TEMPORAL RE-DISTILLATION: A POST-DISTILLATION STABILIZER FOR ONE-STEP VIDEO SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

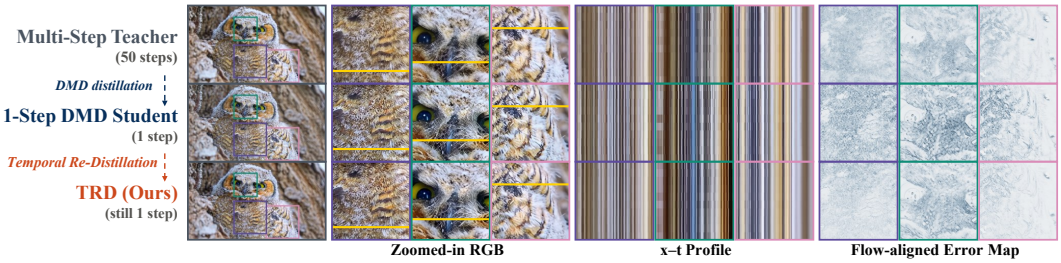

Figure 1: **Temporal consistency comparison.** For each method, we show a zoomed-in RGB patch, the x–t profile along the marked scanline, and the flow-aligned error map obtained by warping all frames to a reference frame and aggregating the residuals. Jagged x–t slices indicate short-term flickering (Bar-Tal et al., 2024), and bright regions in the error maps indicate long-term appearance drift (Thimonier et al., 2021).

ABSTRACT

One-step video super-resolution (VSR) models distilled via Distribution Matching Distillation (DMD) can preserve the frame-wise perceptual quality of multi-step diffusion teachers, yet often exhibit pronounced temporal artifacts, including flickering and texture drift. To understand this discrepancy, we trace temporal refinement along the teacher's denoising trajectory and observe that fine-grained cross-frame consistency is refined predominantly at low noise levels, a regime largely bypassed by one-step distillation. Motivated by this observation, we propose TRD-VSR, a temporal refinement framework that re-perturbs student predictions to low-noise states and distills the teacher's local denoising corrections in this regime. We further introduce Relational Structure Alignment (RSA) to explicitly transfer cross-frame relational structure, complementing the local teacher corrections with an explicit constraint on temporal correspondence. TRD-VSR is used only during training and therefore preserves single-step inference with no additional test-time cost. Experiments on three benchmarks show that TRD-VSR improves both temporal consistency and perceptual quality, achieving the best tLPIPS across all datasets, the best tOF and tLP on two datasets, and the best CLIPIQA and DOVER across all datasets, yielding a favorable perception–consistency trade-off over existing one-step generative VSR methods.

1 INTRODUCTION

Video super-resolution (VSR) aims to recover high-resolution (HR) videos from low-resolution (LR) sequences while synthesizing spatial details and maintaining temporal consistency (Caballero et al., 2017; Chan et al., 2021; Fuoli et al., 2023; Wronski et al., 2019). Recent diffusion-based VSR methods (Zhou et al., 2024; Xie et al., 2025; Wang et al., 2025; Lv et al., 2026; Yu et al., 2026) leverage the strong generative and spatiotemporal priors of pretrained video models such as CogVideoX (Yang et al., 2025), HunyuanVideo (Kong et al., 2024), and Wan2.1 (Team Wan et al., 2025), achieving superior perceptual quality over conventional regression-based approaches (Chan et al., 2022; Liang

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.

et al., 2024; 2022). However, their iterative multi-step sampling incurs high inference costs. Recent studies (Chen et al., 2025; Zhuang et al., 2026; Lv et al., 2026) therefore employ distillation techniques such as Distribution Matching Distillation (DMD) (Yin et al., 2024b) to reduce denoising to few- or single-step inference. Such aggressive step reduction, however, often compromises temporal consistency, causing flickering and temporal drift, as shown in Fig. 1.

Existing methods address this issue by either disentangling spatial synthesis and temporal alignment into separate network branches or enforcing explicit temporal constraints through pixel or feature-level supervision (Yang et al., 2024; Sun et al., 2025). The former relies on the assumption that spatial and temporal objectives are independent and free of feature interference, whereas the latter imposes a rigid deterministic constraint on a stochastic generative process. Both approaches inevitably create a trade-off: prioritizing temporal alignment compromises perceptual quality, such as over-smoothed textures (Dai et al., 2022).

In this work, we instead examine how spatiotemporal interactions evolve along the teacher's multi-step denoising trajectory. Prior work shows that cross-frame attention in video diffusion transformers encodes temporal correspondences that strengthen during denoising (Nam et al., 2025); we extend this analysis to one-step DMD distillation for VSR to reveal what is lost when the trajectory collapses. We find that temporal consistency in multi-step models is associated not merely with the *locality* of spatiotemporal attention, but more critically with how this locality progressively emerges. As denoising proceeds, the teacher's attention increasingly concentrates on nearby spatial regions and cross-frame correspondences, and this tightening accelerates in the low-noise regime, precisely where temporal consistency is largely recovered. Disrupting attention refinement in this regime directly degrades temporal consistency, identifying it as critical for temporal recovery. In contrast, the one-step DMD student exhibits substantially more diffuse attention without such progressive refinement. Motivated by this discrepancy, we revisit the DMD objective and empirically show that it is largely insensitive to temporal incoherence.

Based on these observations, we propose TRD-VSR, a Temporal Re-Distillation (TRD) method that stabilizes a distilled one-step VSR model. It consists of two complementary components: Low-Noise Correction Distillation (LNCD) and Relational Structure Alignment (RSA). LNCD reintroduces teacher supervision in the low-noise regime by distilling teacher-corrected predictions and intermediate features, allowing the student to recover late-stage temporal refinement. RSA aligns the student's cross-frame relational structure with that of the HR sequence; by supervising inter-frame relations rather than feature magnitudes, it improves temporal consistency while preserving spatial details.

Our contributions are summarized as follows:

1. We identify progressive spatiotemporal attention concentration, particularly its accelerated tightening in the low-noise regime, as a key factor underlying temporal recovery in video super-resolution.
2. We propose TRD-VSR, combining LNCD and RSA to restore temporal consistency through low-noise re-distillation and cross-frame relational supervision, with no additional inference cost.
3. Extensive experiments demonstrate that TRD-VSR consistently improves temporal consistency over the initial one-step student across all three benchmarks while maintaining competitive perceptual quality. It also achieves the best tLPIPS and DOVER scores among the compared methods on all benchmarks.

2 Related Work

2.1 Generative Video Super-Resolution

VSR aims to recover high-quality and temporally coherent videos from degraded observations. Conventional methods exploit inter-frame redundancy through recurrent propagation or spatiotemporal modeling (Wang et al., 2019; Chan et al., 2021; 2022; Liang et al., 2024; 2022; Zhang & Yao, 2024), but are typically optimized for pixel-wise fidelity and often produce over-smoothed details under complex degradations. Recent generative approaches alleviate this limitation by leveraging diffusion priors. MGLD-VSR (Yang et al., 2024) and Upscale-A-Video (Zhou et al., 2024) adapt pretrained image diffusion models with motion guidance or temporal modules, while STAR (Xie

et al., 2025) and SeedVR (Wang et al., 2025) further exploit the stronger temporal priors of pre-trained video diffusion models.

The high inference cost of iterative diffusion has motivated one-step generative VSR. DOVE (Chen et al., 2025) distills a video diffusion model into a one-step restorer, while SeedVR2 (Wang et al., 2026), UltraVSR (Liu et al., 2025), and DLoRAL (Sun et al., 2025) balance perceptual quality and temporal consistency via adversarial post-training, explicit temporal modeling, or disentangled learning. FlashVSR (Zhuang et al., 2026) and InfVSR (Zhang et al., 2026b) extend one-step diffusion to streaming restoration, and PS-SR (Wu et al., 2026) reaches single-step inference via speculative diffusion. Whereas these works mainly introduce temporal modeling through architecture or propagation, we treat temporal inconsistency as a consequence of aggressive distillation itself: the motion-dependent temporal structure of a multi-step teacher is weakened when compressed into a one-step VSR model.

2.2 Few-Step Video Generation and Diffusion Distillation

Reducing diffusion sampling cost has also been widely studied in video generation. VideoLCM (Wang et al., 2023c), AnimateDiff-Lightning (Lin & Yang, 2024), and T2V-Turbo (Li et al., 2024) extend consistency-based distillation (Song et al., 2023; Luo et al., 2023) to few-step video synthesis. Another line of work builds on DMD (Yin et al., 2024b) and DMD2 (Yin et al., 2024a), which train fast generators by matching the teacher-induced distribution rather than its full denoising trajectory. CausVid (Yin et al., 2025), Self-Forcing (Huang et al., 2025), and One-Forcing (Feng et al., 2026) further adapt this paradigm to efficient causal video generation.

Recent work has begun to expose the limitations of aggressive video distillation. rCM (Zheng et al., 2026) improves the quality–diversity trade-off by combining forward- and reverse-divergence objectives, while Phased DMD (Fan et al., 2026) addresses degraded complex motion by distributing matching across multiple noise intervals. Most closely related, CoDMD (Zhang et al., 2026a) introduces relational regularization to preserve cross-frame geometry in few-step text-to-video generation. In contrast, we focus on conditional video restoration and show that a DMD-distilled VSR model can retain strong spatial generative quality while insufficiently transferring the teacher's motion-conditioned temporal structure. We therefore introduce a dedicated temporal re-distillation stage that restores this structure without altering the one-step inference process.

3 Analysis

3.1 Preliminaries

Given an LR video, VSR aims to reconstruct its HR counterpart. We formulate VSR in the latent space of the video diffusion model, where the HR video and the upsampled LR video are encoded into latent variables $\boldsymbol{z}_{\mathrm{HR}}$ and $\boldsymbol{z}_{\mathrm{LR}}$, respectively.

Multi-step teacher Φ_{tea}. We first define a linear interpolation operator for adding noise to a latent

$$\mathcal{I}_\sigma(\boldsymbol{z}, \boldsymbol{\epsilon}) := (1-\sigma)\boldsymbol{z} + \sigma\boldsymbol{\epsilon}, \qquad \sigma \in [0,1]. \tag{1}$$

Given an HR latent $\boldsymbol{z}_{\mathrm{HR}}$, the noisy latent along the Flow-Matching path is constructed as

$$\boldsymbol{z}_\sigma = \mathcal{I}_\sigma(\boldsymbol{z}_{\mathrm{HR}}, \boldsymbol{\epsilon}), \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{2}$$

Conditioned on the LR latent $\boldsymbol{z}_{\mathrm{LR}}$, the multi-step teacher predicts the velocity field and is fine-tuned with the Flow-Matching objective (Lipman et al., 2023; Liu et al., 2023):

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\sigma,\boldsymbol{\epsilon}}\left[\left\|\Phi_{\mathrm{tea}}(\boldsymbol{z}_\sigma, \sigma, \boldsymbol{z}_{\mathrm{LR}}) - (\boldsymbol{\epsilon} - \boldsymbol{z}_{\mathrm{HR}})\right\|_2^2\right]. \tag{3}$$

Here, Φ_{tea} is initialized from a pretrained T2V DiT (Peebles & Xie, 2023).

One-step student G_{dmd}. DMD aims to align the distribution of the one-step student G_{dmd} with that of its multi-step teacher Φ_{tea}. Specifically, the student model generates an HR latent prediction

$\boldsymbol{z}_0^{\text{dmd}} = G_{\text{dmd}}(\boldsymbol{\epsilon}, \boldsymbol{z}_{\text{LR}})$. We obtain the noisy latent via $\boldsymbol{z}_\sigma = \mathcal{I}_\sigma(\boldsymbol{z}_0^{\text{dmd}}, \boldsymbol{\eta})$, where $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Given the noisy latent $\boldsymbol{z}_\sigma$, the DMD gradient is given by

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = -\mathbb{E}_{\sigma,\boldsymbol{\epsilon},\boldsymbol{\eta}} \left[(s_{\text{real}}(\boldsymbol{z}_\sigma, \sigma) - s_{\text{fake}}(\boldsymbol{z}_\sigma, \sigma)) \frac{\partial G_{\text{dmd}}(\boldsymbol{\epsilon}, \boldsymbol{z}_{\text{LR}}; \theta)}{\partial \theta} \right], \tag{4}$$

where s_{real} and s_{fake} denote the score estimates from the frozen teacher Φ_{tea} and the auxiliary fake score network trained on their respective distributions. The student G_{dmd} is optimized together with reconstruction and adversarial losses:

$$\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{DMD}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}, \tag{5}$$

where $\mathcal{L}_{\text{rec}}$ denotes pixel reconstruction loss and $\mathcal{L}_{\text{adv}}$ denotes the GAN loss on real HR videos.

3.2 Progressive Temporal Binding along the Teacher Denoising Trajectory

Although DMD substantially accelerates inference, we observe a marked degradation in temporal stability compared to the multi-step teacher. As shown in Fig. 1, the student accumulates significantly higher temporal errors, particularly around intricate texture details. To understand this, we first examine how spatiotemporal token bindings evolve along the teacher's multi-step trajectory.

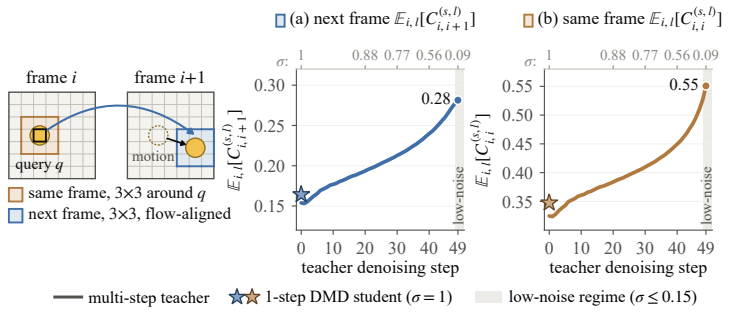

Figure 2: **Attention locality along the teacher's denoising trajectory.** Fraction of attention mass within the 3×3 token neighborhood of each query in its own frame (same frame) and around its flow-warped location in the next frame (cross-frame), averaged over all attention layers and video frames.

Where a token looks, and how it changes while denoising. Similar to Nam et al. (2025), we read out the DiT self-attention and arrange the latent tokens on their spatiotemporal grid. For a query token q in latent frame i at grid position $\boldsymbol{x}_i^q$, let $\bar{\boldsymbol{u}}(\boldsymbol{x}_i^q)$ denote the optical flow estimated by RAFT (Teed & Deng, 2020). The corresponding position in frame $i{+}1$ is:

$$\tilde{\boldsymbol{x}}_{i+1}^q = \boldsymbol{x}_i^q + \bar{\boldsymbol{u}}(\boldsymbol{x}_i^q) \tag{6}$$

To characterize attention concentration, we restrict attention to the target frame, renormalize it, and compute the proportion of attention mass within an r-radius neighborhood:

$$C_{i,i}^{(s,l)} = \mathbb{E}_q \sum_{\|\boldsymbol{x}_i^p - \boldsymbol{x}_i^q\|_\infty \leq r} \hat{A}_{qp}^{(s,l)}, \qquad C_{i,i+1}^{(s,l)} = \mathbb{E}_q \sum_{\|\boldsymbol{x}_{i+1}^p - \tilde{\boldsymbol{x}}_{i+1}^q\|_\infty \leq r} \hat{A}_{qp}^{(s,l)}, \tag{7}$$

where $\hat{\boldsymbol{A}}^{(s,l)}$ denotes the head-averaged attention at denoising step s and layer l, renormalized within the target frame, and $\mathbb{E}_q$ averages over queries in motion regions. $C_{i,i}^{(s,l)}$ and $C_{i,i+1}^{(s,l)}$ measure attention concentration around same-frame neighbors and motion-aligned cross-frame neighbors, respectively.

Fig. 2 reports these statistics at each denoising step, averaged over all frames and blocks. We observe a consistent trend: as denoising progresses, spatiotemporal interactions increasingly localize around true motion paths, with this concentration accelerating in the low-noise regime. This suggests that the resulting tight binding of tokens along motion paths acts as an implicit anchor, preventing adjacent frames from independently generating temporally shifted textures. In contrast, the one-step

DMD student exhibits a diffuse attention pattern akin to the teacher's early denoising stages, failing to capture the progressive token localization through iterative denoising.

To verify the connection between localized attention and temporal stability, we mask these neighboring token interactions during inference. Blocking them—particularly in the low-noise regime—severely impairs temporal consistency, while per-frame visual quality remains largely preserved (Appendix B.2). This finding identifies the late, low-noise stage of the denoising trajectory as critical for temporal stability; in this regime, spatiotemporal bindings tighten most intensely, such that any disruption impairs temporal consistency.

3.3 Why Does Temporal Stability Degrade after DMD

Table 1: **DMD is insensitive to temporal incoherence.** We generate two HR videos from the same LR input with distinct noise seeds, and construct a flickering sample by alternating their frames. This substantially degrades temporal consistency while leaving per-frame quality and DMD loss nearly unchanged. Results are averaged over varying noise levels and 10 video samples.

	Temporal		*Fidelity*		
Video	$E_{\mathbf{warp}}\downarrow$	**tLPIPS**$\downarrow$	**PSNR**$\uparrow$	**LPIPS**$\downarrow$	**DMD loss**
V_1	0.962	0.108	21.19	0.408	0.207 ± 0.052
V_2	0.969	0.108	21.27	0.406	0.206 ± 0.044
$\mathrm{V}_{\text{flicker}}$	1.297	0.159	21.27	0.405	0.199 ± 0.033

The preceding analysis shows that the multi-step teacher Φ_{tea} maintains temporal coherence by progressively localizing spatiotemporal interactions during denoising. We next investigate why this coherence is not preserved by DMD distillation.

To test whether DMD penalizes temporal inconsistency, we construct a controlled flickering video $\mathrm{V}_{\text{flicker}}$ by temporally interleaving two independent student outputs. Given a shared LR input $\boldsymbol{z}_{\text{LR}}$, we generate $\mathrm{V}_k = G_{\text{dmd}}(\boldsymbol{\epsilon}^{(k)}, \boldsymbol{z}_{\text{LR}})$, $k \in \{1, 2\}$, using distinct noise seeds, and then alternate their frames at adjacent temporal indices to synthesize $\mathrm{V}_{\text{flicker}}$. This setup preserves the per-frame marginal distributions while introducing severe frame-to-frame texture jitter. We then feed $\mathrm{V}_{\text{flicker}}$ into the frozen fake and real score models to compute the loss $\mathcal{L}_{\text{DMD}}$ across multiple noise levels σ and report averages over noise levels and videos. Note that we only probe $\mathcal{L}_{\text{DMD}}$, as the reconstruction and adversarial losses in Eq. 5 impose no effective temporal constraints (Lai et al., 2018; Chu et al., 2020).

As shown in Table 1, although $\mathrm{V}_{\text{flicker}}$ exhibits severe flickering, with substantial degradation in both E_{warp} and tLPIPS, the DMD loss $\mathcal{L}_{\text{DMD}}$ remains nearly unchanged. This suggests that the DMD objective optimizes visual quality primarily on a frame-by-frame basis while providing limited supervision for cross-frame fluctuations, leaving one-step students prone to temporal flickering.

4 Method: Temporal Re-Distillation for One-Step VSR

To improve the temporal stability of the one-step model, we propose TRD-VSR based on the above analysis. Fig. 3 presents the overall training framework, consisting of two complementary components: Low-Noise Correction Distillation (LNCD) and Relational Structure Alignment (RSA).

4.1 Low-Noise Correction Distillation

After DMD distillation, we perform a post-distillation stage in the low-noise regime to recover temporal stability. Since the analysis in Sec. 3 identifies the low-noise regime as critical for temporal stability, we propose **Low-Noise Correction Distillation (LNCD)** to reintroduce the teacher's low-noise supervision into the one-step student. As shown in Fig. 3, we freeze the DMD student G_{dmd} and initialize a trainable generator G_{trd} from its weights. The key idea is to locally correct the DMD prediction using the frozen teacher Φ_{tea} after injecting a small amount of noise, and then distill the corrected endpoint into G_{trd}.

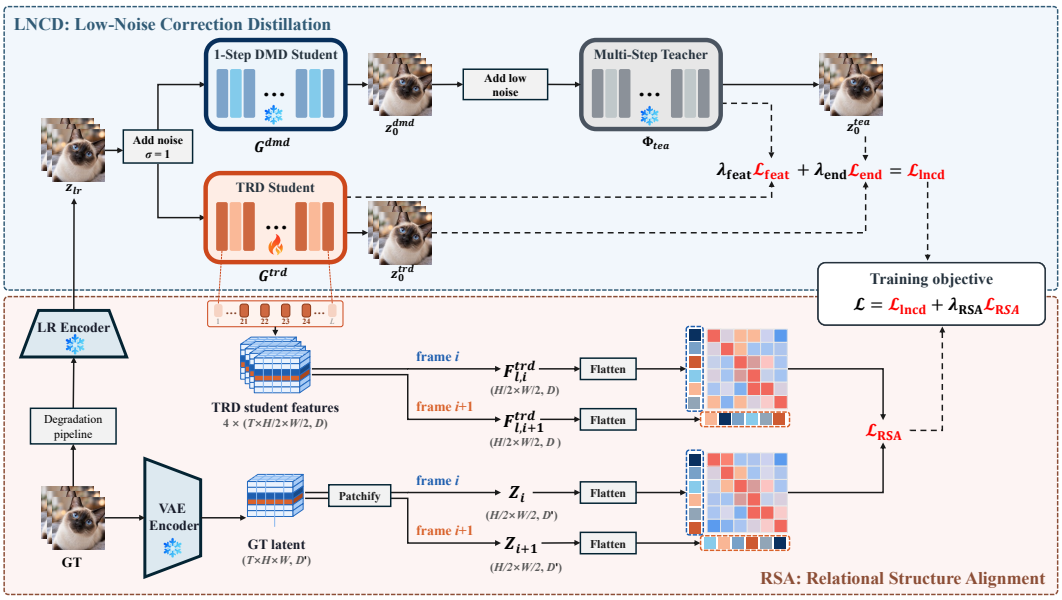

Figure 3: **Overview of TRD-VSR.** Starting from a one-step VSR model obtained by DMD distillation, we introduce two complementary forms of post-distillation supervision. **LNCD** re-distills the teacher's low-noise refinement through corrected endpoint supervision and intermediate feature alignment. **RSA** further constrains temporal structure by matching cross-frame relationships between the prediction and paired HR video. All teacher and auxiliary branches are used only during training, while inference retains the original one-step architecture.

Teacher-corrected target. We perturb the DMD prediction to a fixed small noise level σ_L with independent noise, and let the frozen teacher Φ_{tea}, conditioned on the LR latent, denoise it with M Euler steps on the schedule $\sigma_L = \sigma_M > \cdots > \sigma_0 = 0$:

$$\begin{aligned} \boldsymbol{z}_M^{\text{tea}} &= \mathcal{I}_{\sigma_L}(\boldsymbol{z}_0^{\text{dmd}}, \boldsymbol{\epsilon}'), \quad \boldsymbol{z}_0^{\text{dmd}} = G_{\text{dmd}}(\boldsymbol{\epsilon}, \boldsymbol{z}_{\text{LR}}), \quad \boldsymbol{\epsilon}, \boldsymbol{\epsilon}' \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \\ \boldsymbol{z}_{m-1}^{\text{tea}} &= \boldsymbol{z}_m^{\text{tea}} + (\sigma_{m-1} - \sigma_m)\, \Phi_{\text{tea}}(\boldsymbol{z}_m^{\text{tea}}, \sigma_m, \boldsymbol{z}_{\text{LR}}), \quad m = M, \ldots, 1. \end{aligned} \tag{8}$$

The final state $\boldsymbol{z}_0^{\text{tea}}$ serves as the teacher-corrected target; since σ_L is a small value, this adds only a few teacher evaluations during training and no inference cost.

Endpoint distillation. We then train G_{trd} to directly predict the teacher-corrected endpoint $\boldsymbol{z}_0^{\text{tea}}$ in a single forward pass:

$$\mathcal{L}_{\text{end}} = \mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{\epsilon}', \boldsymbol{z}_{\text{LR}}} \left[\left\| G_{\text{trd}}(\boldsymbol{\epsilon}, \boldsymbol{z}_{\text{LR}}) - \text{sg}\big(\boldsymbol{z}_0^{\text{tea}}\big) \right\|_2^2 \right], \tag{9}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. G_{trd} receives the same initial noise $\boldsymbol{\epsilon}$ used by G_{dmd} to generate $\boldsymbol{z}_0^{\text{dmd}}$, preserving the sample-wise correspondence between the DMD prediction and its teacher-corrected target.

Teacher feature distillation. While endpoint supervision aligns final outputs, it provides no direct guidance for intermediate representations. We thus directly align intermediate features of the student G_{trd} with those generated by the teacher Φ_{tea} during low-noise refinement.

Let $\boldsymbol{F}_l^{\text{trd}} \in \mathbb{R}^{TN \times D}$ and $\boldsymbol{F}_l^{\text{tea}} \in \mathbb{R}^{TN \times D}$ denote the layer-l tokens stacked across T frames for the student G_{trd} and teacher Φ_{tea}, respectively, with $\boldsymbol{F}_{l,i}^{\text{trd}}, \boldsymbol{F}_{l,i}^{\text{tea}} \in \mathbb{R}^{N \times D}$ denoting their tokens of frame i. Similar to REPA (Yu et al., 2025), a lightweight projector P_l maps the student representation into the teacher feature space. For selected layers $l \in \mathcal{S}$, we define

$$\mathcal{L}_{\text{feat}} = \frac{1}{|\mathcal{S}|} \sum_{l \in \mathcal{S}} \frac{1}{TN} \sum_{i=1}^{T} \sum_{n=1}^{N} \left[1 - \cos\big(P_l(\boldsymbol{f}_{l,i,n}^{\text{trd}}), \text{sg}(\boldsymbol{f}_{l,i,n}^{\text{tea}})\big) \right], \tag{10}$$

where $\boldsymbol{f}_{l,i,n}^{\text{trd}}$ and $\boldsymbol{f}_{l,i,n}^{\text{tea}}$ denote the features of the n-th token in frame i (i.e., the n-th row of $\boldsymbol{F}_{l,i}^{\text{trd}}$ and $\boldsymbol{F}_{l,i}^{\text{tea}}$), respectively, and $\cos(\cdot,\cdot)$ denotes cosine similarity along the feature dimension.

4.2 Relational Structure Alignment

Furthermore, we introduce **Relational Structure Alignment (RSA)** to strengthen the limited temporal supervision provided by the DMD objective (Sec. 3.3). Motivated by Sec. 3, where temporal consistency hinges on how tokens in frame i relate to tokens in frame $i{+}1$, RSA supervises these *cross-frame structural relations* against the ground truth, rather than the features themselves.

Specifically, for each selected layer $l \in \mathcal{S}$ and latent frame i, let $\boldsymbol{F}_{l,i} \in \mathbb{R}^{N\times D}$ denote the tokens of G_{trd} at layer l (omitting the superscript from $\boldsymbol{F}_{l,i}^{\text{trd}}$ for brevity). For the ground-truth HR latent $\boldsymbol{z}_{\text{HR}}$, we obtain its patchified latent tokens $\boldsymbol{Z}_i = \text{Patchify}(\boldsymbol{z}_{\text{HR}})_i \in \mathbb{R}^{N\times D'}$ on the same N-token grid. Each token is mean-centered and ℓ_2-normalized. In the spirit of relational knowledge distillation (Park et al., 2019), RSA aligns the cross-frame relation matrices of the two, averaged over the selected layers:

$$\mathcal{L}_{\text{RSA}} = \frac{1}{|\mathcal{S}|}\sum_{l\in\mathcal{S}} \mathbb{E}_i\Big[1 - \cos\big(\text{vec}(\boldsymbol{F}_{l,i}\boldsymbol{F}_{l,i+1}^{\top}),\ \text{vec}(\boldsymbol{Z}_i\boldsymbol{Z}_{i+1}^{\top})\big)\Big], \tag{11}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity, $\text{vec}(\cdot)$ flattens an $N \times N$ matrix into a vector, and $i = 1,\dots,T-1$ ranges over adjacent frame pairs. Ultimately, RSA constrains cross-frame feature coherence to suppress temporal flickering.

4.3 Training Objective

The complete training objective combines teacher-guided low-noise correction with HR-guided relational supervision:

$$\mathcal{L} = \underbrace{\lambda_{\text{end}}\mathcal{L}_{\text{end}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}}}_{\text{LNCD}} + \underbrace{\lambda_{\text{RSA}}\mathcal{L}_{\text{RSA}}}_{\text{RSA}}, \tag{12}$$

where λ_{end}, λ_{feat}, and λ_{RSA} are the loss weights controlling the relative importance of the endpoint, feature, and RSA losses, respectively.

5 Experiments

5.1 Experimental Setup

Training. We train on the HQ-VSR training set from DOVE (Chen et al., 2025), which contains 2,055 high-quality video clips. We adopt Wan2.1-1.3B (Team Wan et al., 2025) as the base model and perform parameter-efficient adaptation using LoRA (Hu et al., 2022). First, we fine-tune the DiT backbone with a flow-matching objective to obtain a multi-step teacher model. Second, we distill the teacher into a one-step student using DMD. Finally, we train our proposed TRD-VSR on HQ-VSR, again using LoRA adaptation, to improve the temporal consistency of one-step generation. Detailed objectives, trainable modules, optimization schedules, and hardware configurations for all three models are provided in Appendix A.

Evaluation. We evaluate on UDM10 (Yi et al., 2019), SPMCS (Tao et al., 2017), and YouHQ40 (Zhou et al., 2024), using the officially released LR–HR pairs. We report PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) for reference-based evaluation; CLIP-IQA (Wang et al., 2023a), MUSIQ (Ke et al., 2021), and DOVER (Wu et al., 2023) for no-reference quality. The temporal metrics are tOF, tLPIPS, and tLP (Chu et al., 2020), with E_{warp} (Lai et al., 2018) additionally reported in the component analysis. More experimental results are provided in Appendix B.4 and B.5.

5.2 Comparison with Existing Methods

As shown in Table 2, our method achieves state-of-the-art perceptual quality while maintaining near-state-of-the-art temporal consistency. Notably, among generative methods with similarly strong

Table 2: **Quantitative comparison on UDM10, SPMCS, and YouHQ40.** Best and second-best results are highlighted in green (bold) and lavender (underlined).

Method	tOF↓	tLPIPS↓	tLP↓	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MUSIQ↑	DOVER↑
UDM10 *10 videos, 32 frames*									
DLoRAL	2.564	0.229	8.89	22.86	0.707	0.425	0.512	64.2	0.487
SeedVR2	0.954	0.141	4.70	24.74	0.752	**0.352**	0.319	51.8	0.434
DOVE	1.515	0.126	2.25	**25.36**	**0.770**	0.365	0.500	61.0	0.567
FlashVSR	0.982	0.107	2.45	23.06	0.719	0.360	0.487	65.8	0.520
PS-SR	1.414	0.132	3.01	23.30	0.738	0.384	0.360	56.4	0.482
UltraVSR	2.157	0.199	5.35	22.81	0.703	0.416	0.414	63.8	0.525
Ours	**0.870**	**0.104**	**1.98**	22.91	0.730	0.376	**0.572**	**66.4**	**0.621**
SPMCS *30 clips, 29 frames*									
DLoRAL	0.938	0.169	10.21	19.67	0.522	0.435	0.478	67.4	0.413
SeedVR2	**0.298**	0.064	2.60	21.13	**0.624**	**0.361**	0.418	66.0	0.433
DOVE	0.500	0.065	**1.41**	**21.60**	0.620	0.398	0.573	69.1	0.496
FlashVSR	0.320	0.059	2.05	20.21	0.552	0.377	0.467	69.5	0.489
PS-SR	0.518	0.069	2.55	20.57	0.592	0.408	0.351	63.5	0.467
UltraVSR	0.756	0.134	6.98	19.33	0.506	0.422	0.414	69.3	0.513
Ours	0.333	**0.058**	1.81	20.06	0.570	0.412	**0.574**	**70.3**	**0.555**
YouHQ40 *40 clips, 29 frames*									
DLoRAL	6.711	0.182	5.61	21.44	0.607	0.445	0.496	66.1	0.631
SeedVR2	1.657	0.121	3.15	21.82	0.654	**0.376**	0.392	65.9	0.708
DOVE	1.272	0.113	2.63	**23.21**	**0.681**	0.407	0.486	61.0	0.702
FlashVSR	1.404	0.114	2.28	21.18	0.602	0.383	0.515	**68.7**	0.684
PS-SR	1.437	0.113	2.77	22.00	0.642	0.407	0.385	58.5	0.669
UltraVSR	4.683	0.161	3.93	21.44	0.597	0.437	0.436	66.3	0.674
Ours	**1.149**	**0.108**	**2.24**	21.24	0.624	0.420	**0.558**	66.4	**0.716**

perceptual quality, our method exhibits the best temporal consistency, producing more natural visual details with fewer temporally discontinuous artifacts. In contrast, our advantage is less pronounced on reference-based fidelity metrics such as PSNR and LPIPS. This is expected for generative restoration, where the underlying inverse problem is inherently one-to-many: visually plausible high-quality details need not exactly match the particular ground-truth realization and can therefore be penalized by reference-based metrics. Overall, our method favors perceptually natural and temporally coherent restoration over strict pixel-level fidelity.

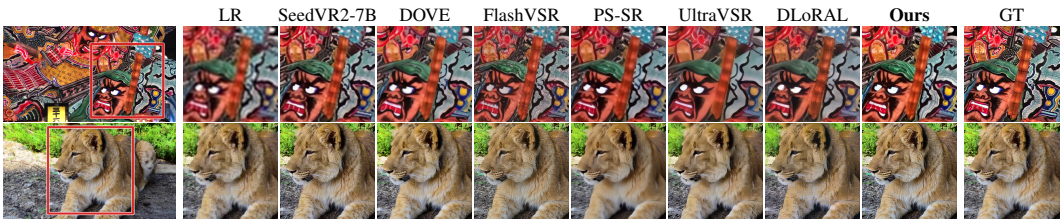

Figure 4: **Qualitative comparison on different datasets.** Zoom in for details.

The qualitative comparisons in Fig. 4 further demonstrate that our method reconstructs sharper fine details while maintaining better temporal consistency than existing approaches. In particular, challenging high-frequency structures exhibit less texture fluctuation across adjacent frames, consistent with the improvements in temporal metrics.

To further examine temporal stability, we visualize temporal profiles in Fig. 5. We fix a horizontal scanline in the frame and stack it across consecutive frames, so that any temporal inconsistency shows up as a discontinuity along the time axis. Competing generative methods produce noticeably jagged and noisy profiles: they hallucinate high-frequency details independently in each frame, and these details are not carried over consistently to the next frame. In contrast, our method yields smooth and continuous stripes that closely follow the ground-truth profile, which indicates that the synthesized details remain coherent over time.

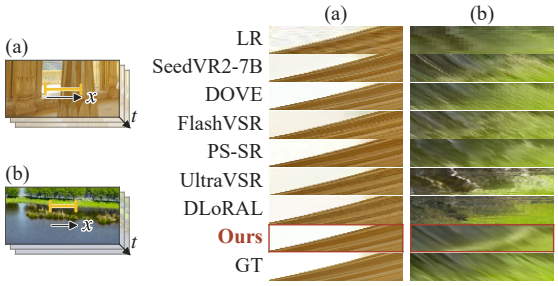

Figure 5: **Temporal profile (x–t slice) comparison on different videos.** The pixel row marked by the red line is stacked over time (vertical axis). Smooth, continuous stripes indicate temporally stable reconstruction, while jagged or broken stripes reveal flickering. Zoom in for details.

5.3 Ablation Study

Table 3: **Ablation of LNCD and RSA.** We compare the one-step student baseline, each module individually, and their combination on UDM10, SPMCS, and YouHQ40. Best and second-best results within each dataset are highlighted in green (bold) and lavender (underlined), respectively.

Variant	LNCD	RSA	tOF↓	E_{warp}↓	tLPIPS↓	CLIPIQA↑	MUSIQ↑	DOVER↑
UDM10 $n = 10$								
Baseline			1.384	0.877	0.1125	0.566	64.4	0.591
LNCD	✓		0.871	0.896	0.1074	**0.579**	66.0	**0.625**
RSA		✓	1.042	0.846	0.1091	0.565	65.1	0.603
Full	✓	✓	**0.870**	**0.821**	**0.1040**	0.572	**66.4**	0.621
SPMCS $n = 30$								
Baseline			0.372	0.791	**0.0573**	0.574	**71.0**	0.552
LNCD	✓		0.338	0.849	0.0580	**0.580**	70.2	0.548
RSA		✓	0.351	0.789	0.0578	0.573	69.7	0.553
Full	✓	✓	**0.333**	**0.785**	0.0576	0.574	70.3	**0.555**
YouHQ40 $n = 40$								
Baseline			1.810	1.602	0.1157	**0.560**	65.5	0.709
LNCD	✓		1.385	1.532	**0.1073**	0.559	**66.4**	**0.727**
RSA		✓	1.472	1.503	0.1112	0.557	65.8	0.712
Full	✓	✓	**1.149**	**1.466**	0.1076	0.558	**66.4**	0.716

Table 3 evaluates LNCD and RSA individually and jointly. All variants are re-distilled from the same one-step DMD student (Baseline) with identical training budgets. LNCD consistently reduces tOF across all three datasets, confirming the benefit of low-noise teacher correction, but increases E_{warp} slightly on UDM10 and SPMCS, likely due to sharper details that remain insufficiently anchored across frames. RSA complements LNCD by reducing E_{warp} on all benchmarks. Combining both modules yields the strongest overall temporal consistency, achieving the best tOF and E_{warp} across all datasets while maintaining competitive perceptual quality.

6 Conclusion

We study post-distillation correction for one-step generative VSR through LNCD, which distills low-noise teacher corrections, and RSA, which introduces cross-frame relational supervision from paired HR videos. Across three benchmarks, TRD-VSR consistently improves temporal consistency over the initial one-step student without additional sampling steps.

Future work could extend this framework to streaming generation, where limited future context may introduce temporal inconsistencies, and explore whether these principles generalize to efficient video generation under aggressive distillation.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods. We have not used generative AI tools to generate synthetic datasets, to help develop theoretical models or conceptual frameworks, to formulate mathematical claims, to propose or refine hypotheses, to design or provide feedback on research methodology or experiments, to assist with translation, to clean and reformat datasets, or to interpret results; providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, to search for and identify relevant literature, to draft parts of the paper and edit it for readability (including notation and equations), and to format references. All figures were created by the authors; AI assistance was limited to minor edits of plotting code. We have reviewed all AI-assisted work: AI-assisted code was checked by the authors, references identified with AI assistance were verified against the original publications, and all AI-edited text and equations were checked against the authors' intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects or the collection of personal data; all experiments use publicly available video super-resolution benchmarks (HQ-VSR, UDM10, SPMCS, YouHQ40, and MVSR4x) for research purposes. As with other generative restoration methods, TRD-VSR synthesizes plausible high-frequency details that are not guaranteed to match the true scene content. Restored videos should therefore not be treated as faithful evidence in forensic, medical, or other safety-critical settings, and such outputs could be misused to create misleading visual content. We encourage clear disclosure whenever generative restoration is applied to videos intended for evidential use.

REPRODUCIBILITY STATEMENT

The construction of the multi-step teacher and the DMD-distilled one-step student is described in Sec. 3.1, and the training objectives of TRD-VSR (LNCD and RSA) are given in Sec. 4. The training and evaluation datasets, degradation pipeline, benchmark protocols, and evaluation metrics are specified in Sec. 5. Appendix A lists the full training configuration of all three stages, including the backbone, LoRA rank, loss weights, correction noise levels, optimizer, learning rate, iteration counts, and hardware. All ablation variants in Sec. 5.3 share the same initialization checkpoint, training budget, and evaluation protocol.

REFERENCES

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. 11, 2024.

Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2848–2857, 2017.

Kelvin C. K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. BasicVSR: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4945–4954, 2021.

Kelvin C. K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5962–5971, 2022.

Zheng Chen, Zichen Zou, Kewei Zhang, Xiongfei Su, Xin Yuan, Yong Guo, and Yulun Zhang. DOVE: Efficient one-step diffusion model for real-world video super-resolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM Transactions on Graphics (TOG)*, 39(4):75, 2020.

Peng Dai, Xin Yu, Lan Ma, Baoheng Zhang, Jia Li, Wenbo Li, Jiajun Shen, and Xiaojuan Qi. Video demoireing with relation-based temporal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17622–17631, 2022.

Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased DMD: Few-step distribution matching distillation via score matching within subintervals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41667–41676, 2026.

Jiaqi Feng, Justin Cui, Yuanhao Ban, and Cho-Jui Hsieh. One-Forcing: Towards stable one-step autoregressive video generation. *arXiv preprint arXiv:2605.23458*, 2026.

Dario Fuoli, Martin Danelljan, Radu Timofte, and Luc Van Gool. Fast online video super-resolution with deformable attention pyramid. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1735–1744, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137, 2021.

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.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *European Conference on Computer Vision (ECCV)*, pp. 179–195, 2018.

Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2V-Turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. VRT: A video restoration transformer. *IEEE Transactions on Image Processing*, 33:2171–2182, 2024.

Shanchuan Lin and Xiao Yang. AnimateDiff-Lightning: Cross-model diffusion distillation. *arXiv preprint arXiv:2403.12706*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Yong Liu, Jinshan Pan, Yinchuan Li, Qingji Dong, Chao Zhu, Yu Guo, and Fei Wang. UltraVSR: Achieving ultra-realistic video super-resolution with efficient one-step diffusion space. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, pp. 7785–7794, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025.

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

Zhengyao Lv, Menghan Xia, Xintao Wang, and Kwan-Yee K. Wong. DUO-VSR: Dual-stream distillation for one-step video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16333–16344, 2026.

Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3967–3976, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

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

Yujing Sun, Lingchen Sun, Shuaizheng Liu, Rongyuan Wu, Zhengqiang Zhang, and Lei Zhang. One-step diffusion for detail-rich and temporally consistent video super-resolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Team Wan, 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, Mengyang Feng, 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. *arXiv preprint arXiv:2503.20314*, 2025.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision (ECCV)*, pp. 402–419, 2020.

Hugo Thimonier, Julien Despois, Robin Kips, and Matthieu Perrot. Learning long term style preserving blind video temporal consistency. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2021. doi: 10.1109/ICME51207.2021.9428445.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023a.

Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, and Lu Jiang. SeedVR: Seeding infinity in diffusion transformer towards generic video restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2161–2172, 2025.

Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, Xuefeng Xiao, Chen Change Loy, and Lu Jiang. SeedVR2: One-step video restoration via diffusion adversarial post-training. In *International Conference on Learning Representations (ICLR)*, 2026.

Ruohao Wang, Xiaohui Liu, Zhilu Zhang, Xiaohe Wu, Chun-Mei Feng, Lei Zhang, and Wangmeng Zuo. Benchmark dataset and effective inter-frame alignment for real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1168–1177, 2023b.

Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. VideoLCM: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023c.

Xintao Wang, Kelvin C. K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (ToG)*, 38(4):1–18, 2019.

Aiqiu Wu, Zhaofan Qiu, Ting Yao, and Tao Mei. PS-SR: Pseudo-single-step video super-resolution via speculative diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20087–20097, 2023.

Rui Xie, Yinhong Liu, Penghao Zhou, Chen Zhao, Jun Zhou, Kai Zhang, Zhenyu Zhang, Jian Yang, Zhenheng Yang, and Ying Tai. STAR: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *European Conference on Computer Vision (ECCV)*, 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, Xu Bin, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations (ICLR)*, 2025.

Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3106–3115, 2019.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623, 2024b.

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

Jiongze Yu, Xiangbo Gao, Pooja Verlani, Akshay Gadde, Yilin Wang, Balu Adsumilli, and Zhengzhong Tu. SparkVSR: Interactive video super-resolution via sparse keyframe propagation. In *European Conference on Computer Vision (ECCV)*, pp. 397–415, 2026.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Wenhu Zhang, Kun Cheng, Changyuan Wang, Shiyao Li, Yuechen Zhang, Wenbo Li, Jiajun Zha, Jingyi Zhang, Kang Zhao, and Jiaya Jia. CoDMD: Copula-aware distribution matching distillation for fast video generation. *arXiv preprint arXiv:2606.21982*, 2026a.

Yuehan Zhang and Angela Yao. RealViformer: Investigating attention for real-world video super-resolution. In *European Conference on Computer Vision (ECCV)*, 2024.

Ziqing Zhang, Kai Liu, Zheng Chen, Xi Li, Yucong Chen, Bingnan Duan, Linghe Kong, and Yulun Zhang. InfVSR: Toward consistency-driven streaming generative video super-resolution. In *International Conference on Machine Learning (ICML)*, 2026b.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In *International Conference on Learning Representations (ICLR)*, 2026.

Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-A-Video: Temporal-consistent diffusion model for real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. FlashVSR: Towards real-time diffusion-based streaming video super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43482–43493, 2026.

APPENDIX

This appendix complements the main paper. Appendix A details our three-stage training pipeline, covering multi-step teacher training, one-step DMD distillation, and TRD-VSR training. Appendix B presents additional experiments: further analysis of temporal consistency (Appendix B.1), attention-blocking interventions (Appendix B.2), an efficiency comparison (Appendix B.3), quantitative results on the real-world MVSR4x benchmark (Appendix B.4), and additional qualitative comparisons (Appendix B.5).

A IMPLEMENTATION DETAILS

Data and backbone. All training uses the HQ-VSR training set introduced in DOVE (Chen et al., 2025), consisting of 2,055 high-quality video clips. We use Wan2.1-1.3B as the backbone throughout the training pipeline. Rather than fully fine-tuning the backbone parameters, we adopt LoRA (Hu et al., 2022) with rank 384 for parameter-efficient adaptation.

Training procedure. TRD-VSR is an independent post-distillation procedure applied to an existing one-step VSR model. For completeness, we describe the preparation of the multi-step teacher and the one-step student, followed by the optimization of TRD-VSR.

(1) Multi-step teacher preparation. We first obtain a multi-step VSR teacher by fine-tuning the DiT backbone with the flow-matching objective. LoRA adapters are applied throughout the DiT backbone, while the remaining backbone parameters are kept frozen. The teacher is trained for 20K iterations and is subsequently used to construct the one-step student and provide low-noise correction supervision.

(2) One-step student preparation. The multi-step teacher is then distilled into a one-step generator using DMD. The student is optimized through LoRA adaptation for 10K iterations, with the pretrained backbone weights kept frozen. The resulting one-step generator serves as the starting point for our method. Importantly, these first two steps only provide the teacher–student pair required by TRD-VSR and are not specific to our temporal re-distillation formulation.

(3) TRD-VSR optimization. Given the pretrained one-step generator and its corresponding multi-step teacher, we optimize TRD-VSR on the same HQ-VSR training set. TRD-VSR is initialized from the one-step generator and trained for 5K iterations with a batch size of 4 using LoRA adaptation. Both $\mathcal{L}_{\text{feat}}$ and $\mathcal{L}_{\text{RSA}}$ use DiT blocks 21–24 as the selected layer set $\mathcal{S}$. The training objective combines endpoint correction distillation, intermediate feature alignment, and relational structure alignment, with $\lambda_{\text{end}} = 50$, $\lambda_{\text{feat}} = 0.25$, and $\lambda_{\text{RSA}} = 0.1$.

For teacher-guided correction, we first perturb the frozen one-step prediction to a low-noise state with $\sigma_L = 0.15$. The teacher then performs a first flow update to $\sigma = 0.07$ and continues the correction toward the clean endpoint. Through this post-distillation optimization, TRD-VSR learns to recover temporal refinement that is insufficiently preserved by aggressive one-step distillation, while retaining the spatial reconstruction capability of the original one-step generator. All teacher-side correction and auxiliary supervision are used only during training, so the resulting model preserves the original one-step inference procedure without introducing additional test-time computation.

Optimization and hardware. All stages are optimized using AdamW (Loshchilov & Hutter, 2019) with a learning rate of 1×10^{-5}. We use LoRA with rank 384 throughout the three-stage training pipeline. Training is conducted on 4 NVIDIA RTX PRO 6000 Blackwell GPUs. The three stages are trained for 20K, 10K, and 5K iterations, respectively. Unless otherwise stated, the backbone architecture and preprocessing configuration are shared across all stages.

B ADDITIONAL EXPERIMENTS

B.1 ADDITIONAL ANALYSIS OF TEMPORAL CONSISTENCY

Fig. 6 provides an additional qualitative comparison of temporal consistency between the multi-step teacher and the one-step student. We align moving patches to the first frame using ground-truth

Table 4: **Attention blocking during correction on UDM10** ($n = 10$)**.** E_{warp} and tLPIPS are normalized to the one-step student G_{dmd} (lower is better) and reported as median [IQR]. ΔPSNR and ΔLPIPS are measured relative to G_{dmd}. Superscripts indicate paired Wilcoxon signed-rank tests against normal correction: $^{*}p < 0.05$, $^{**}p < 0.01$.

Intervention	$\boldsymbol{E_{\text{warp}}}\downarrow$	**tLPIPS**↓	**ΔPSNR**↑	**ΔLPIPS**↓
Student	1.00	1.00	0.00	0.000
Normal correction	0.82 [0.77, 0.87]	0.89 [0.77, 0.93]	+0.15	0.000
Cross-frame blocked	1.05 [0.96, 1.23]**	1.03 [0.89, 1.09]**	+0.04	0.000
Within-frame blocked	1.16 [1.04, 1.23]*	1.22 [1.19, 1.27]**	+0.28	0.020
Both blocked	1.31 [1.27, 1.42]**	1.38 [1.34, 1.39]**	+0.14	0.020

optical flow estimated by RAFT (Teed & Deng, 2020), such that temporal variation after alignment mainly reflects changes in the synthesized texture rather than object motion.

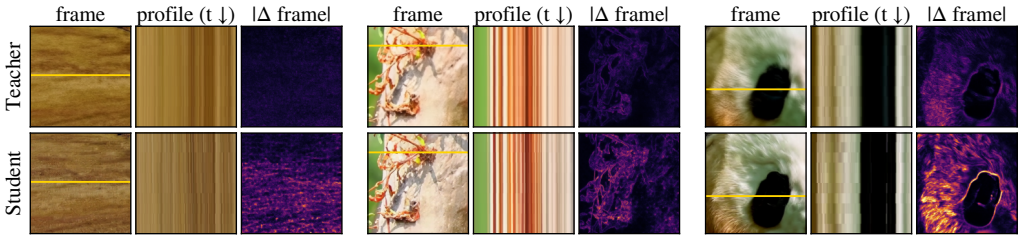

Figure 6: **Temporal consistency after one-step distillation.** Moving patches from YouHQ40 are aligned to the first frame using ground-truth optical flow. The teacher and student use the same LR input and initialization noise. Although the one-step student preserves a similar frame-wise appearance, its synthesized textures exhibit larger temporal variation than those of the 50-step teacher.

B.2 Does Spatiotemporal Interaction Carry the Teacher's Temporal Correction?

Sec. 3.2 showed that interactions among spatiotemporally correlated tokens become progressively stronger during the teacher's denoising trajectory. We further ask whether the resulting temporal improvement genuinely depends on local cross-frame interaction, rather than being merely a by-product of improved per-frame appearance.

Setup. We perform a one-step conditional-mean correction at $\sigma = 0.15$ on 10 UDM10 clips and intervene on all 30 self-attention layers of the frozen teacher. For each query, we selectively block one of the following attention neighborhoods: (i) a flow-aligned 3×3 local neighborhood in other frames, (ii) the 3×3 local neighborhood within the same frame, or (iii) both. All remaining attention connections, the LR condition, and model parameters are kept unchanged.

We compare each intervention with both the normal correction and the one-step student G_{dmd}. Temporal metrics are reported as per-clip ratios relative to G_{dmd}, while ΔPSNR and ΔLPIPS denote changes relative to G_{dmd}. Statistical significance is evaluated using paired Wilcoxon tests against the normal correction.

Cross-frame local interaction is critical for temporal consistency. As shown in Table 4, the normal correction substantially improves temporal consistency over the one-step student, yielding clear reductions in both E_{warp} and tLPIPS.

When the cross-frame local neighborhood is blocked, however, most of this temporal benefit disappears: both E_{warp} and tLPIPS increase slightly above the student baseline and deteriorate significantly relative to the normal correction. Moreover, 7 out of 10 clips become temporally worse than the original one-step student under this intervention.

Table 5: **Inference efficiency across resolutions and sequence lengths.** Each entry reports runtime in seconds, with peak GPU memory in GB in parentheses. NFE counts denoiser forward evaluations: MGLD-VSR uses 50 DDPM (Ho et al., 2020) steps without CFG; STAR uses 15 DPM++ 2M SDE (Lu et al., 2025) steps with CFG, yielding 30 evaluations; PS-SR uses one base step and three draft steps; SeedVR2, DOVE, FlashVSR, UltraVSR, and ours use a single denoiser evaluation. OOM denotes out-of-memory; † denotes CPU offloading. Columns give the output resolution and the number of frames (f).

Method	**NFE**	**540p**		**720p**		**1080p**	
		21f	41f	21f	41f	21f	41f
MGLD-VSR	50	292.4 (19.8)	526.6 (19.9)	295.7 (23.0)	532.6 (23.2)	1318.3 (30.3)	2371.8 (30.7)
STAR	30	52.1 (17.6)	136.6 (20.2)	52.3 (20.6)	137.0 (20.9)	137.5 (32.8)	360.2 (39.6)
SeedVR2-3B	1	6.3 (59.4)	12.4 (61.3)	10.7 (71.3)	20.6 (69.3)	30.4 (93.9)	56.7 (94.8)
SeedVR2-7B	1	6.8 (74.7)	13.0 (79.4)	11.2 (83.3)	21.6 (85.7)	35.5 (95.0)	OOM
SeedVR2-7B†	1	45.9 (55.9)	55.2 (59.6)	56.5 (59.5)	64.9 (60.3)	75.3 (74.0)	101.8 (82.5)
DOVE	1	6.4 (25.6)	10.5 (25.7)	11.8 (25.8)	19.3 (25.9)	26.5 (26.2)	43.9 (26.7)
FlashVSR	1	1.3 (11.8)	2.2 (13.1)	2.4 (17.9)	4.1 (19.4)	5.5 (36.7)	9.8 (40.9)
PS-SR	4	30.3 (15.8)	70.6 (17.3)	36.8 (19.0)	97.0 (19.3)	142.7 (19.0)	382.7 (19.3)
UltraVSR	1	4.8 (18.2)	9.0 (30.9)	14.3 (15.9)	27.9 (26.4)	27.8 (18.5)	54.8 (23.3)
Ours	1	3.1 (19.8)	6.1 (20.1)	5.3 (32.1)	10.5 (32.6)	12.5 (68.1)	25.8 (67.4)

Importantly, this pronounced temporal degradation is not accompanied by a comparable deterioration in per-frame appearance. Frame-wise PSNR and LPIPS remain nearly unchanged, indicating that the loss of temporal consistency cannot be explained simply by a general degradation in individual-frame reconstruction quality. Instead, these results suggest that access to local information across frames directly contributes to the teacher's temporal correction.

Cross-frame and within-frame neighborhoods play complementary roles. Blocking the within-frame local neighborhood also degrades temporal metrics, indicating that cross-frame and within-frame interactions are not completely independent. However, within-frame blocking causes a larger degradation in frame-wise perceptual quality (LPIPS) than cross-frame blocking, suggesting that within-frame local interactions play a more prominent role in maintaining spatial details within individual frames.

When both cross-frame and within-frame neighborhoods are blocked, temporal consistency degrades most severely, with 9 out of 10 clips performing worse than the one-step student.

Overall, these intervention results demonstrate the importance of interactions among spatiotemporally correlated tokens during the teacher correction. Cross-frame local neighborhoods are particularly important for coordinating information across time, whereas within-frame local neighborhoods contribute more strongly to preserving spatial details within individual frames.

B.3 Efficiency and Computational Cost

We compare the inference cost of representative diffusion-based VSR methods in terms of number of function evaluations (NFE), end-to-end runtime, and peak GPU memory. Each denoiser forward pass counts as one NFE; for methods using classifier-free guidance (CFG) (Ho & Salimans, 2022), conditional and unconditional passes are counted separately.

As shown in Table 5, our method requires only one denoiser evaluation and consistently achieves the second-lowest latency across all tested settings. At 1080p with 41 frames, ours takes 25.8 s, compared with 54.8 s for UltraVSR and 382.7 s for PS-SR, while SeedVR2-7B requires offloading to complete inference. The one-step design therefore translates directly into practical end-to-end acceleration.

Table 6: **Quantitative comparison on MVSR4x.** The benchmark contains 15 clips with 100 frames per clip under real-world $\times 4$ video super-resolution. Best and second-best results are highlighted in green (bold) and lavender (underlined).

Method	tOF↓	tLPIPS↓	tLP↓	E_{warp}↓	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MUSIQ↑	DOVER↑
MVSR4x *15 clips, 100 frames, real-world ×4*										
SeedVR2	**1.697**	0.157	5.65	**0.519**	**22.57**	**0.767**	0.410	0.240	31.6	0.181
DOVE	2.354	0.112	4.00	0.550	21.90	0.753	0.402	0.402	62.2	0.515
FlashVSR	2.171	0.115	4.07	1.075	21.83	0.724	**0.387**	0.435	**68.0**	0.544
UltraVSR	3.363	0.162	5.81	1.684	20.86	0.716	0.431	0.472	66.2	0.501
PS-SR	2.081	0.130	4.37	0.993	21.61	0.740	0.413	0.453	67.5	0.487
DLoRAL	3.324	0.184	7.37	0.948	22.10	0.734	0.432	**0.506**	61.4	0.409
Ours	1.749	**0.105**	**3.95**	1.508	21.57	0.731	0.401	0.439	67.9	**0.564**

B.4 Additional Quantitative Results

To further evaluate the performance of our method under real-world degradations, we conduct experiments on the MVSR4x benchmark (Wang et al., 2023b). As shown in Table 6, our method achieves competitive performance on both temporal consistency and perceptual quality metrics. The results suggest that the proposed method maintains stable temporal behavior while preserving favorable perceptual quality under complex real-world degradations, demonstrating its applicability to real-world video super-resolution.

B.5 Additional Qualitative Results

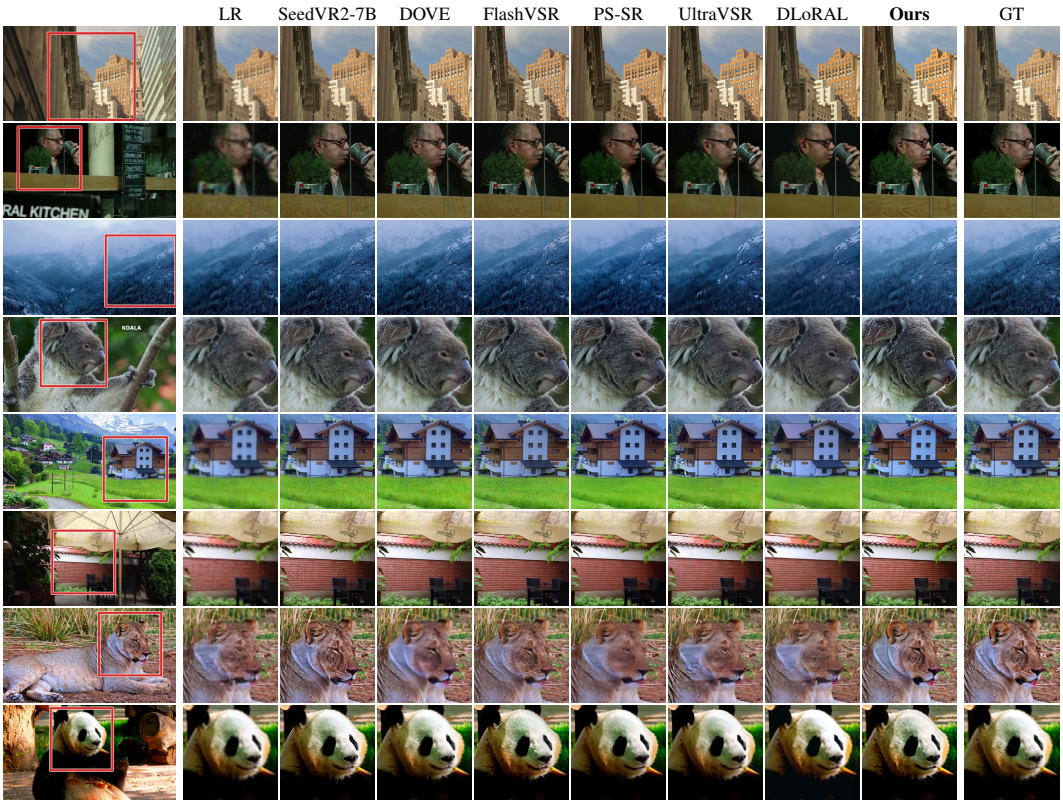

Figure 7: **Additional qualitative comparisons across diverse scenes.** Our method consistently preserves fine-grained details and achieves superior perceptual quality under varying visual conditions.

As shown in Fig. 7, our method exhibits a clear perceptual advantage and consistently retains rich fine-grained details across diverse scenes, demonstrating strong robustness under varying visual conditions.