Stream-DiffVSR: Low-Latency Streamable Video Super-Resolution via Auto-Regressive Diffusion

Hau-Shiang Shiu[1], Chin-Yang Lin[1], Zhixiang Wang[2], Chi-Wei Hsiao[3], Po-Fan Yu[1], Yu-Chih Chen[1], and Yu-Lun Liu[1]

[1] Department of Computer Science, National Yang Ming Chiao Tung University,
[2] Shanda AI Research Tokyo, [3] MediaTek Inc.
xhs0964519.cs12@nycu.edu.tw, yulunliu@cs.nycu.edu.tw

Fig. 1: Comparison of visual quality and inference speed across VSR methods. Stream-DiffVSR achieves superior perceptual quality and comparable runtime to CNN- and Transformer-based models, with significantly reduced Time-to-First-Frame (TTFF) versus offline approaches. Best/second-best in **red**/**green**.

Abstract. Diffusion-based video super-resolution (VSR) methods deliver strong perceptual quality but are often unsuitable for latency-sensitive scenarios due to reliance on future frames and expensive multi-step denoising. We propose Stream-DiffVSR, a causally conditioned diffusion framework for efficient online VSR. Operating strictly on past frames, Stream-DiffVSR integrates a four-step distilled denoiser for fast inference, an Auto-regressive Temporal Guidance (ARTG) module that injects motion-aligned cues during latent denoising, and a lightweight temporal-aware decoder with a Temporal Processor Module (TPM) to enhance detail and temporal coherence. Unlike chunk-wise streaming inference, our strictly frame-by-frame causal design avoids sequence-level waiting, substantially reducing time-to-first-frame and end-to-end latency. Stream-DiffVSR processes 720p frames in 0.328 seconds on an RTX 4090 and consistently outperforms prior diffusion-based baselines. Compared with the online state-of-the-art TMP [120], it improves perceptual quality (LPIPS +0.095). Compared with prior diffusion-based VSR methods such as MGLD-VSR [104], it reduces per-frame runtime by over 130×. Moreover, Stream-DiffVSR substantially lowers time-to-first-frame for diffusion-based VSR, reducing initial delay from over 4600 seconds to 0.328 seconds, making diffusion-based VSR markedly more practical for low-latency online and streaming deployment. Project page: https://jamichss.github.io/stream-diffvsr-project-page/

Keywords: Video super-resolution · Diffusion models · Online/Streaming inference

1 Introduction

Video super-resolution (VSR) aims to reconstruct high-resolution (HR) videos from low-resolution (LR) inputs and is vital in applications such as surveillance, live broadcasting, video conferencing, autonomous driving, and drone imaging. It is increasingly important in low-latency rendering workflows, including neural rendering and resolution upscaling in game engines and AR/VR systems, where latency-aware processing is crucial for visual continuity.

Specifically, latency-sensitive processing involves two key aspects: per-frame inference time (throughput) and end-to-end system latency (delay between receiving an input frame and producing its output). Existing VSR methods often struggle with this trade-off. While CNN- and Transformer-based models offer a balance between efficiency and quality, they fall short in perceptual detail. Diffusion-based models excel in perceptual quality due to strong generative priors, but suffer from high computational cost and reliance on future frames, making them impractical for time-sensitive video applications.

In this paper, we propose **Stream-DiffVSR**, a diffusion-based framework that establishes a fully causal, frame-by-frame paradigm for online video super-resolution, bridging the gap between high-quality but slow diffusion methods and fast but lower-quality CNN- or Transformer-based methods. Unlike previous diffusion-based VSR approaches (e.g., StableVSR [68] and MGLD-VSR [104]) that typically require 50 or more denoising steps and bidirectional temporal information, we adopt a staged training scheme with rollout distillation over full inference trajectories, reducing denoising steps to just four while remaining robust under autoregressive rollout. Additionally, we introduce an Auto-regressive Temporal Guidance mechanism and an Auto-regressive Temporal-aware Decoder, with a learnable gating parameter controlling reliance on the previous frame, to effectively exploit temporal information from previous frames, significantly enhancing temporal consistency and perceptual fidelity.

Fig. 1 illustrates the core advantage of our approach by comparing visual quality and runtime across various categories of video super-resolution methods. Stream-DiffVSR achieves improved perceptual quality, as reflected by lower LPIPS [117] and temporal consistency, outperforming existing unidirectional CNN- and Transformer-based methods (e.g., MIA-VSR [130], RealViformer [119], TMP [120]). Notably, Stream-DiffVSR offers significantly faster per-frame inference than prior diffusion-based approaches (e.g., StableVSR [68], MGLD-VSR [104]), attributed to our use of a distilled 4-step denoising process and a lightweight temporal-aware decoder.

In addition, existing diffusion-based methods, such as StableVSR [68] typically rely on bidirectional or future-frame information, resulting in prohibitively high processing latency that is not suitable for online scenarios. Specifically, for a 100-frame video, StableVSR (46.2 s/frame) would incur an initial latency ex-

Table 1: Comparison of diffusion-based VSR methods. We report online capability, inference steps, runtime (FPS on 720p, RTX 4090), Time-to-First-Frame (TTFF in sec), and whether each method uses distillation, temporal modeling, or offline future frames. OOM denotes out-of-memory; - indicates missing public inference results. Notably, Stream-DiffVSR runs in a strictly online, past-only setting and achieves one of the lowest end-to-end latencies among compared diffusion-based VSR methods under our measurement protocol.

Method	Online	# of Steps	FPS @720p	TTFF (sec)	Distill	Temporal Input	Temporal Decoder
StableVSR [68]	✗	50	0.02	4620	✗	Future/Bi-dir	✗
MGLD-VSR [104]	✗	50	0.02	218	✗	Future/Bi-dir	✓
Upscale-A-Video [129]	✗	30	OOM	-	✗	Future/Bi-dir	✓
VEnhancer [29]	✗	15	OOM	-	✗	Future/Bi-dir	✓
Stream-DiffVSR (ours)	✓	**4**	**3.05**	0.328	✓	**Past-only**	✓

ceeding 4600 seconds on an RTX 4090 GPU, as it requires processing the entire sequence before generating even the first output frame. In contrast, our Stream-DiffVSR operates in a strictly causal, autoregressive manner, conditioning only on the immediately preceding frame. Consequently, the initial frame latency of Stream-DiffVSR corresponds to a single frame's inference time (0.328 s/frame), reducing the latency by more than three orders of magnitude compared to StableVSR. This significant latency reduction demonstrates that Stream-DiffVSR effectively unlocks the potential of diffusion models for practical, low-latency online video super-resolution.

To summarize, the main contributions of this paper are:

- We introduce Stream-DiffVSR, a diffusion-based framework explicitly designed for **streamable, low-latency** video super-resolution, trained via a staged image-to-video scheme and distilled using full-trajectory rollout distillation, enabling efficient inference by reducing diffusion sampling from 50 denoising steps down to 4 steps.
- We propose a novel Auto-regressive Temporal Guidance mechanism and a Temporal-aware Decoder to effectively leverage temporal **cues** *only* from past frames, significantly enhancing perceptual quality and temporal consistency.
- Extensive experiments demonstrate that our approach outperforms existing methods on key perceptual and temporal consistency metrics while maintaining **low-latency inference**, making diffusion-based VSR practical for latency-critical online/streaming and interactive video pipelines.

To contextualize our contributions, Table 1 compares recent diffusion-based VSR methods in terms of online inference capability, runtime efficiency, and temporal modeling. Stream-DiffVSR performs online, low-latency inference in a strictly causal (past-only), frame-by-frame manner, delivering diffusion-level perceptual quality while maintaining strong temporal stability. This substantial latency reduction of over three orders of magnitude compared to prior diffusion-based VSR models makes diffusion-based VSR markedly more practical for latency-critical online applications such as video conferencing and AR/VR.

2 Related Work

Video Super-resolution. VSR methods reconstruct high-resolution videos from low-resolution inputs through CNN-based approaches [6, 7, 79, 80, 89, 102], deformable convolutions [18, 80, 132], online processing [20, 108, 120], recurrent architectures [21, 35, 44, 70, 107], flow-guided methods [27, 56, 112], Transformer-based models [47, 49, 74, 83, 130], and implicit alignment techniques [99]. Causal and streaming paradigms have emerged for low-latency scenarios [25, 63, 127], while structured pruning [95] targets efficient deployment. Despite advances, low-latency online processing with high perceptual quality remains challenging.

VSR under Unknown Degradations. VSR under unknown degradations studies how to handle diverse and poorly specified degradation processes [8,105] through pre-cleaning modules [8,26,55,90], online approaches [119], kernel estimation [37, 62], synthetic degradations [9, 36, 76, 118], new benchmarks [17, 123], real-time systems [5,38], advanced GANs [11,14,82], and Transformer restorers [4,48,114]. Recent efforts leverage text-to-video priors for VSR with complex, non-ideal degradations [97, 124]. Warp error-aware consistency [42] emphasizes temporal error regularisation.

Diffusion-based Image and Video Restoration. Diffusion models provide powerful generative priors [10, 19, 67] for single-image SR [33, 43, 69, 91, 113], inpainting [52, 59, 81, 92], and quality enhancement [23, 32, 88]. Video diffusion methods include StableVSR [68], MGLD-VSR [104], DC-VSR [28], DOVE [15], Ultra-VSR [54], Upscale-A-Video [129], DiffVSR [45], DiffIR2VR-Zero [106], VideoGigaGAN [101], VEnhancer [29], temporal coherence [86], AVID [121], and SeedVR2 [87]. Temporal consistency in video diffusion has also been addressed through optical-flow-guided approaches [16, 46, 103]. Auto-regressive video diffusion has emerged as a promising paradigm for streaming generation. CausVid [111] distills a bidirectional video diffusion transformer into a causal auto-regressive generator with 4-step inference. Self Forcing [34] addresses exposure bias in auto-regressive video diffusion by conditioning on self-generated outputs during training. Diffusion Forcing [12] unifies next-token prediction with full-sequence diffusion through independent per-token noise levels. Other auto-regressive approaches [51, 78, 96, 122] and causal architectures [22, 41] further demonstrate the potential of streaming diffusion. Acceleration techniques include consistency models [2, 24, 30, 40, 57, 75], advanced solvers [1, 58, 126], flow-based methods [39, 53], adversarial distillation [50, 72, 73, 100], distribution matching distillation [109, 110, 128], other distillation approaches [13, 60, 64, 71, 98, 125, 131, 133], video-specific distillation [115], and efficient architectures [3]. Theoretical advances [84, 85] and recent image/offline distillation methods [77, 93, 94, 116] exist. In contrast, *Stream-DiffVSR* combines diffusion distillation with strictly online (past-only) causal temporal modeling to enable low-latency VSR.

3 Method

We propose Stream-DiffVSR, a streamable auto-regressive diffusion framework for efficient video super-resolution (VSR). Its core innovation lies in an auto-

regressive formulation that improves both temporal consistency and inference speed. The framework comprises: (1) a distilled few-step U-Net for accelerated diffusion inference, (2) Auto-regressive Temporal Guidance that conditions latent denoising on previously warped high-quality frames, (3) an Auto-regressive Temporal-aware Decoder that explicitly incorporates temporal cues, and (4) a staged training scheme with rollout distillation over full inference trajectories. Together, these components enable Stream-DiffVSR to produce stable and perceptually coherent videos.

3.1 Diffusion Models Preliminaries

Diffusion Models [31] transform complex data distributions into simpler Gaussian distributions via a forward diffusion process and reconstruct the original data using a learned reverse denoising process. The forward process gradually adds Gaussian noise to the initial data x_0, forming a Markov chain: $q(x_t \mid x_{t-1}) = \mathcal{N}\big(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I\big)$ for $t = 1, \ldots, T$, where β_t denotes a predefined noise schedule. At timestep t, the noised data x_t can be directly sampled from the clean data x_0 as: $x_t = \sqrt{\alpha_t}\, x_0 + \sqrt{1-\alpha_t}\, \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$, where $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$. The reverse process progressively removes noise from x_T, reconstructing the original data x_0 through a learned denoising operation modeled as a Markov chain, i.e., $p_\theta(x_0, \ldots, x_{T-1} \mid x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t)$. Each individual step is parameterized by a neural network-based denoising function $p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(t) I\big)$. Typically, the network predicts the noise component $\epsilon_\theta(x_t, t)$, from which the denoising mean is estimated as $\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\, \epsilon_\theta(x_t, t)\Big)$. Latent Diffusion Models (LDMs) [65] further reduce computational complexity by projecting data into a lower-dimensional latent space using Variational Autoencoders (VAEs), significantly accelerating inference without sacrificing generative quality.

3.2 U-Net Rollout Distillation

We distill a pre-trained Stable Diffusion (SD) $\times 4$ Upscaler [65, 66], originally designed for 50-step inference, into a 4-step variant that balances speed and perceptual quality. To mitigate the training–inference gap of timestep-sampling distillation, we adopt rollout distillation, where the U-Net performs the full 4-step denoising each iteration to obtain a clean latent. Detailed algorithms and implementation are provided in the supplementary material due to page limits.

Unlike conventional distillation that supervises random intermediate timesteps, our method applies loss only on the final denoised latent, ensuring the training trajectory mirrors inference and improving stability and alignment.

Our distillation requires no architectural changes. We train the U-Net by optimizing latent reconstruction with a loss that balances spatial accuracy, perceptual fidelity, and realism:

$$\begin{aligned} \mathcal{L}_{\text{distill}} = & \|\mathbf{z}_{\text{den}} - \mathbf{z}_{\text{gt}}\|_2^2 \\ & + \lambda_{\text{LPIPS}} \cdot \text{LPIPS}\left(D(\mathbf{z}_{\text{den}}), \mathbf{x}_{\text{gt}}\right) \\ & + \lambda_{\text{GAN}} \cdot \mathcal{L}_{\text{GAN}}\left(D(\mathbf{z}_{\text{den}})\right), \end{aligned} \tag{1}$$

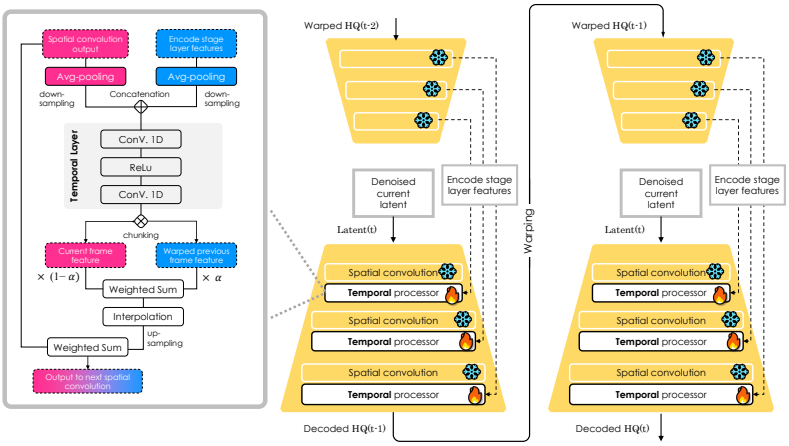

Fig. 2: Overview of Auto-regressive Temporal-aware Decoder. Given the denoised latent and warped previous frame, our decoder enhances temporal consistency using temporal processor modules. This module aligns and fuses these features via interpolation, convolution, and weighted fusion, effectively stabilizing detail reconstruction when decoding into the final RGB frame.

where $\mathbf{z}_{\text{den}}$ and $\mathbf{z}_{\text{gt}}$ are the denoised and ground-truth latent representations. The decoder $D(\cdot)$ maps latent features back to RGB space for perceptual (LPIPS) and adversarial (GAN) loss calculations, encouraging visually realistic outputs.

3.3 Auto-regressive Temporal Guidance

Leveraging temporal information is essential for capturing dynamics and ensuring frame continuity in video super-resolution. However, extensive temporal reasoning often incurs significant computational overhead, increasing per-frame inference time and system latency. Thus, efficient online VSR requires carefully balancing temporal utilization and computational cost to support low-latency processing.

To this end, we propose **Auto-regressive Temporal Guidance (ARTG)**, which enforces temporal coherence during latent denoising. At each timestep t, the U-Net takes both the current noised latent z_t and the warped RGB frame from the previous output, $\hat{x}_{t-1}^{\text{warp}} = \text{Warp}(x_{t-1}^{\text{SR}}, f_{t\leftarrow t-1})$, where $f_{t\leftarrow t-1}$ is the optical flow from frame $t-1$ to t. The denoising prediction is then formulated as:

$$\hat{\epsilon}_\theta = \text{UNet}(z_t, t, \hat{x}_{t-1}^{\text{warp}}), \tag{2}$$

where the warped image $\hat{x}_{t-1}^{\text{warp}}$ serves as temporal conditioning input to guide the denoising process.

We train the **ARTG** module independently using consecutive pairs of low-quality and high-quality frames. The denoising U-Net and decoder are kept fixed during this stage, and the training objective focuses on reconstructing the target latent representation while preserving perceptual quality and visual realism. The

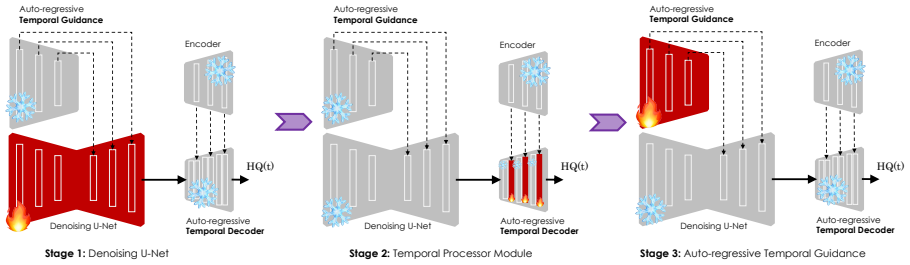

Fig. 3: Training pipeline of Stream-DiffVSR. The training process consists of three sequential stages: (1) Distilling the denoising U-Net to reduce diffusion steps while maintaining perceptual quality with training objective (1); (2) Training the Temporal Processor Module (TPM) within the decoder to enhance temporal consistency at the RGB level with training objective (3); (3) Training the Auto-Regressive Temporal Guidance (ARTG) module to leverage previously restored high-quality frames for improved temporal coherence with training objective (6). Each module is trained separately before integrating them into the final framework.

total loss function is defined as:

$$\begin{aligned}\mathcal{L}_{\text{ARTG}} = &\|\mathbf{z}_{\text{den}} - \mathbf{z}_{\text{gt}}\|_2^2 \\ &+ \lambda_{\text{LPIPS}} \cdot \text{LPIPS}(D(\mathbf{z}_{\text{den}}), \mathbf{x}_{\text{gt}}) \\ &+ \lambda_{\text{GAN}} \cdot \mathcal{L}_{\text{GAN}}(D(\mathbf{z}_{\text{den}})),\end{aligned} \tag{3}$$

where $\mathbf{z}_{\text{den}}$ denotes the denoised latent from DDIM updates with predicted noise $\hat{\boldsymbol{\epsilon}}_\theta$, and $\mathbf{z}_{\text{gt}}$ is the ground-truth latent. The decoder $D(\cdot)$ maps latents to RGB, producing $D(\mathbf{z}_{\text{den}})$ for comparison with the ground-truth image $\mathbf{x}_{\text{gt}}$. The latent ℓ_2 loss enforces alignment, the perceptual loss preserves visual fidelity, and the adversarial loss promotes realism. This design leverages only past frames to propagate temporal context, improving consistency without additional latency.

3.4 Auto-regressive Temporal-aware Decoder

Although the **Auto-regressive Temporal Guidance (ARTG)** improves temporal consistency in the latent space, the features produced by the Stable Diffusion $\times 4$ Upscaler remain at one-quarter of the target resolution. This mismatch may introduce decoding artifacts or misalignment in dynamic scenes.

To address this issue, we propose an **Auto-regressive Temporal-aware Decoder** that incorporates temporal context into decoding to enhance spatial fidelity and temporal consistency. At timestep t, the decoder takes the denoised latent $\mathbf{z}_t^{\text{den}}$ and the aligned feature $\hat{\mathbf{f}}_{t-1}$ derived from the previous super-resolved frame. Specifically, we compute:

$$\hat{\mathbf{x}}_{t-1}^{\text{warp}} = \text{Warp}(\mathbf{x}_{t-1}^{\text{SR}}, f_{t \leftarrow t-1}), \quad \hat{\mathbf{f}}_{t-1} = \text{Enc}(\hat{\mathbf{x}}_{t-1}^{\text{warp}}), \tag{4}$$

where $\mathbf{x}_{t-1}^{\text{SR}}$ is the previously generated RGB output, $f_{t \leftarrow t-1}$ is the optical flow from frame $t-1$ to t, and $\text{Enc}(\cdot)$ is a frozen encoder that projects the warped image into the latent feature space.

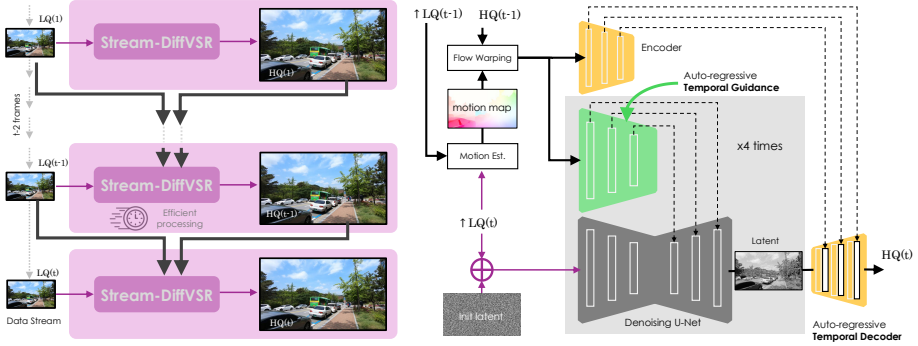

Fig. 4: Overview of our pipeline. Given a low-quality (LQ) input frame, we first initialize its latent representation and employ an autoregressive diffusion model composed of a distilled denoising U-Net, autoregressive temporal Guidance, and an autoregressive temporal Decoder. Temporal guidance utilizes flow-warped high-quality (HQ) results from the previous frame to condition the current frame's latent denoising and decoding processes, significantly improving perceptual quality and temporal consistency in an efficient, online manner.

The decoder then synthesizes the current frame using:

$$\mathbf{x}_t^{\mathrm{SR}} = \mathrm{Decoder}(\mathbf{z}_t^{\mathrm{den}}, \hat{\mathbf{f}}_{t-1}). \tag{5}$$

We adopt a multi-scale fusion strategy inside the decoder to combine current spatial information and prior temporal features across multiple resolution levels, as illustrated in Fig. 2. This design helps reinforce temporal coherence while recovering fine spatial details.

Temporal Processor Module (TPM). We integrate TPM after each spatial convolutional layer in the decoder to explicitly inject temporal coherence, enhancing stability and continuity of reconstructed frames. These modules utilize latent features from the current frame and warped features from the previous frame, optimizing temporal consistency independently from spatial reconstruction. Our training objective for the TPM is defined as:

$$\begin{aligned}\mathcal{L}_{\mathrm{TPM}} = \mathcal{L}_{\mathrm{rec}}(\mathbf{x}_t^{\mathrm{rec}}, \mathbf{x}_t^{\mathrm{GT}}) \\ + \lambda_{\mathrm{flow}} \left\| \mathrm{OF}(\mathbf{x}_t^{\mathrm{rec}}, \mathbf{x}_{t-1}^{\mathrm{rec}}) - \mathrm{OF}(\mathbf{x}_t^{\mathrm{GT}}, \mathbf{x}_{t-1}^{\mathrm{GT}}) \right\|_2^2 \\ + \lambda_{\mathrm{GAN}} \mathcal{L}_{\mathrm{GAN}}(\mathbf{x}_t^{\mathrm{rec}}) \\ + \lambda_{\mathrm{LPIPS}} \mathrm{LPIPS}(\mathbf{x}_t^{\mathrm{rec}}, \mathbf{x}_t^{\mathrm{GT}}),\end{aligned} \tag{6}$$

where $\mathbf{x}_t^{\mathrm{SR}} \in \mathbb{R}^{3\times H\times W}$ is the predicted frame at time t, and $\mathbf{x}_t^{\mathrm{GT}}$ is the ground-truth frame. The reconstruction loss $\mathcal{L}_{\mathrm{rec}} = \mathrm{SmoothL1}(\mathbf{x}_t^{\mathrm{rec}}, \mathbf{x}_t^{\mathrm{GT}})$ enforces spatial fidelity, the adversarial loss $\mathcal{L}_{\mathrm{GAN}}$ improves realism, and the optical-flow term $\mathrm{OF}(\cdot,\cdot)$ reduces temporal discrepancies, yielding consistent and perceptually faithful outputs.

3.5 Training and Inference Stages

Our training pipeline consists of three independent stages (Fig. 3), while our inference process and the Auto-Regressive Diffusion-based VSR algorithm are illustrated in Fig. 4 and detailed in the appendix due to page constraints, respectively.

Distilling the Denoising U-Net. We first distill the denoising U-Net using pairs of low-quality (LQ) and high-quality (HQ) frames to optimize per-frame super-resolution and latent-space consistency.

Training the Temporal Processor Module (TPM). In parallel, we train the Temporal Processor Module (TPM) in the decoder using ground-truth frames, keeping all other weights fixed. This enhances the decoder's capability to incorporate temporal information into the final RGB reconstruction.

Training Auto-regressive Temporal Guidance. After training and freezing the U-Net and decoder, we train the ARTG, which leverages flow-aligned previous outputs to enhance temporal coherence without degrading spatial quality. This staged training strategy progressively refines spatial fidelity, latent consistency, and temporal smoothness in a decoupled manner.

Inference. Given a sequence of low-quality (LQ) frames, our method auto-regressively generates high-quality (HQ) outputs. For each frame t, denoising is conditioned on the previous output HQ_{t-1}, warped via optical flow to capture temporal motion. To balance quality and efficiency, we employ a **4-step DDIM** scheme using a distilled U-Net. By combining motion alignment with reduced denoising steps, our inference pipeline achieves efficient and stable temporal consistency.

4 Experiment

Due to space limitations, we provide the experimental setup in the appendix.

We quantitatively compare Stream-DiffVSR with state-of-the-art VSR methods on REDS4 [61], Vimeo-90K-T [102] and VideoLQ [8], covering diverse scene content and motion characteristics. Tabs. 2, 4 and 7 report perceptual and temporal-consistency metrics across CNN-, Transformer-, and diffusion-based approaches under both bidirectional (offline) and unidirectional (online) settings, including comparisons against memory-intensive diffusion-based baselines. Tabs. 3, 5 and 8 report the corresponding runtime, Time-to-First-Frame, and peak memory for each method, further highlighting Stream-DiffVSR's quality–latency trade-off under practical memory budgets.

On REDS4, Stream-DiffVSR achieves strong perceptual quality (LPIPS=0.099) compared with representative CNN (BasicVSR++ [7], RealBasicVSR [8]), Transformer (RVRT [49]), and diffusion-based baselines (StableVSR [68], MGLD-VSR [104]), while maintaining competitive temporal consistency (tLP=4.198, tOF=3.638). Importantly, these results are obtained with substantially lower runtime (0.328s/frame vs. 43–46s/frame for diffusion-based baselines), matching the low-latency objective.

Table 2: Quantitative comparison on the REDS4 dataset. We compare bidirectional/offline and unidirectional/online methods. ↑/↓ indicate higher/lower is better. **Dir.**: **B** for bidirectional/offline, **U** for unidirectional/online. **tLP** and **tOF** are scaled by 100× and 10×. Best and second-best results within each direction group are marked in red and blue. Gray-shaded entries are reported from FlashVSR.

Dir.	Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑	NIQE↓	NRQM↑	BRISQUE↓	VMAF↑	tLP↓	tOF↓
Bidirectional/Offline Methods												
-	Bicubic	25.501	0.712	0.460	0.187	27.362	7.360	3.459	60.256	46.132	21.603	4.241
B	BasicVSR++	32.386	0.907	0.132	0.069	67.013	3.850	6.363	38.641	99.580	9.017	2.490
B	RealBasicVSR	27.042	0.778	0.134	0.065	67.033	2.530	6.769	18.046	88.197	6.422	4.759
B	RVRT	32.701	0.911	0.130	0.067	67.251	3.793	6.366	38.038	99.681	9.133	2.421
B	MIA-VSR	32.790	0.912	0.123	0.064	68.140	3.742	6.451	37.099	99.669	8.870	2.354
B	StableVSR	27.928	0.793	0.102	0.047	67.058	2.713	6.960	16.249	95.303	5.755	2.742
B	MGLD-VSR	26.53	0.749	0.151	0.065	66.081	2.972	6.701	15.291	78.255	18.139	5.910
B	UAV	23.077	0.608	0.495	0.249	30.912	5.917	3.666	39.751	20.600	30.546	13.582
B	UAV	24.840	0.644	0.41	–	53.000	3.104	–	–	–	–	–
B	SeedVR2	22.774	0.661	0.246	0.111	65.011	2.812	6.707	29.948	64.447	30.296	12.79
B	SeedVR2	24.830	0.704	0.302	–	61.830	3.066	–	–	–	–	–
B	DOVE	25.027	0.692	0.291	0.149	54.060	4.087	5.753	25.606	58.329	10.260	10.960
B	VEnhancer	20.712	0.560	0.384	0.156	56.932	4.142	5.567	24.906	24.743	32.913	16.643
Unidirectional/Online Methods												
U	TMP	30.672	0.871	0.194	0.090	63.818	4.378	5.796	43.394	98.586	10.424	2.480
U	RealViformer	26.763	0.761	0.129	0.065	64.585	2.731	7.028	17.272	60.509	11.261	4.037
U	StableVSR*	27.174	0.763	0.111	0.051	66.428	2.572	6.944	15.805	88.675	11.107	3.925
U	FlashVSR	21.484	0.569	0.283	0.124	62.178	6.574	6.574	25.821	40.556	15.69	20.548
U	FlashVSR	23.920	0.649	0.343	–	68.97	2.425	–	–	–	–	–
U	FlashVSR-tiny	21.887	0.577	0.300	0.141	55.204	3.463	6.188	28.589	38.634	17.838	23.992
U	FlashVSR-tiny	24.110	0.651	0.343	–	67.43	2.680	–	–	–	–	–
U	Ours	27.256	0.766	0.099	0.062	65.586	3.114	7.055	17.117	88.751	4.265	3.620

On Vimeo-90K-T, Stream-DiffVSR similarly attains favorable perceptual performance (LPIPS=0.056, DISTS=0.105) and improved temporal consistency (tLP=4.307, tOF=2.689) with a runtime of 0.041s/frame, supporting efficient causal inference for online/streaming usage.

4.1 Generalization to Mixed Degradations

The REDS4 and Vimeo-90K-T evaluations above use bicubic downsampling, which isolates throughput, latency, and temporal-consistency comparisons from confounds introduced by stochastic degradation pipelines. To examine whether this controlled setting limits Stream-DiffVSR to a narrow degradation regime, we additionally fine-tune our model on REDS using the degradation pipeline introduced by DOVE [15] to synthesize realistic low-quality/high-quality pairs covering blur, noise, and compression artifacts. Tab. 6 compares the fine-tuned model against real-world-oriented baselines under this setting. Stream-DiffVSR remains competitive in perceptual quality (PSNR, SSIM, LPIPS) and attains the best or second-best temporal consistency (tLP, tOF) among online methods, indicating that our architecture and training procedure are not tied to the bicubic setting used elsewhere in this paper. We further evaluate this fine-tuned model on the real-world VideoLQ in Tabs. 7 and 8, providing additional evidence of generalization in the absence of paired ground truth.

In addition to speed, Stream-DiffVSR achieves a markedly lower memory footprint. We emphasize that Stream-DiffVSR performs strictly frame-by-frame

Table 3: Efficiency comparison on the REDS4 dataset. Runtime, TTFF, and peak memory are measured per 720p frame; memory-intensive methods are measured on an RTX Pro 6000, others on an RTX 4090. **TTFF** (time-to-first-frame) is the wait from the first input frame to the first output frame, measured over 100-frame sequences; for offline/bidirectional methods this scales with sequence length. Best and second-best results within each GPU/direction group are marked in red and blue.

Dir.	Method	Runtime (s)↓	TTFF (s)↓	Mem (GB)↓	Dir.	Method	Runtime (s)↓	TTFF (s)↓	Mem (GB)↓
Measured on an NVIDIA RTX 4090 GPU									
B	BasicVSR++	0.098	9.8	-	U	TMP	0.041	0.041	-
B	RealBasicVSR	0.064	6.4	-	U	RealViformer	0.099	9.9	-
B	RVRT	0.498	2.49	-	U	StableVSR*	46.2	4620	-
B	MIA-VSR	0.768	76.8	-	U	Ours	0.328	0.328	-
B	StableVSR	46.2	4620	-					
B	MGLD-VSR	43.6	218	-					
Measured on an NVIDIA RTX Pro 6000 GPU									
B	UAV	6.017	601.740	57.094	U	FlashVSR	0.283	1.698	19.630
B	SeedVR2	0.479	47.840	69.592	U	FlashVSR-tiny	0.157	0.942	12.390
B	DOVE	0.760	75.983	42.208	U	Ours	0.243	0.243	19.580
B	VEnhancer	0.672	671.650	53.200					

Table 4: Quantitative comparison on the Vimeo-90K-T dataset.

Dir.	Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑	NIQE↓	NRQM↑	BRISQUE↓	VMAF↑	tLP↓	tOF↓
Bidirectional/Offline Methods												
-	Bicubic	29.282	0.864	0.297	0.209	23.433	8.735	3.588	61.714	42.928	11.606	2.49
B	BasicVSR++	37.479	0.956	0.098	0.117	51.940	7.077	5.509	47.792	92.905	4.691	1.57
B	RealBasicVSR	29.388	0.857	0.156	0.149	56.986	5.069	7.413	23.822	79.781	10.947	3.46
B	RVRT	37.815	0.955	0.093	0.105	49.937	7.205	5.393	48.352	94.660	4.873	1.429
B	MIA-VSR	37.598	0.957	0.086	0.101	51.402	7.116	5.569	47.865	95.113	4.696	1.419
B	StableVSR	31.823	0.878	0.095	0.111	54.582	4.745	7.265	20.039	86.936	26.224	3.108
B	MGLD-VSR	29.651	0.865	0.151	0.137	57.788	5.340	7.217	20.761	71.509	12.550	4.661
B	UAV	25.263	0.72	0.226	0.177	55.562	5.602	7.247	24.766	46.489	9.546	7.576
B	SeedVR2	24.585	0.719	0.177	0.152	58.809	4.204	8.107	20.447	41.430	21.191	12.460
B	DOVE	28.611	0.834	0.12	0.116	57.749	5.734	7.182	23.388	77.997	8.214	4.387
B	VEnhancer	23.712	0.712	0.234	0.189	57.822	6.175	7.910	34.906	60.920	32.913	16.643
Unidirectional/Online Methods												
U	TMP	36.482	0.946	0.109	0.118	48.374	7.368	5.096	49.192	92.001	4.870	1.603
U	RealViformer	30.291	0.877	0.130	0.140	53.107	5.515	6.711	24.628	54.689	8.232	2.769
U	StableVSR*	31.729	0.875	0.072	0.113	54.447	4.698	7.280	19.836	86.162	30.858	3.144
U	FlashVSR	26.114	0.789	0.136	0.118	59.451	5.249	7.317	29.288	78.575	19.018	9.351
U	FlashVSR-tiny	26.154	0.791	0.136	0.121	57.845	5.249	7.317	29.288	76.376	19.881	9.793
U	Ours	32.593	0.900	0.056	0.105	52.755	4.403	7.672	29.297	88.311	4.307	2.689

causal inference, whereas FlashVSR [133] adopts chunk-wise streaming. Consequently, FlashVSR's per-frame runtime can appear low, but its time-to-first-frame is bounded by chunk buffering, while Stream-DiffVSR avoids sequence-level waiting and provides immediate frame-level response. As shown in Tabs. 3, 5 and 8, prior diffusion-based VSR methods such as DOVE [15], SeedVR2 [87], and Upscale-A-Video(UAV) [129] often require large GPU memory budgets (e.g., exceeding 42 GB on REDS4/VideoLQ under our settings) or run into OOM on single RTX 4090. In contrast, Stream-DiffVSR operates within 19.58 GB and 22.80 GB peak memory on REDS4 and VideoLQ, respectively, and consistently achieves lower runtime than every memory-intensive baseline across both datasets, with speedups ranging from approximately 2× over the closest com-

Table 5: Efficiency comparison on the Vimeo-90K-T dataset. Runtime, TTFF, and peak memory are measured per 448×256 frame.

Dir.	Method	Runtime (s)↓	TTFF (s)↓	Mem (GB)↓	Dir.	Method	Runtime (s)↓	TTFF (s)↓	Mem (GB)↓
Measured on an NVIDIA RTX 4090 GPU									
B	BasicVSR++	0.012	0.084	-	U	TMP	0.006	0.006	-
B	RealBasicVSR	0.008	0.056	-	U	RealViformer	0.013	0.091	-
B	RVRT	0.061	0.305	-	U	StableVSR*	5.749	40.243	-
B	MIA-VSR	0.096	0.672	-	U	Ours	0.041	0.041	-
B	StableVSR	5.749	40.243	-					
B	MGLD-VSR	5.426	27.130	-					
Measured on an NVIDIA RTX Pro 6000 GPU									
B	UAV	0.499	3.490	37.987	U	FlashVSR	0.025	0.150	5.340
B	SeedVR2	0.039	0.270	41.366	U	FlashVSR-tiny	0.019	0.114	4.280
B	DOVE	0.223	1.560	31.934	U	Ours	0.036	0.036	9.720
B	VEnhancer	1.834	12.840	35.855					

Table 6: REDS4 comparison under multi-degradations against real-world-oriented baselines.

Dir.	Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑	tLP↓	tOF↓
B	MGLD-VSR	23.305	0.594	0.246	0.112	61.139	37.737	39.039
B	UAV	21.721	0.534	0.419	0.22	45.714	14.687	30.589
B	SeedVR2	22.043	0.587	0.28	0.111	62.213	30.273	14.906
B	DOVE	23.128	0.608	0.383	0.207	52.498	11.129	16.869
U	FlashVSR	21.163	0.527	0.287	0.124	63.891	14.723	24.353
U	FlashVSR-tiny	21.462	0.536	0.307	0.142	57.593	20.221	27.106
U	Ours	22.377	0.549	0.22	0.128	63.222	14.797	16.785

Table 7: Quantitative comparison on the VideoLQ dataset. VideoLQ is a real-world benchmark without ground-truth references; we therefore report no-reference quality metrics only. Best and second-best results within each direction group are marked in red and blue. Gray text indicates entries reported from FlashVSR.

Dir.	Method	NIQE↓	NRQM↑	BRISQUE↓	Dir.	Method	NIQE↓	NRQM↑	BRISQUE↓
B	BasicVSR++	5.909	3.745	56.800	U	TMP	6.751	3.511	59.841
B	RealBasicVSR	3.973	6.095	30.158	U	RealViformer	4.070	6.066	28.266
B	RVRT	6.939	3.493	60.557	U	StableVSR*	3.982	6.122	23.814
B	MIA-VSR	5.860	3.810	58.513	U	FlashVSR	–	–	–
B	StableVSR	3.973	6.154	22.973	U	FlashVSR	3.803	–	–
B	MGLD-VSR	4.163	5.761	29.497	U	FlashVSR-tiny	4.569	5.164	42.514
B	UAV	6.299	3.652	44.139	U	FlashVSR-tiny	4.070	–	–
B	UAV	4.889	–	–	U	Ours	3.929	6.140	23.176
B	SeedVR2	4.661	5.523	37.975					
B	SeedVR2	5.205	–	–					
B	DOVE	5.090	5.214	36.631					
B	VEnhancer	6.221	3.85	48.1					

Table 8: Efficiency comparison on the VideoLQ dataset under a single NVIDIA RTX Pro 6000 GPU.

Dir.	Method	Runtime (s)↓	TTFF (s)↓	Mem (GB)↓	Dir.	Method	Runtime (s)↓	TTFF (s)↓	Mem (GB)↓
B	VEnhancer	9.544	477.237	47.985	U	FlashVSR	–	–	OOM
B	SeedVR2	1.126	56.28	76.094	U	FlashVSR-tiny	0.204	1.224	44.180
B	UAV	8.081	404.07	55.897	U	Ours	0.454	0.454	22.800
B	DOVE	1.735	86.774	46.344					

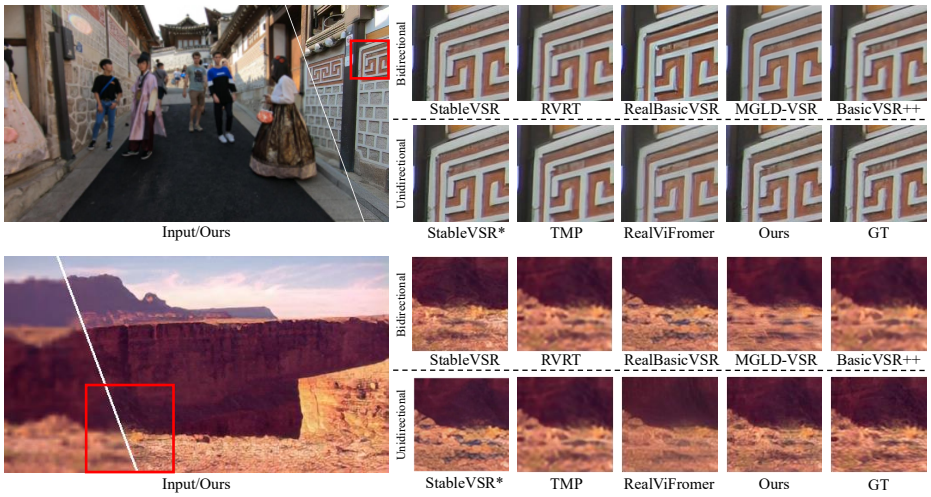

Fig. 5: Qualitative comparison on REDS4 and Vimeo-90K-T datasets. Our method demonstrates superior visual quality with sharper details compared to unidirectional methods (TMP [120], RealViformer [119]) and competitive performance against bidirectional methods (StableVSR [68], MGLD-VSR [104], RVRT [49], BasicVSR++ [7], RealBasicVSR [8]).

Table 9: Ablation study of temporal modules in Stream-DiffVSR.

Component	LPIPS↓	DISTS↓	MUSIQ↑	NIQE↓	NRQM↑	BRISQUE↓	tLP↓	tOF↓	WarpErr↓	FPS↑	TTFF(s)↓
Per-frame	0.099	0.071	65.981	3.249	6.969	21.655	7.261	4.201	25.668	4.807	0.208
w/o ARTG	0.117	0.070	63.347	3.194	6.980	19.027	6.132	3.910	16.598	4.255	0.235
w/o TPM	0.116	0.078	67.110	3.197	7.007	20.279	12.847	4.639	21.990	3.33	0.300
TPM (unwarped)	0.122	0.082	63.849	3.201	7.159	14.063	12.846	5.689	17.143	3.125	0.320
Ours	0.099	0.062	65.586	3.111	7.256	17.667	4.265	3.620	14.909	3.408	0.328

petitor (SeedVR2) to over 20× over the slowest baselines (UAV, VEnhancer), underscoring its efficiency and deployability.

Results on VideoLQ further show consistent perceptual and temporal performance, indicating stable behavior across diverse content and motion patterns.

4.2 Qualitative Comparisons

We provide qualitative comparisons in Fig. 5, where Stream-DiffVSR generates sharper details and fewer artifacts than prior methods. Additional visualizations of temporal consistency and flow coherence are included in the supplemental material. A qualitative comparison with DOVE, SeedVR2, Upscale-A-Video (UAV) is included in the appendix and supplementary.

4.3 Ablation Study

We ablate key components of Stream-DiffVSR including denoising-step reduction, ARTG, TPM, timestep selection, and training-stage combinations on REDS4 to ensure consistent evaluation of perceptual quality and temporal stability.

Table 10: Ablation study on training strategy.

Stage combination	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑	tLP↓	tOF↓	WarpErr↓
stage 1 and 2	25.442	0.702	0.156	0.100	67.528	21.781	6.37	27.307
stage 1 and 3	26.307	0.753	0.121	0.077	64.902	13.094	4.09	21.689
stage 2 and 3	26.906	0.758	0.132	0.077	64.751	10.510	4.225	15.726
All stage jointly	26.135	0.736	0.124	0.073	67.35	17.816	4.596	24.298
Separate (Ours)	27.256	0.766	0.099	0.062	65.586	4.265	3.620	14.909

Table 11: Ablation study on denoising step count within Stream-DiffVSR. We evaluate 50, 10, 1, and 4 steps. Our 4-step design achieves a favorable balance between perceptual quality and runtime.

Step(s)	LPIPS↓	DISTS↓	MUSIQ↑	NIQE↓	NRQM↑	BRISQUE↓	tLP↓	tOF↓	Runtime (s)↓
50	0.102	0.068	66.061	2.804	7.026	9.925	18.798	3.826	3.460
10	0.122	0.072	64.900	2.869	6.917	12.461	9.990	3.625	0.718
1	0.138	0.076	63.915	3.843	6.984	29.552	9.899	3.882	0.106
4 (Ours)	0.099	0.062	65.586	3.111	7.256	17.667	4.265	3.620	0.328

Table 12: Ablation study on Rollout Training. Comparison of random timestep distillation vs. rollout training across fidelity and perceptual metrics.

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑	GPU Hours↓
Random Timestep Selection	26.27	0.743	0.099	0.071	65.981	60.5
Rollout Distillation	26.36	0.753	0.095	0.075	66.391	21

Fig. 6: Ablation study on the Temporal Processor Module (TPM). Integrating TPM improves motion stability and reduces temporal artifacts by leveraging warped previous-frame features, enhancing temporal consistency in video super-resolution.

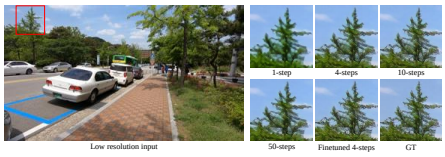

Fig. 7: Ablation study on inference steps. The 4-step model yields the best quality–efficiency trade-off, validating our distillation strategy.

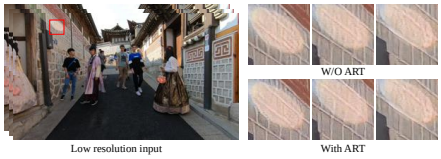

Fig. 8: Ablation study on Autoregressive Temporal Guidance (ARTG). ARTG enhances temporal consistency and perceptual quality by leveraging warped previous frames, reducing flickering, and improving structural coherence.

We perform ablation studies on training strategies in Tab. 12 and Tab. 10. For stage-wise training, partial or joint training yields inferior results, while our separate stage-wise scheme achieves the best trade-off across fidelity, perceptual, and temporal metrics. For distillation, rollout training outperforms random timestep selection in both quality and efficiency, reducing training cost from 60.5 to 21 GPU hours on 4×A6000 GPUs.

We assess the runtime–quality trade-off by varying DDIM inference steps while keeping model weights fixed. As shown in Tab. 11 and Fig. 7, fewer steps increase efficiency but reduce perceptual quality, whereas more steps improve fidelity with higher latency. A 4-step setting provides the best balance.

Tab. 9 and Fig. 8 show the effectiveness of ARTG and TPM. The *per-frame* baseline uses only the distilled U-Net with both ARTG and TPM disabled. In the ablation labels, *w/o* indicates that a module is fully removed; for instance, *TPM (unwarp)* feeds TPM the previous HR frame without flow-based warping, removing motion alignment. ARTG improves perceptual quality and temporal consistency. TPM further enhances temporal coherence through temporal-feature warping and fusion, yielding additional gains in tLP100. These results highlight the complementary roles of latent-space guidance and decoder-side temporal modeling.

5 Conclusion

We propose Stream-DiffVSR, an efficient online video super-resolution framework using diffusion models. By integrating a distilled U-Net, Auto-Regressive Temporal Guidance, and Temporal-aware Decoder, Stream-DiffVSR achieves superior perceptual quality, temporal consistency, and practical inference speed for low-latency applications.

Limitations. Stream-DiffVSR remains heavier than CNN and Transformer models, and its use of optical flow can introduce artifacts under fast motion. Its auto-regressive design may also degrade the earliest frames, motivating stronger initialization to reduce cold-start effects. Robustness to diverse, unknown degradations remains future work.

Acknowledgements

This research was funded by the National Science and Technology Council, Taiwan, under Grants NSTC 112-2222-E-A49-004-MY2, 113-2628-E-A49-023-, and 115-2222-E-A49-005-MY2. The authors are grateful to Google, NVIDIA, and MediaTek Inc. for their generous donations. Yu-Lun Liu acknowledges the Yushan Young Fellow Program by the MOE in Taiwan. Yu-Chih Chen acknowledges the Yushan Young Fellow Program by the MOE in Taiwan (Grant MOE-114-YSFEE-0010-008-P1).

References

1. Lu et al., C.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. NeurIPS **35**, 5775–5787 (2022)
2. Luo et al., S.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
3. Bai, W., Xu, S., Ren, Y., Hao, J., Sun, M., Chen, W., Sun, H.: Instantvir: Real-time video inverse problem solver with distilled diffusion prior. arXiv preprint arXiv:2511.14208 (2025)
4. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6228–6237 (2018)
5. Cao, Y., Wang, C., Song, C., Tang, Y., Li, H.: Real-time super-resolution system of 4k-video based on deep learning. In: 2021 IEEE 32nd International Conference on Application-specific Systems, Architectures and Processors (ASAP). pp. 69–76. IEEE (2021)
6. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: Basicvsr: The search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4947–4956 (2021)
7. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: 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 (2022)
8. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Investigating tradeoffs in real-world video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5962–5971 (2022)
9. Chang, K.C., Wang, R., Lin, H.J., Liu, Y.L., Chen, C.P., Chang, Y.L., Chen, H.T.: Learning camera-aware noise models. In: European Conference on Computer Vision. pp. 343–358. Springer (2020)
10. Chao, C.H., Sun, W.F., Cheng, B.W., Lo, Y.C., Chang, C.C., Liu, Y.L., Chang, Y.L., Chen, C.P., Lee, C.Y.: Denoising likelihood score matching for conditional score-based data generation. arXiv preprint arXiv:2203.14206 (2022)
11. Chen, B., Li, G., Wu, R., Zhang, X., Chen, J., Zhang, J., Zhang, L.: 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 (2025)
12. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. Advances in Neural Information Processing Systems **37**, 24081–24125 (2024)
13. Chen, J., Xue, S., Zhao, Y., Yu, J., Paul, S., Chen, J., Cai, H., Han, S., Xie, E.: Sana-sprint: One-step diffusion with continuous-time consistency distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16185–16195 (2025)
14. Chen, R., Mu, Y., Zhang, Y.: High-order relational generative adversarial network for video super-resolution. Pattern Recognition **146**, 110059 (2024)
15. Chen, Z., Zou, Z., Zhang, K., Su, X., Yuan, X., Guo, Y., Zhang, Y.: Dove: Efficient one-step diffusion model for real-world video super-resolution. arXiv preprint arXiv:2505.16239 (2025)

16. Chu, E., Huang, T., Lin, S.Y., Chen, J.C.: Medm: Mediating image diffusion models for video-to-video translation with temporal correspondence guidance. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1353–1361 (2024)
17. Conde, M.V., Lei, Z., Li, W., Bampis, C., Katsavounidis, I., Timofte, R., Luo, Q., Song, J., Jiang, L., Lei, H., et al.: Aim 2024 challenge on efficient video super-resolution for av1 compressed content. In: European Conference on Computer Vision. pp. 304–325. Springer (2024)
18. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp. 764–773 (2017)
19. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
20. Fuoli, D., Danelljan, M., Timofte, R., Van Gool, L.: Fast online video super-resolution with deformable attention pyramid. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1735–1744 (2023)
21. Fuoli, D., Gu, S., Timofte, R.: Efficient video super-resolution through recurrent latent space propagation. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 3476–3485. IEEE (2019)
22. Gao, K., Shi, J., Zhang, H., Wang, C., Xiao, J., Chen, L.: Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing. arXiv preprint arXiv:2411.16375 (2024)
23. Gao, S., Liu, X., Zeng, B., Xu, S., Li, Y., Luo, X., Liu, J., Zhen, X., Zhang, B.: Implicit diffusion models for continuous super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10021–10030 (2023)
24. Geng, Z., Pokle, A., Luo, W., Lin, J., Kolter, J.Z.: Consistency models made easy. arXiv preprint arXiv:2406.14548 (2024)
25. Ghasemabadi, A., Janjua, M.K., Salameh, M., Niu, D.: Learning truncated causal history model for video restoration. Advances in Neural Information Processing Systems **37**, 27584–27615 (2024)
26. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
27. Guo, Z., Li, W., Loy, C.C.: Generalizable implicit motion modeling for video frame interpolation. Advances in Neural Information Processing Systems **37**, 63747–63770 (2024)
28. Han, J., Sim, G., Kim, G., Lee, H.S., Choi, K., Han, Y., Cho, S.: Dc-vsr: Spatially and temporally consistent video super-resolution with video diffusion prior. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025)
29. He, J., Xue, T., Liu, D., Lin, X., Gao, P., Lin, D., Qiao, Y., Ouyang, W., Liu, Z.: Venhancer: Generative space-time enhancement for video generation. arXiv preprint arXiv:2407.07667 (2024)
30. Heek, J., Hoogeboom, E., Salimans, T.: Multistep consistency models. arXiv preprint arXiv:2403.06807 (2024)
31. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

32. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. Journal of Machine Learning Research **23**(47), 1–33 (2022)
33. Hsiao, C.W., Liu, Y.L., Yang, C.K., Kuo, S.P., Jou, K., Chen, C.P.: Ref-ldm: A latent diffusion model for reference-based face image restoration. Advances in Neural Information Processing Systems **37**, 74840–74867 (2024)
34. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
35. Isobe, T., Jia, X., Gu, S., Li, S., Wang, S., Tian, Q.: Video super-resolution with recurrent structure-detail network. arXiv preprint arXiv:2008.00455 (2020)
36. Jeelani, M., Cheema, N., Illgner-Fehns, K., Slusallek, P., Jaiswal, S., et al.: Expanding synthetic real-world degradations for blind video super resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1199–1208 (2023)
37. Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., Huang, F.: Real-world super-resolution via kernel estimation and noise injection. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 466–467 (2020)
38. Jiang, Y., Nawała, J., Feng, C., Zhang, F., Zhu, X., Sole, J., Bull, D.: Rtsr: A real-time super-resolution model for av1 compressed content. In: 2025 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 1–5. IEEE (2025)
39. Jin, Y., Sun, Z., Li, N., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. arXiv preprint arXiv:2410.05954 (2024)
40. Kim, D., Lai, C.H., Liao, W.H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., Ermon, S.: Consistency trajectory models: Learning probability flow ode trajectory of diffusion. arXiv preprint arXiv:2310.02279 (2023)
41. Kim, J., Kang, J., Choi, J., Han, B.: Fifo-diffusion: Generating infinite videos from text without training. Advances in Neural Information Processing Systems **37**, 89834–89868 (2024)
42. Lei, C., Xing, Y., Chen, Q.: Blind video temporal consistency via deep video prior. Advances in Neural Information Processing Systems **33**, 1083–1093 (2020)
43. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022)
44. Li, W., Tao, X., Guo, T., Qi, L., Lu, J., Jia, J.: Mucan: Multi-correspondence aggregation network for video super-resolution. arXiv preprint arXiv:2007.11803 (2020)
45. Li, X., Liu, Y., Cao, S., Chen, Z., Zhuang, S., Chen, X., He, Y., Wang, Y., Qiao, Y.: Diffvsr: Enhancing real-world video super-resolution with diffusion models for advanced visual quality and temporal consistency. arXiv preprint arXiv:2501.10110 (2025)
46. Liang, F., Wu, B., Wang, J., Yu, L., Li, K., Zhao, Y., Misra, I., Huang, J.B., Zhang, P., Vajda, P., et al.: Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8207–8216 (2024)
47. Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L.: Vrt: A video restoration transformer. arXiv preprint arXiv:2201.12288 (2022)
48. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)

49. Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Gool, L.V.: Recurrent video restoration transformer with guided deformable attention. Advances in Neural Information Processing Systems **35**, 378–393 (2022)
50. Lin, S., Wang, A., Yang, X.: Sdxl-lightning: Progressive adversarial diffusion distillation. arXiv preprint arXiv:2402.13929 (2024)
51. Liu, H., Liu, S., Zhou, Z., Xu, M., Xie, Y., Han, X., Pérez, J.C., Liu, D., Kahatapitiya, K., Jia, M., et al.: Mardini: Masked autoregressive diffusion for video generation at scale. arXiv preprint arXiv:2410.20280 (2024)
52. Liu, K.H., Yang, C.K., Chen, M.H., Liu, Y.L., Lin, Y.Y.: Corrfill: Enhancing faithfulness in reference-based inpainting with correspondence guidance in diffusion models. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1618–1627. IEEE (2025)
53. Liu, X., Zhang, X., Ma, J., Peng, J., et al.: Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In: The Twelfth International Conference on Learning Representations (2023)
54. Liu, Y., Pan, J., Li, Y., Dong, Q., Zhu, C., Guo, Y., Wang, F.: Ultravsr: Achieving ultra-realistic video super-resolution with efficient one-step diffusion space. arXiv preprint arXiv:2505.19958 (2025)
55. Liu, Y.L., Lai, W.S., Yang, M.H., Chuang, Y.Y., Huang, J.B.: Learning to see through obstructions with layered decomposition. IEEE transactions on pattern analysis and machine intelligence **44**(11), 8387–8402 (2021)
56. Liu, Y.L., Liao, Y.T., Lin, Y.Y., Chuang, Y.Y.: Deep video frame interpolation using cyclic frame generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8794–8802 (2019)
57. Lu, C., Song, Y.: Simplifying, stabilizing and scaling continuous-time consistency models. arXiv preprint arXiv:2410.11081 (2024)
58. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. Machine Intelligence Research pp. 1–22 (2025)
59. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
60. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14297–14306 (2023)
61. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Lee, K.M.: Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In: CVPR Workshops (June 2019)
62. Pan, J., Bai, H., Dong, J., Zhang, J., Tang, J.: Deep blind video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4811–4820 (2021)
63. Qu, Z., Jiang, X., Yang, Y., Li, D., Zhao, C.: Online video quality enhancement with spatial-temporal look-up tables. In: European Conference on Computer Vision. pp. 449–465. Springer (2024)
64. Ren, Y., Xia, X., Lu, Y., Zhang, J., Wu, J., Xie, P., Wang, X., Xiao, X.: Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. Advances in neural information processing systems **37**, 117340–117362 (2024)

65. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
66. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
67. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
68. Rota, C., Buzzelli, M., van de Weijer, J.: Enhancing perceptual quality in video super-resolution through temporally-consistent detail synthesis using diffusion models. In: European Conference on Computer Vision. pp. 36–53. Springer (2024)
69. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE transactions on pattern analysis and machine intelligence **45**(4), 4713–4726 (2022)
70. Sajjadi, M.S., Vemulapalli, R., Brown, M.: Frame-recurrent video super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6626–6634 (2018)
71. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
72. Sauer, A., Boesel, F., Dockhorn, T., Blattmann, A., Esser, P., Rombach, R.: Fast high-resolution image synthesis with latent adversarial diffusion distillation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
73. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: European Conference on Computer Vision. pp. 87–103. Springer (2024)
74. Shi, S., Gu, J., Xie, L., Wang, X., Yang, Y., Dong, C.: Rethinking alignment in video super-resolution transformers. arXiv preprint arXiv:2207.08494 (2022)
75. Song, Y., Dhariwal, P.: Improved techniques for training consistency models. arXiv preprint arXiv:2310.14189 (2023)
76. Song, Y., Wang, M., Yang, Z., Xian, X., Shi, Y.: Negvsr: Augmenting negatives for generalized noise modeling in real-world video super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 10705–10713 (2024)
77. Sun, L., Wu, R., Ma, Z., Liu, S., Yi, Q., Zhang, L.: Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach (2025)
78. Sun, M., Wang, W., Li, G., Liu, J., Sun, J., Feng, W., Lao, S., Zhou, S., He, Q., Liu, J.: Ar-diffusion: Asynchronous video generation with auto-regressive diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7364–7373 (2025)
79. Sun, Y.C., Yeo, C.Y., Chu, E., Chen, J.C., Liu, Y.L.: Fiper: Factorized features for robust image super-resolution and compression. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
80. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: Tdan: Temporally-deformable alignment network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3360–3369 (2020)
81. Tsai, S.R., Chang, W.C., Lee, J.Y., Su, C.H., Liu, Y.L.: Lightsout: Diffusion-based outpainting for enhanced lens flare removal. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6353–6363 (2025)
82. Tsai, Y.J., Liu, Y.L., Qi, L., Chan, K.C., Yang, M.H.: Dual associated encoder for face restoration. arXiv preprint arXiv:2308.07314 (2023)

83. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
84. Wang, F.Y., Huang, Z., Bergman, A., Shen, D., Gao, P., Lingelbach, M., Sun, K., Bian, W., Song, G., Liu, Y., et al.: Phased consistency models. Advances in neural information processing systems **37**, 83951–84009 (2024)
85. Wang, F.Y., Yang, L., Huang, Z., Wang, M., Li, H.: Rectified diffusion: Straightness is not your need in rectified flow. arXiv preprint arXiv:2410.07303 (2024)
86. Wang, H., Liu, Y., Liu, H., Wang, C.C., Guo, Y., Li, H., Wang, B., Sun, J.: Temporal-consistent video restoration with pre-trained diffusion models. arXiv preprint arXiv:2503.14863 (2025)
87. Wang, J., Lin, S., Lin, Z., Ren, Y., Wei, M., Yue, Z., Zhou, S., Chen, H., Zhao, Y., Yang, C., et al.: Seedvr2: One-step video restoration via diffusion adversarial post-training. arXiv preprint arXiv:2506.05301 (2025)
88. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision **132**(12), 5929–5949 (2024)
89. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: 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)
90. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: International Conference on Computer Vision Workshops (ICCVW) (2021)
91. Wang, Y., Yang, W., Chen, X., Wang, Y., Guo, L., Chau, L.P., Liu, Z., Qiao, Y., Kot, A.C., Wen, B.: Sinsr: diffusion-based image super-resolution in a single step. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25796–25805 (2024)
92. Weng, S., Zheng, H., Zhan, P., Hong, Y., Jiang, H., Li, S., Shi, B.: Vires: Video instance repainting with sketch and text guidance. arXiv preprint arXiv:2411.16199 (2024)
93. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. arXiv preprint arXiv:2406.08177 (2024)
94. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: Seesr: Towards semantics-aware real-world image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25456–25467 (2024)
95. Xia, B., He, J., Zhang, Y., Wang, Y., Tian, Y., Yang, W., Van Gool, L.: Structured sparsity learning for efficient video super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22638–22647 (2023)
96. Xie, D., Xu, Z., Hong, Y., Tan, H., Liu, D., Liu, F., Kaufman, A., Zhou, Y.: Progressive autoregressive video diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6322–6332 (2025)
97. Xie, R., Liu, Y., Zhou, P., Zhao, C., Zhou, J., Zhang, K., Zhang, Z., Yang, J., Yang, Z., Tai, Y.: Star: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17108–17118 (2025)
98. Xie, S., Xiao, Z., Kingma, D., Hou, T., Wu, Y.N., Murphy, K.P., Salimans, T., Poole, B., Gao, R.: Em distillation for one-step diffusion models. Advances in Neural Information Processing Systems **37**, 45073–45104 (2024)

99. Xu, K., Yu, Z., Wang, X., Mi, M.B., Yao, A.: Enhancing video super-resolution via implicit resampling-based alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2546–2555 (2024)
100. Xu, Y., Zhao, Y., Xiao, Z., Hou, T.: Ufogen: You forward once large scale text-to-image generation via diffusion gans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8196–8206 (2024)
101. Xu, Y., Park, T., Zhang, R., Zhou, Y., Shechtman, E., Liu, F., Huang, J.B., Liu, D.: Videogigagan: Towards detail-rich video super-resolution (2024)
102. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. International Journal of Computer Vision **127**(8), 1106–1125 (2019)
103. Yang, S., Zhou, Y., Liu, Z., Loy, C.C.: Fresco: Spatial-temporal correspondence for zero-shot video translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8703–8712 (2024)
104. Yang, X., He, C., Ma, J., Zhang, L.: Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In: European Conference on Computer Vision. pp. 224–242. Springer (2024)
105. Yang, X., Xiang, W., Zeng, H., Zhang, L.: Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4781–4790 (2021)
106. Yeh, C.H., Lin, C.Y., Wang, Z., Hsiao, C.W., Chen, T.H., Shiu, H.S., Liu, Y.L.: Diffir2vr-zero: Zero-shot video restoration with diffusion-based image restoration models. arXiv preprint arXiv:2407.01519 (2024)
107. Yi, P., Wang, Z., Jiang, K., Jiang, J., Ma, J.: Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3106–3115 (2019)
108. Yin, G., Qu, Z., Jiang, X., Jiang, S., Han, Z., Zheng, N., Yang, H., Liu, X., Yang, Y., Li, D., et al.: Online streaming video super-resolution with convolutional look-up table. IEEE Transactions on Image Processing **33**, 2305–2317 (2024)
109. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, B.: Improved distribution matching distillation for fast image synthesis. Advances in neural information processing systems **37**, 47455–47487 (2024)
110. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024)
111. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22963–22974 (2025)
112. Youk, G., Oh, J., Kim, M.: Fma-net: Flow-guided dynamic filtering and iterative feature refinement with multi-attention for joint video super-resolution and deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 44–55 (2024)
113. Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by residual shifting. Advances in neural information processing systems **36**, 13294–13307 (2023)

114. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5728–5739 (2022)
115. Zhai, Y., Lin, K., Yang, Z., Li, L., Wang, J., Lin, C.C., Doermann, D., Yuan, J., Wang, L.: Motion consistency model: Accelerating video diffusion with disentangled motion-appearance distillation. Advances in Neural Information Processing Systems **37**, 111000–111021 (2024)
116. Zhang, A., Yue, Z., Pei, R., Ren, W., Cao, X.: Degradation-guided one-step image super-resolution with diffusion priors (2024), `https://arxiv.org/abs/2409.17058`
117. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
118. Zhang, R., Gu, J., Chen, H., Dong, C., Zhang, Y., Yang, W.: Crafting training degradation distribution for the accuracy-generalization trade-off in real-world super-resolution (2023)
119. Zhang, Y., Yao, A.: Realviformer: Investigating attention for real-world video super-resolution. In: European Conference on Computer Vision. pp. 412–428. Springer (2024)
120. Zhang, Z., Li, R., Guo, S., Cao, Y., Zhang, L.: Tmp: Temporal motion propagation for online video super-resolution. IEEE Transactions on Image Processing (2024)
121. Zhang, Z., Wu, B., Wang, X., Luo, Y., Zhang, L., Zhao, Y., Vajda, P., Metaxas, D., Yu, L.: Avid: Any-length video inpainting with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7162–7172 (2024)
122. Zhang, Z., Liu, K., Chen, Z., Li, X., Chen, Y., Duan, B., Kong, L., Zhang, Y.: Infvsr: Breaking length limits of generic video super-resolution. arXiv preprint arXiv:2510.00948 (2025)
123. Zhao, W., Zhou, J., Zhu, X., Chen, W., Zhang, X.Y., Lei, Z., Wang, F.: Realisvsr: Detail-enhanced diffusion for real-world 4k video super-resolution. arXiv preprint arXiv:2507.19138 (2025)
124. Zhao, W., Zhou, J., Zhu, X., Chen, W., Zhang, X.Y., Lei, Z., Wang, F.: Realisvsr: Detail-enhanced diffusion for real-world 4k video super-resolution. arXiv preprint arXiv:2507.19138 (2025)
125. Zheng, J., Hu, M., Fan, Z., Wang, C., Ding, C., Tao, D., Cham, T.J.: Trajectory consistency distillation: Improved latent consistency distillation by semi-linear consistency function with trajectory mapping. arXiv preprint arXiv:2402.19159 (2024)
126. Zheng, K., Lu, C., Chen, J., Zhu, J.: Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics. Advances in Neural Information Processing Systems **36**, 55502–55542 (2023)
127. Zheng, M., Sun, L., Dong, J., Pan, J.: Efficient video super-resolution for real-time rendering with decoupled g-buffer guidance. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11328–11337 (2025)
128. Zhou, M., Zheng, H., Wang, Z., Yin, M., Huang, H.: Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In: Forty-first International Conference on Machine Learning (2024)
129. Zhou, S., Yang, P., Wang, J., Luo, Y., Loy, C.C.: Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2535–2545 (2024)
130. Zhou, X., Zhang, L., Zhao, X., Wang, K., Li, L., Gu, S.: Video super-resolution transformer with masked inter&intra-frame attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25399–25408 (2024)
131. Zhou, Z., Chen, D., Wang, C., Chen, C., Lyu, S.: Simple and fast distillation of diffusion models. Advances in Neural Information Processing Systems **37**, 40831–40860 (2024)
132. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9308–9316 (2019)
133. Zhuang, J., Guo, S., Cai, X., Li, X., Liu, Y., Yuan, C., Xue, T.: Flashvsr: Towards real-time diffusion-based streaming video super-resolution. arXiv preprint arXiv:2510.12747 (2025)