Zhang, Z., Wu, X., Feng, C.M., Zhang, L., Zuo, W.: Benchmark dataset and effective inter-frame alignment for real-world video super-resolution. In: CVPR. pp. 1168–1177 (2023)
34. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: EDVR: Video restoration with enhanced deformable convolutional networks. In: CVPRW. pp. 247–256 (2019)
35. Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., Lin, W.: Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In: ICCV. pp. 20144–20154 (2023)
36. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. In: NeurIPS. vol. 37, pp. 92529–92553 (2024)
37. Xiao, Z., Xiong, Z., Fu, X., Liu, D., Zha, Z.J.: Space-time video super-resolution using temporal profiles. In: ACM MM. pp. 664–672 (2020)
38. Xie, R., Liu, Y., Zhou, P., Zhao, C., Zhou, J., Zhang, K., Zhang, Z., Yang, J., Yang, Z., Tai, Y.: Star: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. In: ICCV. pp. 17108–17118 (2025)
39. Xu, Y., Park, T., Zhang, R., Zhou, Y., Shechtman, E., Liu, F., Huang, J.B., Liu, D.: VideoGigaGAN: Towards detail-rich video super-resolution. In: CVPR. pp. 2139–2149 (2024)
40. Yang, X., He, C., Ma, J., Zhang, L.: Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In: ECCV. pp. 224–242. Springer (2024)
41. Yang, Z., Teng, J., Zheng, W., et al.: CogVideoX: Text-to-video diffusion models with an expert transformer. In: ICLR (2025)
42. Yi, P., Wang, Z., Jiang, K., Jiang, J., Ma, J.: Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In: ICCV. pp. 3106–3115 (2019)
43. Yue, Z., Wang, J., Loy, C.C.: ResShift: Efficient diffusion model for image super-resolution by residual shifting. In: NeurIPS. vol. 36, pp. 13294–13307 (2023)
44. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
46. Zhang, Y., Yao, A.: RealViformer: Investigating attention for real-world video super-resolution. In: ECCV. pp. 412–428. Springer (2024)
47. Zhou, S., Yang, P., Wang, J., Luo, Y., Loy, C.C.: Upscale-A-Video: Temporal-consistent diffusion model for real-world video super-resolution. In: CVPR. pp. 2535–2545 (2024)
48. Zhou, X., Zhang, L., Zhao, X., Wang, K., Li, L., Gu, S.: Video super-resolution transformer with masked inter&intra-frame attention. In: CVPR. pp. 25399–25408 (2024)
49. Zhuang, J., Guo, S., Cai, X., Li, X., Liu, Y., Yuan, C., Xue, T.: FlashVSR: Towards real-time diffusion-based streaming video super-resolution. arXiv preprint arXiv:2510.12747 (2025)