RECURRENT SECOND-ORDER LEARNING FOR EFFICIENT VIDEO SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video super-resolution (VSR) seeks to reconstruct high-resolution frames from low-resolution inputs. While diffusion-based methods have substantially improved perceptual quality, extending them to video remains challenging for two reasons: strong generative priors can introduce temporal instability, and multi-frame diffusion pipelines are often too expensive for practical deployment. To address both challenges simultaneously, we propose InstaVSR, a lightweight diffusion framework for efficient video super-resolution. InstaVSR combines three ingredients: (1) a pruned one-step diffusion backbone that removes several costly components from conventional diffusion-based VSR pipelines, (2) recurrent training with flow-guided temporal regularization to improve frame-to-frame stability, and (3) dual-space adversarial learning in latent and pixel spaces to preserve perceptual quality after backbone simplification. On an NVIDIA RTX 4090, InstaVSR processes a 30-frame video at 2K×2K resolution in under one minute with only 7 GB of memory usage, substantially reducing the computational cost compared to existing diffusion-based methods while maintaining favorable perceptual quality with significantly smoother temporal transitions.

1 INTRODUCTION

Efficient video super-resolution (VSR) aims to reconstruct high-resolution (HR) video frames from low-resolution (LR) observations while maintaining practical computational cost. It supports a wide range of applications, including video streaming (Liu et al., 2017; Zhang et al., 2020; 2021a), surveillance enhancement (Farooq et al., 2021; Liu et al., 2024; Ding et al., 2020a), film restoration (Berlincioni et al., 2025; Peng et al., 2025), and medical imaging and related reconstruction tasks (Wang et al., 2022; Hu et al., 2024; Cammarasana et al., 2023).

Despite the great success achieved by existing VSR methods through lightweight architectures (Fuoli et al., 2019; Isobe et al., 2020), recurrent propagation (Sajjadi et al., 2018; Yan et al., 2019; Chan et al., 2021; Liang et al., 2022), and multi-frame alignment (Wang et al., 2019; Chan et al., 2022a), these techniques are typically studied in isolation. Moreover, when diffusion models are adopted for efficient one-step VSR, their strong generative priors may introduce temporal flickering by synthesizing inconsistent details across frames. Thus, one question is raised: *how can we systematically unify these designs within a diffusion framework to achieve a better efficiency–quality trade-off while balancing spatial synthesis and temporal consistency?*

To address this issue, we systematically revisit the design space of efficient video super-resolution. By an extensive analysis of over 10 core design components, we propose InstaVSR following two key principles including temporal consistency and spatial efficiency. Specifically, InstaVSR is a simple multi-input single-output (MISO) framework with lightweight flow-based alignment among frames, avoiding costly spatiotemporal 3D modeling. Then, we streamline the diffusion backbone by removing redundant VAE and text-conditioning components and adopting one-step generation, while dual-space adversarial supervision preserves its generative capability.

Even with an optimized module combination identified through extensive ablations, we observe that InstaVSR still exhibits unstable synthesis across frames under varying noise levels. We attribute this limitation to first-order learning, which directly learns the mapping from LR observations to HR targets, whereas stable reconstruction may require iterative refinement over multiple rounds to progressively converge. Motivated by this observation, we introduce recurrent second-order learning,

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

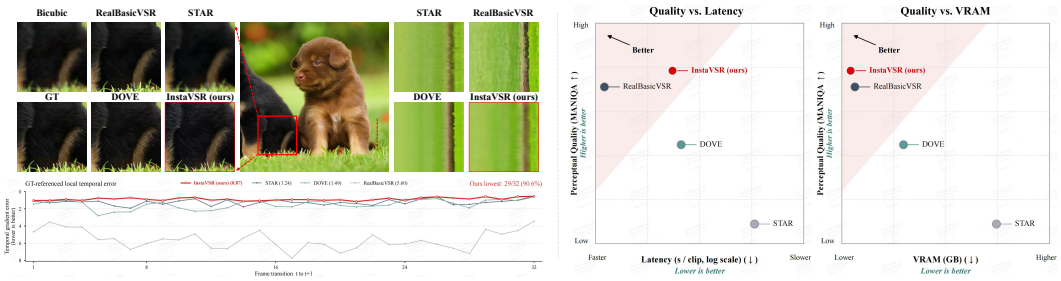


Figure 1: **InstaVSR delivers detailed and temporally consistent video super-resolution at low computational cost.** On the illustrated sequence, InstaVSR reconstructs fine spatial details and achieves the lowest GT-referenced local temporal gradient error on 90.6% of frame transitions. It also offers higher perceptual quality with lower inference latency and memory consumption.

which feeds previous SR predictions back as new observations for subsequent reconstruction. Given a sliding window of aligned LR frames, each predicted SR frame is written back to the temporal buffer to replace its LR counterpart and participate in subsequent predictions, allowing previously super-resolved frames to be revisited within an SR-enhanced context to yield second-order SR predictions. Furthermore, we couple this recurrent process with flow-guided temporal regularization to explicitly enforce the consistency of motion-aligned predictions across neighboring frames.

At last, extensive experiments on multiple VSR benchmarks demonstrate that InstaVSR achieves a favorable balance among efficiency, perceptual quality, and temporal consistency through systematic module integration and recurrent second-order learning. In particular, the current implementation contains only 0.45B parameters and processes a 30-frame sequence at 2K×2K resolution in approximately 0.77 minutes on an RTX 4090, with only 7 GB of GPU memory, demonstrating the practical efficiency of the proposed design.

Our contributions can be summarized as follows:

- We investigate and optimize key components for efficient VSR, including a lightweight one-step backbone, aligned multi-frame observations, simplified conditioning, and dual-space adversarial learning, and integrate them into a unified framework InstaVSR.
- Beyond conventional LR-to-SR supervision, we introduce recurrent second-order learning that progressively feeds first-order SR predictions back as observations for second-order reconstruction. This recurrent refinement exposes prediction errors and promotes temporally stable details without introducing heavy temporal architectures.
- Extensive experiments on standard VSR benchmarks demonstrate that InstaVSR achieves competitive perceptual quality and temporal consistency with substantially lower memory consumption and inference latency than existing diffusion-based VSR methods.

2 Related Work

Efficient VSR must balance two objectives: exploiting neighboring frames without temporal flickering or recurrent error accumulation, and keeping spatial reconstruction tractable at high resolution. Accordingly, we review prior work on temporal consistency and spatial efficiency.

2.1 Temporal Consistency in Video Super-Resolution

Temporal consistency is commonly achieved through frame alignment, recurrent propagation, or explicit temporal modeling. Early frame-recurrent methods reuse reconstructed frames or features for temporal coherence and efficiency. FRVSR (Sajjadi et al., 2018) propagates the previously estimated HR frame to reconstruct the next frame, while FFCVSR (Yan et al., 2019) combines local multi-frame information with previously reconstructed frames and features. RSDN (Isobe et al., 2020) decomposes recurrent representations into structure and detail components for robustness to appearance changes and accumulated errors. EDVR (Wang et al., 2019) instead uses deformable alignment and temporal-spatial attention for neighboring-frame aggregation. BasicVSR (Chan et al.,

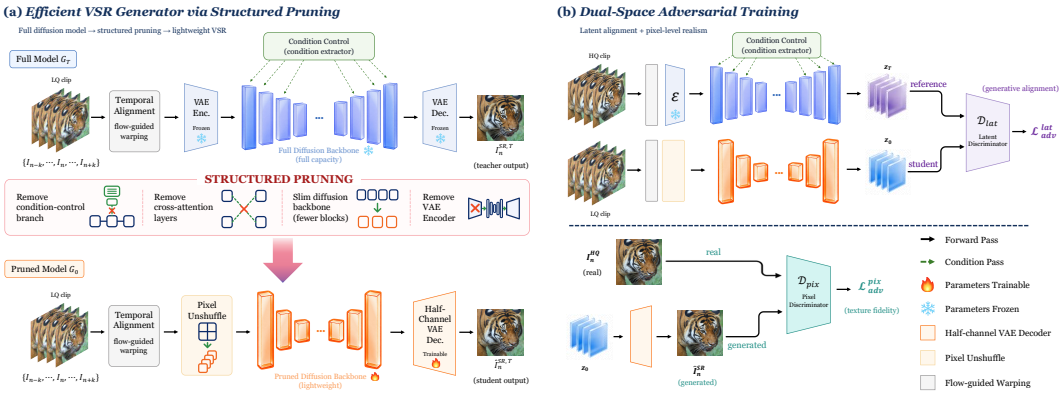


Figure 2: **Overview of Our Pruned Video Super-Resolution Framework and Dual-Space Adversarial Training Strategy.** (a) We prune the base model by removing the conditional control module for efficiency. A multi-frame input, single-frame output architecture enhances temporal stability, eliminating the need for the 3D VAE structure. (b) Our dual-space adversarial training uses latent and pixel discriminators. The latent discriminator aligns the pruned model's generative capacity with the full model, while the pixel discriminator enhances texture fidelity and prevents over-generation of fine details in reconstructed frames.

2021) propagates features bidirectionally, while BasicVSR++ (Chan et al., 2022a) introduces second-order grid propagation and flow-guided deformable alignment. RVRT (Liang et al., 2022) combines recurrent propagation with local attention for long-sequence restoration. Despite exploiting temporal redundancy, these methods may accumulate prediction and alignment errors under complex motion and diverse real-world degradation patterns.

Generative VSR adds a challenge because independently synthesized high-frequency details may vary between adjacent frames. Frame-wise methods such as StableSR (Wang et al., 2024) and DiffBIR (Lin et al., 2024b) provide strong perceptual priors but lack explicit temporal correspondence. Upscale-A-Video (Zhou et al., 2024) incorporates temporal layers and flow-guided latent propagation, whereas MGLD-VSR (Yang et al., 2024) uses motion-guided latent diffusion. These methods improve temporal coherence, but temporal layers, latent propagation, and iterative diffusion sampling raise the computational cost of inference.

Our recurrent second-order learning differs from conventional feature propagation. Rather than propagating only hidden features, we write reconstructed SR frames back to the observation buffer for reuse. A target frame is first reconstructed using the current buffer, then revisited after additional neighboring SR predictions have been written back. Here, "second-order" denotes reconstruction order, not the second-order feature propagation of BasicVSR++ (Chan et al., 2022a). Flow-guided regularization further constrains motion-aligned outputs without heavy 3D temporal modules.

2.2 Spatial Efficiency in Generative Super-Resolution

Spatial efficiency is critical because VSR computation and memory grow rapidly with frame resolution. Conventional methods reduce these costs through compact operators or efficient temporal propagation. RLSP (Fuoli et al., 2019) propagates temporal information through a compact recurrent latent state. ECBSR (Zhang et al., 2021b) uses structural re-parameterization, while ShuffleMixer (Sun et al., 2022) uses lightweight channel-mixing operations. SwinIR (Liang et al., 2021) introduces window-based attention for image restoration, and VRT (Liang et al., 2024) extends transformer modeling to video restoration. Although more efficient than large generative models, their regression-oriented objectives tend to favor distortion metrics and offer limited ability to synthesize realistic high-frequency details in reconstructed video frames.

Diffusion models provide stronger generative priors but incur substantial spatial overhead. Latent diffusion reduces feature resolution through a VAE (Rombach et al., 2022), while StableSR (Wang et al., 2024) and DiffBIR (Lin et al., 2024b) reuse pretrained diffusion priors for detail generation. However, iterative denoising and auxiliary conditioning remain expensive for individual video frames. Video-native diffusion architectures also use 3D VAEs, temporal attention, or spatiotemporal

transformers to process clips, further increasing high-resolution memory consumption (Yang et al., 2025; Brooks et al., 2024; Zheng et al., 2024; Lin et al., 2024a).

One-step generation offers a more efficient alternative. AdcSR (Chen et al., 2025a) shows that an SD2.1-based image super-resolution model produces perceptually realistic results in one forward pass, while AdcVSR (Chen et al., 2026) explores efficient video diffusion through teacher–student distillation. Nevertheless, directly extending an image model to video does not resolve the cost of multi-frame spatial processing or the temporal instability of synthesized details.

3 METHODOLOGY

3.1 PROBLEM FORMULATION AND OVERVIEW

Given an LR video $V^l \in \mathbb{R}^{T\times H_l\times W_l\times 3}$, we aim to recover its corresponding HR video $V^h \in \mathbb{R}^{T\times H_h\times W_h\times 3}$, where T is the number of frames, H_l and W_l are the height and width of each LR frame, and $H_h = rH_l$ and $W_h = rW_l$ for an upscaling factor r.

Our goal is to learn a VSR model G_θ that reconstructs $\hat{V}^h$ from V^l. To achieve efficient and temporally consistent reconstruction, InstaVSR combines a lightweight multi-frame one-step diffusion backbone, recurrent second-order learning, and dual-space adversarial supervision.

3.2 BASIC FRAMEWORK DESIGN

As illustrated in Figure 2, InstaVSR adopts a multi-input single-output framework. For each target LR frame I^l_t, we select $2k+1$ neighboring frames and align them to the target coordinate system:

$$\tilde{I}^l_{t+i} = \mathcal{W}\left(I^l_{t+i}, F_{t+i\to t}\right), \qquad i \in [-k, k], \tag{1}$$

where $F_{t+i\to t}$ is the optical flow from frame $t+i$ to frame t, and $\mathcal{W}$ denotes backward warping. The aligned frames are concatenated and fed into the generator:

$$\hat{I}^h_t = G_\theta\left(\text{Concat}[\tilde{I}^l_{t-k}, \ldots, I^l_t, \ldots, \tilde{I}^l_{t+k}]\right). \tag{2}$$

To reduce computation, we employ a pruned one-step diffusion backbone. The VAE encoder in the generator input path is replaced by Pixel Unshuffle, while text–image cross-attention, timestep embeddings, and redundant bottleneck blocks are removed. During training, reconstructed frames are progressively written back into the temporal buffer and reused in subsequent predictions. This recurrent process enables second-order learning, while flow-guided temporal regularization and dual-space adversarial supervision improve temporal consistency and perceptual quality.

3.3 RECURRENT SECOND-ORDER LEARNING

Frame-wise reconstruction losses supervise spatial quality but do not explicitly constrain consistency across frames. We introduce recurrent second-order learning that feeds SR predictions back into the observation buffer and reconstructs an earlier frame using the updated context. The resulting second-order prediction is paired with a neighboring first-order prediction for flow-guided temporal supervision across neighboring reconstructed frames.

Recurrent prediction. For the five-frame model, we sample seven consecutive LR frames $\{I^l_{t-3}, \ldots, I^l_{t+3}\}$ and initialize a temporal buffer with $B^{[0]}_j = I^l_j$. As illustrated in Figure 3, we sequentially reconstruct frames $t-1$, t, and $t+1$ using their aligned five-frame windows. Each SR prediction replaces its corresponding observation before the next window is processed. For $q = 0, 1, 2$ and $j = t-1+q$, this process is written as

$$\hat{I}^{h,(1)}_j = G_\theta\Big(\mathcal{A}_j(B^{[q]}_{j-2:j+2})\Big), \qquad B^{[q+1]}_j = \mathcal{R}\Big(\hat{I}^{h,(1)}_j\Big), \tag{3}$$

while all other buffer entries remain unchanged. Here, $B^{[q]}$ denotes the buffer after q updates, $\mathcal{A}_j$ denotes flow-guided alignment, and $\mathcal{R}$ converts an SR prediction to the representation required by the input buffer. The superscript (1) denotes the first reconstruction of a frame, which may already use previously reconstructed neighboring frames.

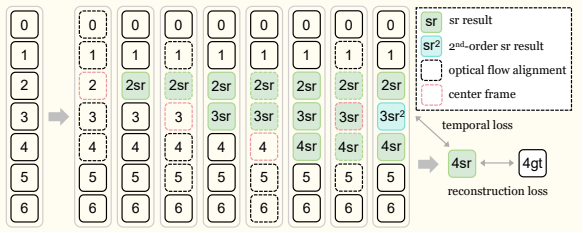


Figure 3: **Recurrent second-order training for VSR.** Frames 2, 3, and 4 are sequentially reconstructed from flow-aligned five-frame windows, with each SR prediction written back into the temporal buffer. Frame 3 is then reconstructed again using the updated context. Temporal loss compares its second-order result with the first-order result of frame 4 after flow alignment, while reconstruction loss supervises the latter against its ground truth.

Second-order reconstruction. After generating $\hat{I}^{h,(1)}_{t+1}$, we revisit frame t using the temporal buffer updated by preceding predictions:

$$\hat{I}^{h,(2)}_{t} = G_\theta\Big(\mathcal{A}_t(B^{[3]}_{t-2:t+2})\Big)\,. \tag{4}$$

The updated window contains the first-order SR results of frames $t-1$, t, and $t+1$. Reconstructing frame t again under this SR-enhanced context produces its second-order prediction. In Figure 3, frames $t-1$, t, and $t+1$ correspond to frames 2, 3, and 4, respectively. Here, "second-order" refers to reconstruction order, rather than the second-order feature propagation used for recurrent feature aggregation in BasicVSR++ (Chan et al., 2022a).

Reconstruction loss. We apply standard reconstruction supervision to the first-order prediction of frame $t+1$ at the final sequential reconstruction step:

$$\mathcal{L}_{\text{rec}} = \ell_{\text{rec}}\Big(\hat{I}^{h,(1)}_{t+1}, I^{h}_{t+1}\Big)\,, \tag{5}$$

where ℓ_{rec} denotes the reconstruction criterion and I^{h}_{t+1} is the corresponding HR target. This is the supervision between the first-order SR result of frame 4 and its ground truth in Figure 3.

Flow-guided temporal regularization. Temporal supervision compares the second-order prediction $\hat{I}^{h,(2)}_{t}$ with the neighboring first-order prediction $\hat{I}^{h,(1)}_{t+1}$. We estimate optical flow from the degraded LR frames using a lightweight estimator distilled from PWC-Net (Sun et al., 2018), and align the SR predictions before penalizing their appearance differences:

$$\mathcal{L}_{\text{temp}} = \frac{\sum_p \omega_{t\to t+1}(p)\rho\Big(\hat{I}^{h,(1)}_{t+1}(p) - \mathcal{W}\Big(\hat{I}^{h,(2)}_{t}, F_{t\to t+1}\Big)(p)\Big)}{\sum_p \omega_{t\to t+1}(p) + \epsilon}, \tag{6}$$

where $\mathcal{W}$ warps frame t into the coordinate system of frame $t+1$, and $\rho(x) = \sqrt{x^2+\epsilon^2}$ is the Charbonnier penalty. The confidence weight $\omega_{t\to t+1}$ incorporates a visibility mask obtained through forward–backward flow consistency to suppress unreliable correspondences. This loss encourages agreement between neighboring predictions generated at different reconstruction orders, complementing the spatial supervision provided by $\mathcal{L}_{\text{rec}}$. Further implementation details are provided for the complete recurrent training procedure in Appendix A.1.

3.4 Efficient One-Step Diffusion Architecture

Standard latent diffusion models use a VAE encoder to map an image into a compact latent space before denoising. In one-step VSR, however, the LR input already provides strong structural information. We therefore remove the VAE encoder from the generator input path to avoid additional learned compression and computational overhead.

Instead, the aligned multi-frame input is transformed using Pixel Unshuffle (Shi et al., 2016):

$$X_t = \text{PixelUnshuffle}\Big(\text{Concat}[\tilde{I}^{l}_{t-k}, \ldots, I^{l}_{t}, \ldots, \tilde{I}^{l}_{t+k}], s\Big)\,, \tag{7}$$

Table 1: **Quantitative comparison between InstaVSR and other methods.** The best, second-best, and third-best results of each dataset are marked in red, blue, and green, respectively.

Dataset	Method	LPIPS↓	DISTS↓	FID↓	NIQE↓	MANIQA↑	MUSIQ↑	E_{warp} ↓	DOVER↑
SPMCS	RealESRGAN Wang et al. (2021)	0.4339	0.2147	95.7204	3.472	0.6088	69.2294	3.1010	71.9639
	RealViformer Zhang & Yao (2024)	0.2144	0.1500	49.3322	3.793	0.6111	66.0845	2.3705	65.7693
	RealBasicVSR Chan et al. (2022b)	0.2076	0.1278	48.6680	3.292	0.6249	69.4399	2.8075	69.9027
	AdcSR Chen et al. (2025a)	0.2553	0.1533	71.9473	3.500	0.6382	72.0246	3.9597	71.5490
	UpscaleAVideo Zhou et al. (2024)	0.3420	0.1995	95.4628	4.821	0.4886	53.8774	2.7397	52.0868
	DOVE Chen et al. (2025b)	0.2192	0.1287	56.5175	4.463	0.6320	70.1333	2.2717	48.3011
	STAR Xie et al. (2025)	0.2857	0.1517	63.2338	4.288	0.6376	67.0929	3.0611	62.6324
	DLoRAL Sun et al. (2025)	0.2494	0.1596	58.9229	3.615	0.6471	70.7384	3.8593	70.5281
	InstaVSR (Ours)	0.2623	0.1475	59.2432	3.441	0.6779	71.3068	3.3329	72.5530
YouHQ	RealESRGAN Wang et al. (2021)	0.3188	0.1560	40.8991	3.293	0.5101	62.4595	8.3914	87.4597
	RealViformer Zhang & Yao (2024)	0.2702	0.1540	23.7465	3.273	0.5498	64.1624	5.1029	86.6196
	RealBasicVSR Chan et al. (2022b)	0.3167	0.1674	32.3203	3.051	0.5675	67.2988	6.0415	88.4188
	AdcSR Chen et al. (2025a)	0.3236	0.1606	29.8406	3.245	0.5671	70.4673	10.2355	86.5633
	UpscaleAVideo Zhou et al. (2024)	0.4331	0.2244	51.6875	5.900	0.3357	34.3326	3.5438	62.5278
	DOVE Chen et al. (2025b)	0.2336	0.1230	18.6305	4.519	0.4836	63.0005	3.6251	87.2314
	STAR Xie et al. (2025)	0.2121	0.1457	102.8143	6.678	0.3685	43.3237	5.1003	60.7029
	DLoRAL Sun et al. (2025)	0.3338	0.1700	34.5033	3.574	0.6003	71.0752	8.1360	86.5118
	InstaVSR (Ours)	0.3051	0.1254	20.9868	3.463	0.5910	66.0882	6.7386	87.2016
UDM10	RealESRGAN Wang et al. (2021)	0.3576	0.1934	65.0686	4.074	0.5786	64.7056	3.7062	81.58
	RealViformer Zhang & Yao (2024)	0.1290	0.0878	23.3183	4.181	0.5696	61.8556	3.1018	80.06
	RealBasicVSR Chan et al. (2022b)	0.1810	0.1073	36.3774	3.988	0.6065	66.0118	3.1118	81.55
	AdcSR Chen et al. (2025a)	0.2356	0.1410	46.5959	4.069	0.6149	70.5955	4.8715	82.27
	UpscaleAVideo Zhou et al. (2024)	0.3134	0.1785	93.0232	6.199	0.3600	39.4257	3.4148	57.29
	STAR Xie et al. (2025)	0.1693	0.0898	38.0273	4.689	0.5425	61.8154	3.1244	76.15
	DOVE Chen et al. (2025b)	0.1518	0.0911	35.4606	5.023	0.5078	62.4765	2.9801	78.75
	DLoRAL Sun et al. (2025)	0.2279	0.1412	44.1321	3.718	0.6254	69.9747	4.6991	81.67
	InstaVSR (Ours)	0.2427	0.1051	36.2891	4.466	0.6209	68.2072	3.1096	81.75

Table 2: **Efficiency comparison for generating a 30-frame $2K \times 2K$ video.** A800 results follow the common protocol; RTX 4090 results are additionally reported for methods that fit within 24 GB of GPU memory during inference. Peak memory is measured on the A800.

	AdcSR	UAV	DOVE	STAR	DLoRAL	Ours
Params (B)	0.46	14.44	10.55	2.49	1.30	**0.45**
A800 time (min)	1.3	25	1.4	24.5	3.5	**1.0**
RTX 4090 time (min)	1.0	–	–	–	–	**0.77**
Peak memory (GB)	17.23	38.85	31.79	75.25	24.93	**7.10**

where s is the Pixel Unshuffle factor. Pixel Unshuffle rearranges each $s \times s$ spatial region into the channel dimension. It is an invertible rearrangement and does not discard input values. The input stem of the UNet is modified to accept the resulting channel dimension.

We further remove text–image cross-attention layers because InstaVSR uses no text condition, and remove timestep embeddings because the model performs a fixed one-step reconstruction rather than iterative diffusion sampling. Finally, the bottleneck is reduced to a single residual block. These simplifications reduce the parameter count and inference cost while retaining the generative prior of the pretrained diffusion backbone. Architecture and initialization details are given in Appendix A.2.

3.5 Dual-Space Adversarial Learning

Reconstruction and temporal losses encourage fidelity and consistency but may produce conservative or oversmoothed results, especially after aggressive backbone simplification. A single discriminator is insufficient in this setting: latent-space supervision captures high-level image statistics but may overlook local artifacts, whereas pixel-space supervision is sensitive to fine textures but provides weaker semantic guidance. We therefore introduce two complementary discriminators operating in latent and pixel spaces. Both are used only during training and introduce no additional inference cost or test-time model parameters.



Figure 4: **Qualitative comparison of InstaVSR and other VSR methods.** InstaVSR reconstructs realistic textures and regular structures while avoiding the oversmoothing and distorted patterns observed in competing methods. Zoom in for the best view.

Latent-space discrimination. A frozen SD2.1 VAE encoder $\mathcal{E}$ maps generated and real HR frames into the same latent space:

$$z_{\text{gen}} = \mathcal{E}(\hat{I}), \qquad z_{\text{real}} = \mathcal{E}(I^{\text{GT}}). \tag{8}$$

Only the VAE encoder is used in this branch. The VAE decoder is unnecessary because both inputs are already RGB images and the purpose of $\mathcal{E}$ is only to obtain their latent representations.

We reuse the pretrained SD2.1 UNet as the latent discriminator D_{latent}. Its original parameters are frozen, and only lightweight LoRA modules (Hu et al., 2022) are optimized. LoRA allows the discriminator to adapt to artifacts produced by the pruned VSR generator without the computational cost and potential prior degradation of fully fine-tuning the UNet.

Pixel-space discrimination. Latent-space supervision may overlook small artifacts weakened by VAE compression. We therefore introduce a pixel discriminator D_{pixel} based on a frozen DINOv3-Large backbone (Siméoni et al., 2025) and lightweight trainable discrimination heads. Operating on RGB frames allows it to detect local defects such as jagged edges, blocky structures, unnatural textures, and residual high-frequency image noise.

The generator-side dual-space adversarial objective is

$$\mathcal{L}_G^{\text{dual}} = \lambda_{\text{latent}} \mathcal{L}_{\text{adv}}^{\text{latent}} + \lambda_{\text{pixel}} \mathcal{L}_{\text{adv}}^{\text{pixel}}. \tag{9}$$

The complete generator objective is

$$\mathcal{L}_G = \mathcal{L}_{\text{rec}} + \lambda_{\text{temp}} \mathcal{L}_{\text{temp}} + \mathcal{L}_G^{\text{dual}}. \tag{10}$$

Latent discriminator preserves high-level generative statistics, while the pixel discriminator improves local texture realism. Their complete training objectives are provided in Appendix A.3.

4 Experiments

4.1 Experimental Settings

We train a $4\times$ VSR model for 200 epochs on a single NVIDIA H100 GPU, using an initial learning rate of 2×10^{-5}, halved every 50 epochs, and a batch size of 2. The generator is initialized from the same-name and same-shape parameters of AdcSR Chen et al. (2025a). Each sample contains seven consecutive frames, from which five-frame windows are constructed for recurrent prediction. LQ–GT pairs are generated with the RealESRGAN degradation pipeline, using shared degradation parameters

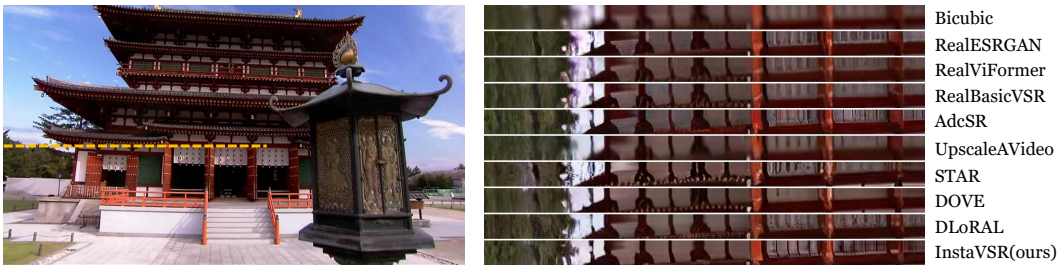


Figure 5: **Temporal profiles of video reconstruction.** InstaVSR produces smoother profiles while preserving straight lines and fine structural details across frames. Zoom in for the best view.



(a) Dual-space adversarial learning

(b) Recurrent learning and temporal regularization

Figure 6: **Qualitative results of the ablation study.** Zoom in for the best view.

within each sequence. A lightweight PWC-Net-distilled flow network is used for alignment and temporal supervision across neighboring video frames.

The training set contains approximately 20,000 YouTube videos at resolutions from 720p to 2K. We evaluate on SPMCS Tao et al. (2017), YouHQ Zhou et al. (2024), and UDM10 using LPIPS, DISTS Ding et al. (2020b), FID Heusel et al. (2017), NIQE Mittal et al. (2012), MANIQA Yang et al. (2022), MUSIQ Ke et al. (2021), E_{warp}, and DOVER Wu et al. (2023). Unless otherwise specified, efficiency is measured by generating a 30-frame $2K \times 2K$ video on an NVIDIA A800. RTX 4090 results are reported only for methods fitting within its 24 GB memory budget.

4.2 Quantitative Results

Table 1 shows that InstaVSR achieves a favorable perception–fidelity–temporal trade-off across all three benchmarks. On SPMCS, it obtains the best MANIQA and DOVER scores, the second-best NIQE and MUSIQ scores, and the third-best DISTS score. On YouHQ, InstaVSR ranks second in DISTS, FID, and MANIQA, while remaining competitive in DOVER. On UDM10, it ranks second in MANIQA and DOVER and third in FID, MUSIQ, and E_{warp}. These results indicate that InstaVSR produces perceptually strong videos with competitive overall video quality, although regression-oriented methods retain an advantage on several distortion metrics.

As shown in Table 2, InstaVSR contains only 0.45B parameters and requires 7.10 GB of peak GPU memory, both the lowest among the compared methods. It processes a 30-frame $(2K)^2$ video in 1.0 minute on an A800 and 0.77 minute on an RTX 4090. The lightweight flow estimator costs 4.821 ms per frame, corresponding to approximately 0.31% of the complete RTX 4090 runtime. Thus, flow estimation accounts for only a small fraction of the measured runtime while supporting aligned multi-frame reconstruction across neighboring low-resolution video frames.

4.3 Qualitative Results

Figure 4 compares spatial restoration quality on real-world videos. InstaVSR recovers fine-grained stone, brick, foliage, and soil textures while preserving the underlying scene structure. In contrast, competing methods either oversmooth these regions or generate irregular and geometrically inconsistent patterns in the illustrated image regions.

Figure 5 presents temporal profiles formed by stacking the same horizontal scanline from consecutive frames. InstaVSR produces smooth trajectories with well-preserved structural lines, whereas competing methods either remove fine structures or introduce jagged and temporally varying patterns.

Table 3: **Component ablation study on SPMCS.** Results report differences from the matched Full configuration, Δ = Variant − Full. Teal and red indicate improvement and degradation, respectively.

Configuration	PSNR↑	SSIM↑	NIQE↓	MANIQA↑	CLIPIQA↑	MUSIQ↑	E_{warp} ↓	DOVER↑
Full (reference)	21.764	0.557	3.441	0.678	0.609	71.307	3.333	72.553
Dual-space adversarial learning								
w/o $\mathcal{D}_{\mathrm{latent}}$	+0.268	+0.016	+0.196	+0.003	+0.006	+0.108	−0.355	−0.255
w/o $\mathcal{D}_{\mathrm{pixel}}$	+0.887	+0.039	+0.692	−0.039	−0.014	−2.674	−0.893	−6.164
Recurrent learning and temporal regularization								
First-order recurrence	+0.254	+0.013	+0.135	−0.006	−0.004	−0.929	−0.218	−1.945
w/o recurrent learning	+0.150	+0.013	+0.288	+0.015	+0.038	+0.130	−0.433	−0.155
w/o temporal loss	+0.255	+0.014	+0.046	−0.008	−0.007	−0.929	−0.153	−2.014
w/o recurrent and temporal loss	+0.257	+0.014	+0.342	−0.021	−0.058	−4.266	−0.483	−7.544
Flow-guided temporal alignment								
w/o flow alignment	+0.578	+0.025	+0.350	−0.058	−0.087	−5.771	−0.851	−10.732
One-step backbone design								
w/ VAE encoder	+0.260	+0.026	−0.071	+0.008	−0.021	+0.420	+0.164	+0.160
w/ cross-attention and time embedding	+0.090	+0.013	−0.034	+0.005	−0.018	+0.310	+0.195	−0.120

This demonstrates that the proposed method balances temporal smoothness with geometric fidelity in the illustrated temporal profiles.

4.4 Ablation Study

Adversarial supervision. Removing the pixel discriminator substantially reduces MANIQA, CLIP-IQA, MUSIQ, and DOVER, despite improving distortion-oriented metrics. Removing the latent discriminator produces smaller and mixed changes, indicating that the two discriminators provide complementary local and semantic supervision.

Recurrent and flow-guided learning. First-order recurrence underperforms the complete second-order formulation on perceptual and video-quality metrics. Removing both recurrent learning and temporal loss causes the largest degradation in this group, reducing MUSIQ and DOVER by 4.266 and 7.544, respectively. The particularly large DOVER decrease without flow alignment supports the use of motion compensation when exploiting neighboring observations. The occlusion mask further prevents unreliable correspondences from dominating temporal supervision.

Backbone design. Restoring the VAE encoder or cross-attention and timestep embeddings produces only mixed and modest quality changes, motivating their removal in our efficiency-oriented design. Temporal context size is analyzed in Appendix A.7. The qualitative results in Figure 6 support the quantitative observations. Without the pixel discriminator, outputs exhibit oversmoothed or blocky textures, while removing latent-space supervision increases content-inconsistent details. Without recurrent learning and temporal regularization, structures drift under repeated prediction and temporal variations become more visible. The complete configuration provides a favorable balance between realistic detail, structural fidelity, and temporal stability.

5 Conclusion

We presented InstaVSR, a lightweight one-step diffusion framework that balances spatial efficiency and temporal consistency for video super-resolution. Through systematic component analysis, we integrate a pruned backbone, flow-aligned multi-frame inputs, and dual-space adversarial supervision to reduce computational cost while preserving perceptual quality. Our recurrent second-order learning writes SR predictions back into the observation buffer and revisits target frames under an SR-enhanced context. Together with flow-guided temporal regularization, this promotes stable reconstruction without heavy spatiotemporal architectures. Experiments on three benchmarks demonstrate a favorable balance among perceptual quality, temporal consistency, and efficiency. With 0.45B parameters, InstaVSR processes a 30-frame $2K \times 2K$ sequence in 0.77 minutes on an RTX 4090 under the reported inference protocol.

Despite these gains, InstaVSR does not consistently outperform regression-oriented methods on fidelity metrics. Future work will investigate how to improve reconstruction fidelity while retaining perceptual quality and temporal stability under a limited computational budget.

Reproducibility Statement

Section 4.1 reports the training schedule, model initialization, datasets, and evaluation hardware. Appendix A.1 details recurrent prediction and temporal supervision; Appendices A.2 and A.3 specify the architecture, initialization, and adversarial objectives. Data preparation and evaluation metrics are described in Appendices A.4 and A.5. The experimental tables state the reference configurations and distinguish benchmark-resolution latency from the $2K \times 2K$ efficiency protocol.

AI Use Statement

Generative AI assistance was limited to language polishing and manuscript formatting. The authors remain responsible for verifying the revisions and for the scientific claims, experimental results, and final content of this work.

References

Lorenzo Berlincioni, Marco Bertini, and Alberto Del Bimbo. High-speed image enhancement: Real-time super-resolution and artifact removal for degraded analog footage. *Journal of Industrial Information Integration*, 45:100798, 2025. 1

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024. 4

Simone Cammarasana, Paolo Nicolardi, and Giuseppe Patanè. Super-resolution of 2d ultrasound images and videos. *Medical & Biological Engineering & Computing*, 61(10):2511–2526, 2023. 1

Kelvin CK 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*, pp. 4947–4956, 2021. 1, 2

Kelvin CK 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*, pp. 5972–5981, 2022a. 1, 3, 5

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Realbasicvsr: Investigating tradeoffs in real-world video super-resolution. In *CVPR*, 2022b. 6

Bin Chen, Gehui Li, Rongyuan Wu, Xindong Zhang, Jie Chen, Jian Zhang, and Lei Zhang. Adversarial diffusion compression for real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28208–28220, 2025a. 4, 6, 7

Bin Chen, Weiqi Li, Shijie Zhao, Xuanyu Zhang, Junlin Li, Jian Zhang, et al. Improved adversarial diffusion compression for real-world video super-resolution. In *The Fourteenth International Conference on Learning Representations*, 2026. 4

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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. 6

Dandan Ding, Junchao Tong, and Lingyi Kong. A deep learning approach for quality enhancement of surveillance video. *Journal of Intelligent Transportation Systems*, 24(3):304–314, 2020a. 1

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020b. 8

Muhammad Farooq, Matthew N Dailey, Arif Mahmood, Jednipat Moonrinta, and Mongkol Ekpanyapong. Human face super-resolution on poor quality surveillance video footage. *Neural Computing and Applications*, 33(20):13505–13523, 2021. 1

Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3476–3485, 2019. doi: 10.1109/ICCVW.2019.00431. 1, 3

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 8

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 7, 15

Jintong Hu, Hui Che, Zishuo Li, and Wenming Yang. Residual dense swin transformer for continuous depth-independent ultrasound imaging. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2280–2284. IEEE, 2024. 1

Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *ECCV*, 2020. 1, 2

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*, pp. 5148–5157, 2021. 8

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1833–1844, 2021. 3

Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35:378–393, 2022. 1, 3, 17

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. 3

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024a. 4

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *European Conference on Computer Vision*, pp. 430–448. Springer, 2024b. 3

Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2507–2515, 2017. 1

Zhifeng Liu, Zheng He, Gang Ye, and Wenqian Zhu. Advancing surveillance video clarity and transmission: a real-time video super-resolution model with background information awareness. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 304–318. Springer, 2024. 1

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 8

Jiaxin Peng, Siwang Zhou, Chengqing Li, Yucheng Li, and Dunyun Chen. Mitigating delivery artifacts in real-world video super-resolution. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3114–3123, 2025. 1

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. 3

Mehdi S. M. Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6626–6634, 2018. 1, 2

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883, 2016. 5

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 7, 16

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943, 2018. 5

Long Sun, Jinshan Pan, and Jinhui Tang. Shufflemixer: An efficient convnet for image super-resolution. *Advances in Neural Information Processing Systems*, 35:17314–17326, 2022. 3

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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 6

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*, pp. 4472–4480, 2017. 8

Huang Wang, Lu-Lu Li, Jin Shang, Jian Song, and Bin Liu. Application of deep learning image reconstruction in low-dose chest ct scan. *The British Journal of Radiology*, 95(1133):20210380, 2022. 1

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024. 3

Xintao Wang, Kelvin CK 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*, pp. 0–0, 2019. 1, 2

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1905–1914, 2021. 6

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*, pp. 20144–20154, 2023. 8

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*, pp. 17108–17118, 2025. 6

Bo Yan, Chuming Lin, and Weimin Tan. Frame and feature-context video super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5597–5604, 2019. doi: 10.1609/aaai.v33i01.33015597. 1, 2

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1191–1200, 2022. 8

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*, pp. 224–242. Springer, 2024. 3

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. 4

Aoyang Zhang, Qing Li, Ying Chen, Xiaoteng Ma, Longhao Zou, Yong Jiang, Zhimin Xu, and Gabriel-Miro Muntean. Video super-resolution and caching—an edge-assisted adaptive video streaming solution. *IEEE Transactions on Broadcasting*, 67(4):799–812, 2021a. 1

Xindong Zhang, Hui Zeng, and Lei Zhang. Edge-oriented convolution block for real-time super resolution on mobile devices. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4034–4043, 2021b. 3

Yinjie Zhang, Yuanxing Zhang, Yi Wu, Yu Tao, Kaigui Bian, Pan Zhou, Lingyang Song, and Hu Tuo. Improving quality of experience by adaptive video streaming with super-resolution. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pp. 1957–1966. IEEE, 2020. 1

Yuehan Zhang and Angela Yao. Realviformer: Investigating attention for real-world video super-resolution. In *European Conference on Computer Vision*, pp. 412–428. Springer, 2024. 6

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 4

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*, pp. 2535–2545, 2024. 3, 6, 8

A Additional Methodological Details

A.1 Recurrent Prediction and Temporal Regularization

Seven-frame training sequence. We sample seven consecutive LR frames $\{I^l_{t-3}, \dots, I^l_{t+3}\}$ and initialize $B^{[0]}_j = I^l_j$. The buffer is updated by sequentially reconstructing frames $t-1$, t, and $t+1$:

$$\begin{aligned} \hat{I}^{h,(1)}_{t-1} &= G_\theta\Big(\mathcal{A}_{t-1}(B^{[0]}_{t-3:t+1})\Big), \\ B^{[1]}_{t-1} &= \mathcal{R}\Big(\hat{I}^{h,(1)}_{t-1}\Big), \end{aligned} \tag{11}$$

$$\begin{aligned} \hat{I}^{h,(1)}_{t} &= G_\theta\Big(\mathcal{A}_{t}(B^{[1]}_{t-2:t+2})\Big), \\ B^{[2]}_{t} &= \mathcal{R}\Big(\hat{I}^{h,(1)}_{t}\Big), \end{aligned} \tag{12}$$

and

$$\begin{aligned} \hat{I}^{h,(1)}_{t+1} &= G_\theta\Big(\mathcal{A}_{t+1}(B^{[2]}_{t-1:t+3})\Big), \\ B^{[3]}_{t+1} &= \mathcal{R}\Big(\hat{I}^{h,(1)}_{t+1}\Big). \end{aligned} \tag{13}$$

At each update, all buffer entries other than the reconstructed center frame remain unchanged. Thus, the first prediction of frame t already uses the SR result of frame $t-1$, and the first prediction of frame $t+1$ uses both preceding SR results.

We then reconstruct frame t again using the updated buffer:

$$\hat{I}^{h,(2)}_{t} = G_\theta\Big(\mathcal{A}_{t}(B^{[3]}_{t-2:t+2})\Big). \tag{14}$$

This second-order prediction uses a window containing the first-order SR results of frames $t-1$, t, and $t+1$. These frames correspond to frames 2, 3, and 4 in Figure 3.

Supervision. The reconstruction loss in Eq. 5 compares $\hat{I}^{h,(1)}_{t+1}$ with its HR target I^h_{t+1}. The temporal loss in Eq. 6 compares $\hat{I}^{h,(2)}_{t}$ with $\hat{I}^{h,(1)}_{t+1}$ after flow alignment. The recurrent process constructs the first- and second-order predictions used for this supervision; it does not introduce a separate rollout loss during the recurrent training process.

Flow processing. The optical-flow estimator operates on degraded LR inputs. For an upscaling factor r, the estimated flow is resized to the HR spatial resolution, and both displacement components are multiplied by r before warping the SR predictions.

Let $F_{t\to t+1}$ denote the resulting flow used to warp the second-order prediction of frame t into the coordinate system of frame $t+1$. Forward–backward flow consistency provides a visibility mask $M_{t\to t+1}$, which excludes occluded regions and unreliable correspondences. The resulting per-pixel confidence weight is defined as follows:

$$\omega_{t\to t+1}(p) = M_{t\to t+1}(p)\exp\left(-\frac{\|F_{t\to t+1}(p)\|_2}{\sigma_m}\right), \tag{15}$$

where σ_m controls the weighting of motion magnitude. This reduces the contribution of large-motion regions, where flow estimation may be less reliable.

The weighted temporal loss penalizes differences between the aligned second-order and first-order predictions, while reconstruction supervision anchors the first-order prediction to its HR target.

A.2 One-Step Architecture and Initialization

Pixel Unshuffle input transformation. For $K = 2k+1$ aligned RGB frames, concatenation produces an input with $3K$ channels:

$$X^{\text{cat}}_t = \text{Concat}[\tilde{I}^l_{t-k}, \dots, I^l_t, \dots, \tilde{I}^l_{t+k}]. \tag{16}$$

Given a Pixel Unshuffle factor s, the transformation maps

$$X_t^{\text{cat}} \in \mathbb{R}^{H_l \times W_l \times 3K} \tag{17}$$

to

$$X_t \in \mathbb{R}^{\frac{H_l}{s} \times \frac{W_l}{s} \times 3Ks^2}. \tag{18}$$

Pixel Unshuffle changes only the arrangement of the input values and is therefore invertible. Unlike a learned VAE encoder, it does not introduce additional nonlinear compression before processing with the one-step diffusion UNet.

The input convolution of the UNet is modified to accept the expanded channel dimension. The aligned frames share the same target coordinate system before Pixel Unshuffle, allowing the input stem to jointly process their spatial and temporal information.

Backbone pruning. The original diffusion backbone contains text–image cross-attention layers for injecting text embeddings and timestep-embedding modules for representing different denoising steps. Neither component is required in InstaVSR because the aligned LR frames provide the conditioning signal and the generator performs a fixed one-step reconstruction.

We therefore remove the text–image cross-attention layers and timestep embeddings. We also reduce the UNet bottleneck to a single residual block. These modifications decrease the parameter count and computational cost while retaining the spatial representation inherited from the pretrained diffusion backbone used for model initialization.

The VAE is removed only from the generator input path. During training, the latent discriminator still uses a frozen VAE encoder to map generated and real RGB images into the pretrained latent space. No VAE decoder is used by the latent discriminator, and neither the latent nor the pixel discriminator is retained during inference.

Parameter initialization. The generator is initialized from AdcSR by loading parameters whose names and tensor shapes match the corresponding InstaVSR modules. Parameters associated with the expanded multi-frame input stem and structurally modified modules are initialized separately and trained for InstaVSR. The recurrent buffer, flow-guided temporal objectives, and adversarial components are also newly introduced and are not inherited from AdcSR.

Although concatenating multiple aligned frames increases the channel dimension of the input stem, the middle and bottleneck layers are simplified. As a result, InstaVSR remains slightly smaller than the single-frame AdcSR generator while exploiting multi-frame information.

A.3 Complete Dual-Space Adversarial Objectives

Latent-space discriminator. Given a generated frame $\hat{I}$ and its ground-truth HR frame I^{GT}, the frozen SD2.1 VAE encoder $\mathcal{E}$ produces

$$z_{\text{gen}} = \mathcal{E}(\hat{I}), \qquad z_{\text{real}} = \mathcal{E}(I^{\text{GT}}). \tag{19}$$

The pretrained SD2.1 UNet is reused as the latent discriminator D_{latent}. Its dense outputs are interpreted as patch-wise discrimination scores and aggregated when computing the adversarial loss. The original UNet parameters remain frozen, while trainable LoRA modules (Hu et al., 2022) adapt the discriminator to artifacts produced by the VSR generator.

The latent discriminator is optimized with the non-saturating logistic loss:

$$\begin{aligned}\mathcal{L}_D^{\text{latent}} = \mathbb{E}_{I^{\text{GT}}}\left[\text{softplus}\left(-D_{\text{latent}}(z_{\text{real}})\right)\right] \\ + \mathbb{E}_{\hat{I}}\left[\text{softplus}\left(D_{\text{latent}}\left(\text{sg}(z_{\text{gen}})\right)\right)\right],\end{aligned} \tag{20}$$

where $\text{sg}(\cdot)$ stops gradients from propagating to the generator during the discriminator update. The corresponding adversarial objective for the generator is

$$\mathcal{L}_{\text{adv}}^{\text{latent}} = \mathbb{E}_{\hat{I}}\left[\text{softplus}\left(-D_{\text{latent}}(z_{\text{gen}})\right)\right]. \tag{21}$$

Pixel-space discriminator. The pixel discriminator D_{pixel} uses a frozen DINOv3-Large backbone (Siméoni et al., 2025). Features extracted from three selected backbone levels are passed to lightweight trainable ConvNeXt heads, producing patch-wise real/fake predictions at multiple feature levels. The DINOv3 backbone remains frozen and in evaluation mode, while only the discriminator heads are optimized. DiffAugment is applied consistently to real and generated images during discriminator training at each optimization step.

The pixel discriminator is trained with the hinge loss:

$$\begin{aligned}\mathcal{L}_D^{\text{pixel}} = \mathbb{E}_{I^{\text{GT}}}\left[\max\left(0,\, 1 - D_{\text{pixel}}(I^{\text{GT}})\right)\right] \\ + \mathbb{E}_{\hat{I}}\left[\max\left(0,\, 1 + D_{\text{pixel}}\left(\text{sg}(\hat{I})\right)\right)\right].\end{aligned} \tag{22}$$

The corresponding generator objective is

$$\mathcal{L}_{\text{adv}}^{\text{pixel}} = -\mathbb{E}_{\hat{I}}\left[D_{\text{pixel}}(\hat{I})\right]. \tag{23}$$

The complete dual-space adversarial objective is

$$\mathcal{L}_G^{\text{dual}} = \lambda_{\text{latent}}\mathcal{L}_{\text{adv}}^{\text{latent}} + \lambda_{\text{pixel}}\mathcal{L}_{\text{adv}}^{\text{pixel}}. \tag{24}$$

Combining reconstruction, flow-guided temporal supervision, and dual-space adversarial learning yields the following complete generator training objective:

$$\mathcal{L}_G = \mathcal{L}_{\text{rec}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} + \lambda_{\text{latent}}\mathcal{L}_{\text{adv}}^{\text{latent}} + \lambda_{\text{pixel}}\mathcal{L}_{\text{adv}}^{\text{pixel}}. \tag{25}$$

The latent discriminator encourages the generated frames to follow the high-level statistics represented by the pretrained diffusion model, whereas the pixel discriminator penalizes local RGB artifacts that may be weakened by VAE encoding. Since both discriminators are removed after training, the dual-space objective does not increase inference parameters or runtime.

A.4 Training Data and Evaluation Protocol

We collect approximately 20,000 training videos from YouTube, with resolutions ranging from 720p to 2K. Before generating training pairs, the source videos are enhanced using a $1\times$ restoration model to reduce compression artifacts in the pseudo-ground-truth data. LQ sequences are then synthesized using the RealESRGAN degradation pipeline. The degradation parameters are shared across all frames of each sequence, which avoids unrealistic frame-wise changes in blur, noise, and compression across neighboring video frames.

In addition to the three quantitative benchmarks, we collect a real-world aerial-video set for qualitative analysis. It contains 10 scenes at 2K resolution, with each sequence consisting of 30 frames captured at 30 fps. The frame rate describes the source-video acquisition setting and does not imply real-time inference by the proposed model.

Training and inference hardware are reported separately. The model is trained on one NVIDIA H100 GPU, whereas the main efficiency comparison uses an NVIDIA A800 under the same 30-frame $2K \times 2K$ protocol. RTX 4090 measurements are additionally provided for methods that fit within 24 GB of memory. The use of an H100 for training therefore does not affect the fairness of the reported comparison of inference efficiency across methods.

A.5 Evaluation Metrics

We evaluate spatial fidelity using LPIPS, DISTS, and FID, and assess no-reference perceptual quality using NIQE, MANIQA, CLIPIQA, and MUSIQ. We report DOVER for overall video quality and use the following flow-guided warping error to quantify motion-compensated temporal consistency:

$$E_{\text{warp}} = \frac{1}{T-1}\sum_{t=1}^{T-1}\left\|\mathcal{W}\left(\hat{I}_{t+1}, F_{t+1\to t}\right) - \hat{I}_t\right\|_1, \tag{26}$$

where $\mathcal{W}$ denotes backward warping and $F_{t+1\to t}$ is the estimated optical flow from frame $t+1$ to frame t. A lower E_{warp} indicates stronger motion-compensated temporal consistency. DOVER complements this low-level measure by assessing overall perceptual video quality.

Table 4: **Comparison with the efficient non-diffusion baseline RVRT on benchmark-resolution SPMCS clips.** Latency is measured over 30 frames and is not directly comparable with the $2K \times 2K$ protocol used in the main paper.

Method	LPIPS↓	DISTS↓	NIQE↓	MANIQA↑	CLIPIQA↑	MUSIQ↑	E_{warp} ↓	DOVER↑	Latency↓ (s/30 frames)
RVRT Liang et al. (2022)	0.1835	0.1028	5.1069	0.6304	0.4476	63.2708	1.4734	61.3796	18.312
InstaVSR	0.2623	0.1475	3.4410	0.6779	0.5550	71.3068	3.3329	72.5530	8.167

Table 5: **Temporal context size ablation on SPMCS.** The five-frame configuration reports absolute values; three- and seven-frame configurations report differences from this reference, Δ = Variant − 5-frame. Teal and dark red indicate improvement and degradation, respectively.

Setting	Flow time (ms/frame)↓	Total time (s/frame)↓	MANIQA↑	E_{warp} ↓
5-frame (Ours)	4.821	1.54	0.678	3.333
3-frame	−1.070	−0.11	−0.017	+0.194
7-frame	+1.088	+0.14	−0.030	−0.305

A.6 Comparison with RVRT

Table 4 compares InstaVSR with RVRT, an efficient non-diffusion VSR method. RVRT performs better on LPIPS, DISTS, and warping error, whereas InstaVSR achieves substantially stronger NIQE, MANIQA, CLIPIQA, MUSIQ, and DOVER scores. Under this benchmark-resolution protocol, InstaVSR also requires less than half the measured latency. The results reflect the different objectives of deterministic regression and generative restoration: RVRT prioritizes pixel fidelity, while InstaVSR provides stronger perceptual quality and overall video quality.

A.7 Temporal Context Size

Table 5 compares three-, five-, and seven-frame input windows. Three frames reduce runtime but yield worse MANIQA and warping error, whereas seven frames improve warping error at the cost of runtime and perceptual quality. We therefore use five frames as the default balance.

For the five-frame configuration, the lightweight flow estimator costs 4.821 ms per frame, approximately 0.31% of the complete RTX 4090 runtime, indicating a small flow-estimation overhead.

A.8 Temporal Range and Occlusion Masking

Table 6: **Additional temporal ablations on SPMCS.** Results report differences from the matched Full configuration used in Table 3, Δ = Variant − Full. Teal and red indicate improvement and degradation relative to the Full configuration, respectively.

Configuration	PSNR↑	SSIM↑	NIQE↓	MANIQA↑	CLIPIQA↑	MUSIQ↑	E_{warp} ↓	DOVER↑
Temporal range $D = 2$	+0.141	+0.008	+0.438	+0.004	−0.014	−1.001	−0.221	−2.192
w/o occlusion mask	−0.018	0.000	+0.275	−0.012	−0.015	−0.948	−0.085	−4.115

Table 6 examines temporal range and occlusion masking. Setting $D = 2$ produces mixed changes: PSNR, SSIM, MANIQA, and warping error improve, while NIQE, CLIPIQA, MUSIQ, and DOVER worsen. In particular, MUSIQ and DOVER decrease by 1.001 and 2.192, respectively.

Removing the occlusion mask reduces MANIQA, CLIPIQA, MUSIQ, and DOVER, with a DOVER drop of 4.115, despite a small improvement in warping error. These results support retaining occlusion masking for perceptual and overall video quality, while showing that lower warping error alone does not necessarily indicate better perceptual quality.