WANSR: DISTILLING VIDEO DIFFUSION FOR TRAJECTORY-AWARE REAL-WORLD IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world image super-resolution (Real-ISR) aims to reconstruct high-resolution (HR) images from low-resolution (LR) inputs perturbed by unknown and mixed degradations. Although diffusion-based methods enable generating perceptually satisfying results, their intermediate denoising states are typically treated as numerical variables for sampling, making it hard to interpret and explicitly constrain the restoration process. Therefore, we ask the question: *Can the intermediate restoration process be explicitly modeled?* In this paper, we propose WANSR, a trajectory-aware framework that reformulates Real-ISR as a conditional video generation problem. Instead of directly predicting a single HR image, WANSR models a sequence of restoration states that explicitly describes the progressive evolution toward the final high-quality distribution. To stabilize this evolution, we introduce anchor-guided trajectory alignment by aligning transition directions and anchor-level magnitudes. We further propose quality-verified trajectory refinement, which evaluates restoration progress of each latent state and enforces monotonic quality improvement along the generated trajectory. Extensive experiments on multiple Real-ISR benchmarks demonstrate that our WANSR achieves superior reconstruction performance, while explicitly revealing the progressive process of degradation removal and details reconstruction.

1 INTRODUCTION

Real-world image super-resolution (Real-ISR) aims to reconstruct high-resolution (HR) images from degraded low-resolution (LR) observations without relying on fixed degradation assumptions (*e.g.*, bicubic downsampling and known blur kernels). The LR inputs are usually corrupted by unknown mixing degradations, where degradations of multiple types (*e.g.*, blur, noise and compression artifacts) and severity are mixed in different orders. Earlier methods Cai et al. (2019); Wei et al. (2020); Wang et al. (2021); Zhang et al. (2021) reduce the gap between ideal training data and real-world inputs with strategies such as real-degraded pair collection and degradation synthesis.

While recent diffusion-based Real-ISR methods Wang et al. (2024a); Lin et al. (2024); Wu et al. (2024b); Wang et al. (2024b); Wu et al. (2024a); Li et al. (2025) demonstrate remarkable performance in modeling realistic textures and high-frequency details, the generated content is likely to suffer from low fidelity. We attribute such a tendency to the endpoint-oriented characteristic of diffusion models (as shown in Fig. 1(a)). Although latents are refined through an iterative denoising process, the intermediate states are treated as numerical variables for sampling instead of meaningful restoration stages. As a result, it is difficult to explicitly constrain how textures should be progressively recovered, especially for few-step methods where the rapid generation process leaves limited room for trajectory regularization.

Can the intermediate restoration process be explicitly modeled? This is non-trivial in Real-ISR scenarios, where only the final degraded LR is available, while the underlying degradation chain and sequential states remain unknown. Therefore, instead of recovering a physically plausible degradation chain, we aim to construct an observable and controllable LR-to-HR restoration trajectory, where its intermediate states serve as meaningful proxies for progressive restoration. Specifically, three challenges remain: *(1) Evolving process opacity:* existing methods represent intermediate states as

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.

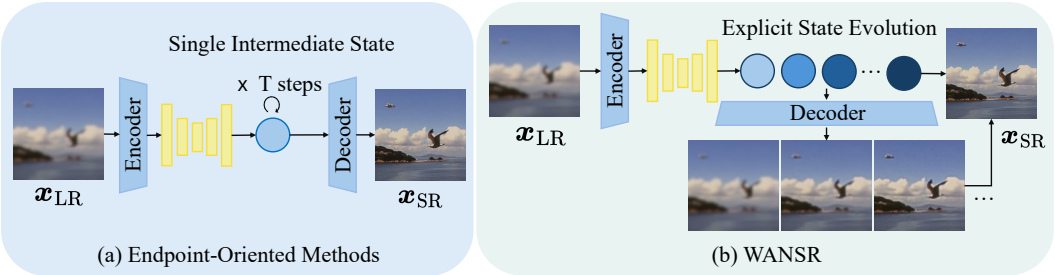

Figure 1: (a) Endpoint-oriented methods directly map a degraded LR input to a single HR output through an implicit restoration process. Their intermediate denoising states remain inaccessible, offering limited insight into how reconstruction evolves. (b) WANSR reformulates Real-ISR as explicit state evolution, where a sequence of intermediate latent states represents successive reconstruction stages and gradually transforms from LR to HR distribution.

noisy latents, making the LR-to-HR evolution hard to observe and interpret; *(2) Direction-agnostic trajectory:* transitions between states are not explicitly constrained by a coherent restoration direction, which may leads to discontinuous or detoured evolution instead of a stable progression of degradation removal and detail reconstruction; *(3) Progress-unverified intermediate states:* intermediate states are rarely evaluated with criteria (*e.g.*, degradation level or perceptual quality), making it unclear whether each state is genuinely approaching the target HR distribution.

To address these challenges, we propose WANSR, a trajectory-aware Real-ISR framework that reformulates Real-ISR as a video generation problem. Instead of predicting a single HR output, our WANSR generates a sequence of reconstruction frames that explicitly depict the transition from degraded LR inputs to the final HR result (see Fig. 1(b)). To further stabilize the progressive restoration process, we introduce anchor-guided trajectory alignment that encourages the predicted transitions to follow a pseudo restoration direction, making the trajectory more consistent and straight. Furthermore, we propose a quality-verified trajectory refinement strategy, where a state judger evaluates the intermediate stages and enforces a monotonic LR-to-HR progression. This enables explicit measurement of intermediate states while guiding the latent state toward a high-quality HR distribution. Our contributions are summarized as follows:

- We propose WANSR, a trajectory-aware Real-ISR framework that reformulates Real-ISR as conditional video generation. Instead of predicting a single HR output, WANSR generates an observable and temporally ordered sequence of restoration states, explicitly revealing the progressive evolution from degraded LR input to restored HR output.
- We introduce anchor-guided trajectory alignment to regularize the latent restoration trajectory. By aligning transition directions and anchor evolution magnitudes with a pseudo restoration trajectory, it encourages stable and coherent progression while preserving meaningful restoration dynamics.
- We develop a quality-verified trajectory refinement strategy, where a pre-trained state judge evaluates intermediate latent states and enforces monotonically evolution toward the HR distribution, thereby improving the reliability of the predicted trajectory.
- Extensive experiments on both synthetic and real-world benchmarks demonstrate that WANSR achieves superior reconstruction fidelity while providing an explicit and progressive restoration trajectory.

2 Related Work

2.1 Real-World Image Super-Resolution

Early Real-ISR methods Cai et al. (2019); Wei et al. (2020); Wang et al. (2021); Zhang et al. (2021) focus on narrowing the domain gap between synthetic training data and real degradations, either by collecting real paired benchmarks Cai et al. (2019); Wei et al. (2020), or by designing more realistic blind degradation pipelines Wang et al. (2021); Zhang et al. (2021). Recent methods Wang et al. (2024a); Lin et al. (2024); Wu et al. (2024a); Duan et al. (2025) increasingly explore generative priors from pre-trained diffusion models to improve perceptual quality under complex unknown

degradations. StableSR Wang et al. (2024a) introduces a time-aware encoder and controllable feature wrapping to exploit pre-trained Stable Diffusion Rombach et al. (2022) while balancing fidelity and perceptual quality. PASD Yang et al. (2024) improves structural consistency through pixel-aware cross-attention and degradation-robust guidance. SeeSR Wu et al. (2024b) employs a degradation-aware prompt extractor to provide complementary soft and hard semantic guidance. To reduce the high sampling cost of diffusion process, OSEDiff Wu et al. (2024a) performs one-step generation from the LQ latent through score distillation. AddSR Tai et al. (2026), TSD SR Dong et al. (2025), TADSR Zhang et al. (2025), and InvSR Yue et al. (2025) further investigate adversarial distillation, target score supervision, timestep-aware generation, and diffusion inversion for efficient restoration. DiT4SR Duan et al. (2025) extends Real-ISR to diffusion transformers by injecting LR guidance into the DiT attention layers. However, most existing Real-ISR methods treat restoration as a black box. They mainly focus on aspects of improving realism, semantic fidelity and efficiency, while providing limited interpretability on how the image is progressively restored. In contrast, our method explicitly models the restoration trajectory, making the evolution from a degraded LR input to a high-quality output both observable and directly constrainable.

2.2 Visual Chain of State Modeling

Recent studies Meng et al. (2023); Yin et al. (2026); Zheng et al. (2026) start to externalize implicit visual processes as ordered intermediate states. CoI Meng et al. (2023) represents reasoning steps through generated images, while GVCoT Yin et al. (2026) introduces intermediate visual cues for localizing and planning image edits, making semantic reasoning more observable. IRSDE Luo et al. (2023) models restoration as a reverse stochastic process from degraded observations to clean images. Efficient differential equation-based methods Zhu et al. (2025); Li et al. (2025) optimize and distill restoration trajectories for faster inference. However, their intermediate states serve as numerical sampling variables rather than visible and explicitly constrained restoration states. DiffIR2VR-Zero Yeh et al. (2024) transfers image restoration diffusion models to video restoration by latent warping and token merging, which focuses more on temporal consistency than modeling the restoration process of a single image. Chain-of-Zoom Kim et al. (2026) decomposes extreme super-resolution into a sequence of progressively enlarged scaling processes, while our method operates at a fixed target scale and models quality evolution from degraded LR to high-quality HR. V-Bridge Zheng et al. (2026) formulates image restoration as a video generation problem, where the intermediate sequence serves as a bridge to incorporate video priors into few-shot restoration tasks instead of explicitly modeling and constraining the underlying restoration dynamics. In contrast, our intermediate states directly represent successive stages of degradation removal and detail recovery, turning Real-ISR from endpoint prediction into an explicit and controllable chain of reconstruction.

3 Preliminaries

Distribution Matching Distillation. Distribution Matching Distillation (DMD) distills a pretrained diffusion model into an efficient generator by matching the generator's output distribution to real data distribution. The DMD framework Yin et al. (2024a;b); Fan et al. (2026) typically includes a learnable generator $\boldsymbol{G_\phi}$, a learnable fake score estimator $\boldsymbol{F_\theta}$ and a frozen pretrained teacher score estimator $\boldsymbol{T_{\theta'}}$. $\boldsymbol{T_{\theta'}}$ estimates the real data distribution, whereas $\boldsymbol{F_\theta}$ estimates the distribution induced by the generator. $\boldsymbol{G_\phi}$ is optimized by minimizing the discrepancy between real and generated distributions, thereby approximating the real distribution. Given a Gaussian noise input $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, the generator produces a clean sample:

$$\boldsymbol{x}_0 = \boldsymbol{G_\phi}(\boldsymbol{z}), \qquad \boldsymbol{x}_0 \sim p_{\text{fake}}(\boldsymbol{x}_0), \tag{1}$$

where p_{fake} denotes the generated distribution. DMD aims to minimize the reverse Kullback-Leibler (KL) divergence $D_{\text{KL}}(p_{\text{fake}} \| p_{\text{real}})$ between generated and real distributions:

$$\begin{aligned} \nabla_\phi D_{KL}\left(p_{\text{fake}} \| p_{\text{real}}\right) &= E_{\boldsymbol{z},\boldsymbol{x}_0}\left[\left(\nabla_{\boldsymbol{x}_0} \log p_{\text{fake}}\left(\boldsymbol{x}_0\right) - \nabla_{\boldsymbol{x}_0} \log p_{\text{real}}\left(\boldsymbol{x}_0\right)\right) \frac{d\boldsymbol{G}}{d\boldsymbol{\phi}}\right], \\ &= \mathbb{E}_{\boldsymbol{z}}\left[\left(\boldsymbol{s}_{\text{fake}}(\boldsymbol{x}_0) - \boldsymbol{s}_{\text{real}}(\boldsymbol{x}_0)\right) \frac{d\boldsymbol{G}}{d\boldsymbol{\phi}}\right], \end{aligned} \tag{2}$$

where $\boldsymbol{s}_p(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$ denotes the score function of distribution p. When using diffusion models as score estimators to approximate the data distribution, with the generated $\boldsymbol{x}_0$ diffused as

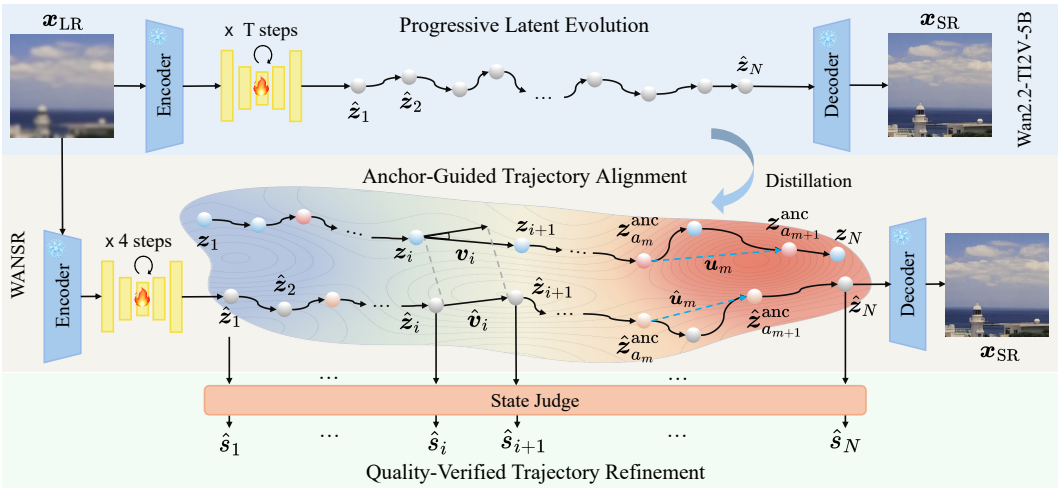

Figure 2: The proposed WANSR. Given a degraded LR input, our WANSR models a sequence of latent states that progressively evolves from an LR distribution toward an HR distribution. The predicted trajectory is constrained by an anchor-guided trajectory alignment from aspects of transition direction and anchor magnitude. A pre-trained state judge further enforces the intermediate states to follow a monotonic progression toward higher quality.

$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}$, D_{KL} is further approximated as follows:

$$\nabla_{\phi} D_{\mathrm{KL}} \approx \mathbb{E}_{\substack{\boldsymbol{z},\boldsymbol{x}_0,\\ t,\boldsymbol{\epsilon},\boldsymbol{x}_t}} \left[w_t \Big(\boldsymbol{T}_{\boldsymbol{\theta}'}(\boldsymbol{x}_t, t) - \boldsymbol{F}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) \Big) \frac{\mathrm{d}\boldsymbol{G}}{\mathrm{d}\boldsymbol{\phi}} \right], \tag{3}$$

where $w_t = \frac{\alpha_t^2}{\alpha_t \sigma_t + \sigma_t^2}$. The fake score estimator is alternately optimized with $\boldsymbol{G}_{\phi}$-generated samples with objective below:

$$\mathcal{L}_{\mathrm{fake}} = \mathbb{E}_{\substack{\boldsymbol{z},\boldsymbol{x}_0,\\ t,\boldsymbol{\epsilon},\boldsymbol{x}_t}} \left[\| \boldsymbol{F}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - (\boldsymbol{\epsilon} - \boldsymbol{x}_0) \|_2^2 \right]. \tag{4}$$

4 Method

Given a degraded low-resolution (LR) input $\boldsymbol{x}_{\mathrm{LR}}$, we aim to reconstruct a high-resolution (HR) output $\boldsymbol{x}_{\mathrm{HR}}$ at a scale factor. Our generator $\boldsymbol{G}_{\phi}$ consists of a frozen VAE encoder $\boldsymbol{E}$, a diffusion module parameterized by ϕ, and a frozen VAE decoder $\boldsymbol{D}$. The proposed framework consists of the following components: *progressive latent evolution*, *anchor-guided trajectory alignment*, and *quality-verified trajectory refinement*. Progressive latent evolution models the reconstruction process with a sequence of latent states that gradually evolve from a severely degraded distribution toward the high-quality distribution. During training, anchor-guided trajectory alignment constrains transitions with pseudo restoration anchors, while quality-verified trajectory refinement further enforces a monotonic progression toward higher-quality states.

4.1 Progressive Latent Evolution

As shown in Fig. 2, WANSR reformulates Real-ISR as an LR-conditioned sequence generation problem. Instead of predicting a single HR image, WANSR models an ordered latent sequence under the condition of $\boldsymbol{x}_{\mathrm{LR}}$ and text embeddings:

$$\hat{\boldsymbol{Z}} = \boldsymbol{G}_{\phi}(\boldsymbol{h}) = \{\hat{\boldsymbol{z}}_i\}_{i=1}^{N}, \tag{5}$$

where $\boldsymbol{h} = (\boldsymbol{E}(\boldsymbol{x}_{\mathrm{LR}}), \boldsymbol{c})$ denotes the condition embedding. $\hat{\boldsymbol{z}}_1$ and $\hat{\boldsymbol{z}}_N$ denote the initial degraded state and final restored state, respectively. Decoding the sequential states with decoder $\boldsymbol{D}$ further reveals the progressive evolution and obtains the super-resolved result $\boldsymbol{x}_{\mathrm{SR}}$ from the last frame. Learning such a progressive evolution requires an ordered restoration trajectory. However, real intermediate states between an LR image and its HR target are generally unavailable. We therefore construct pseudo restoration trajectories by reversing a synthetically generated degradation sequence. Starting

from the clean HR target, we sequentially apply randomly selected degradation operations (*e.g.*, blur and noise) from Real-ESRGAN pipeline Wang et al. (2021), where the degraded result after each operation is retained as an anchor frame. The recorded degraded anchor frames are then reversed to form an LR-to-HR sequence, where the first and last frame correspond to the degraded LR input and clean HR target, respectively. Pixelwise linear interpolation is performed between adjacent anchor frames to reach target frame number, after which all frames are encoded by the frozen VAE encoder to form the pseudo latent trajectory $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i=1}^{N}$. These pseudo trajectories provide explicit supervision for learning progressive degradation removal and detail recovery. To equip the teacher model with such capability, we first fine-tune a pretrained video foundation model $\boldsymbol{T}_{\boldsymbol{\theta}'}$ on the constructed trajectories. Under the flow-matching formulation, the teacher is optimized as follows:

$$\mathcal{L}_{\text{teacher}} = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{T}_{\boldsymbol{\theta}'}\left(\boldsymbol{Z}_t, t, \boldsymbol{h}\right) - (\boldsymbol{\epsilon} - \boldsymbol{Z})\right\|_2^2\right], \tag{6}$$

where $\boldsymbol{Z}_t = \alpha_t \boldsymbol{Z} + \sigma_t \boldsymbol{\epsilon}$. After this adaptation, the teacher is frozen and we adopt the following DMD gradient to distill the teacher's generative knowledge into WANSR:

$$\nabla_{\boldsymbol{\phi}} \mathcal{L}_{\text{DMD}} = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[w_t\Big(\boldsymbol{T}_{\boldsymbol{\theta}'}\left(\hat{\boldsymbol{Z}}_t, t, \boldsymbol{h}\right) - \boldsymbol{F}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{Z}}_t, t, \boldsymbol{h}\right)\Big)\frac{\partial \hat{\boldsymbol{Z}}}{\partial \boldsymbol{\phi}}\right], \tag{7}$$

where $\hat{\boldsymbol{Z}}_t = \alpha_t \hat{\boldsymbol{Z}} + \sigma_t \boldsymbol{\epsilon}$ denotes the diffused noisy latent sequence. The fake estimator $\boldsymbol{F}_{\boldsymbol{\theta}}$ is optimized below:

$$\mathcal{L}_{\text{fake}} = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{F}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{Z}}_t, t, \boldsymbol{h}\right) - (\boldsymbol{\epsilon} - \hat{\boldsymbol{Z}})\right\|_2^2\right]. \tag{8}$$

During training, $\mathcal{L}_{\text{DMD}}$ and $\mathcal{L}_{\text{fake}}$ are alternatively optimized to update $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, respectively, while teacher $\boldsymbol{T}_{\boldsymbol{\theta}'}$, VAE encoder $\boldsymbol{E}$ and decoder $\boldsymbol{D}$ remain frozen.

4.2 Anchor-Guided Trajectory Alignment

The DMD objective distills the teacher's restoration capability into the student by matching the output distribution. However, it does not explicitly constrain how the intermediate latent states evolve from the degraded state $\hat{\boldsymbol{z}}1$ to the restored state $\hat{\boldsymbol{z}}N$. Consequently, the student may produce sequences with plausible distribution but irregular or drifted transitions. To regularize this evolution, we leverage the pseudo latent trajectory $\boldsymbol{Z}$ as intermediate guidance, introducing *transition-based direction alignment* and *anchor-based magnitude alignment* to learn the progressive restoration process, while avoiding learning artificial uniform transitions.

Transition-based direction alignment. Given the predicted trajectory $\{\hat{\boldsymbol{z}}_i\}_{i=1}^{N}$, the local transition vector $\{\hat{\boldsymbol{v}}_i\}_{i=1}^{N-1}$ (*i.e.*, $\hat{\boldsymbol{v}}_i = \hat{\boldsymbol{z}}_{i+1} - \hat{\boldsymbol{z}}_i$) describes the local evolution from current state toward a cleaner state. Consecutive transitions provide denser direction guidance compared to anchor-to-anchor directions, which cover longer intervals and provide sparse directions. We align predicted transition vectors and pseudo ones (*i.e.*, $\boldsymbol{v}_i = \boldsymbol{z}_{i+1} - \boldsymbol{z}_i$) with following cosine similarity:

$$\mathcal{L}_{\text{dir}} = \frac{\sum_{i=1}^{N-1} w_i \left[1 - \cos\left(\hat{\boldsymbol{v}}_i, \boldsymbol{v}_i\right)\right]}{\sum_{i=1}^{N-1} w_i + \epsilon}, \tag{9}$$

where $w_i = \|\boldsymbol{v}_i\|_2$ weights each transition by its displacement magnitude, reducing the influence of small transitions and emphasizing directions with substantial changes. ϵ is a small constant for numerical stability.

Anchor-based magnitude alignment. Direction alignment specifies the evolution direction but does not determine how far the latent state should move along that direction, potentially causing insufficient or excessive reconstruction. Magnitude constraint is therefore necessary to calibrate the restoration progress. However, matching every transition magnitude is suboptimal, as intermediate states are uniformly interpolated, making the transition step dependent on interpolation density rather than real restoration dynamics. We therefore constrain transition magnitude only between anchor states $\{\hat{\boldsymbol{z}}_{a_m}^{\text{anc}}\}_{m=1}^{M}$, which reflects meaningful restoration states without introducing artificial linear interpolation. With anchor transition vector $\hat{\boldsymbol{u}}_m = \hat{\boldsymbol{z}}_{a_{m+1}}^{\text{anc}} - \hat{\boldsymbol{z}}_{a_m}^{\text{anc}}$ (denoted with blue dashed lines in Fig. 2), the anchor transition magnitude is then regularized as:

$$\mathcal{L}_{\text{mag}} = \frac{1}{M-1}\sum_{m=1}^{M-1} \frac{\left|\|\hat{\boldsymbol{u}}_m\|_2 - \|\boldsymbol{u}_m\|_2\right|}{\|\boldsymbol{u}_m\|_2 + \epsilon}, \tag{10}$$

Table 1: Quantitative comparison ($\times 4$) on DIV2K-val, DRealSR and RealSR datasets, where the best, second best and third best results are highlighted with red, blue, and **bold**, respectively. Our method achieves superior performance on reference-based metrics, while remaining competitive with representative methods such as StableSR and SeeSR on non-reference metrics, demonstrating a strong balance between reconstruction fidelity and perceptual quality.

Dataset	Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	MANIQA↑	MUSIQ↑	NIQE↓	CLIP-IQA↑	LIQE↑	Q-Align↑	TOPIQ↑
DIV2K-val	StableSR	21.61	0.5340	0.3119	0.2047	25	0.4214	65.90	4.78	0.6787	3.8830	3.5322	0.5991
	DiffBIR	21.48	0.5050	0.3668	0.2209	33	0.5667	69.87	4.99	**0.7300**	**4.3497**	4.1017	0.7039
	SeeSR	21.97	0.5672	0.3194	0.1968	26	**0.5038**	68.68	4.81	0.6937	4.2753	**3.9773**	**0.6858**
	ResShift	22.66	0.5888	0.3077	0.2136	31	0.3693	58.90	6.92	0.5717	3.0824	3.2573	0.5059
	SinSR	**22.51**	0.5678	0.3240	0.2066	36	0.4214	62.82	6.02	0.6470	3.4914	3.5081	0.5706
	OSEDiff	22.06	**0.5735**	0.2942	0.1975	26	0.4411	67.97	**4.71**	0.6681	4.1187	3.8357	0.6188
	AdcSR	22.07	0.5637	**0.2853**	**0.1899**	26	0.4468	67.99	4.36	0.6763	4.1328	3.8612	0.6526
	TSD-SR	21.38	0.5436	0.2673	0.1821	23	0.4837	71.69	4.32	0.7416	4.4498	3.9430	0.6679
	TADSR	22.14	0.5669	0.3078	0.2110	30	0.5364	**69.64**	4.87	0.7353	4.3610	4.0774	0.7044
	Ours	23.35	0.6110	0.2622	0.1833	**25**	0.4062	66.05	5.35	0.5735	3.7653	3.5796	0.5940
DRealSR	StableSR	25.81	0.7222	0.3286	0.2270	149	0.3874	58.51	6.53	0.6362	3.2430	3.0610	0.5323
	DiffBIR	24.05	0.5831	0.4669	0.2881	181	0.5544	66.14	**6.34**	**0.7072**	4.1035	3.7159	0.6730
	SeeSR	25.82	0.7405	0.3175	0.2316	147	**0.5131**	65.09	6.40	0.6912	**4.1279**	3.5866	**0.6576**
	ResShift	24.85	0.7038	0.3871	0.2632	160	0.3247	51.23	8.64	0.5394	2.6762	2.8597	0.4530
	SinSR	25.86	0.7182	0.3664	0.2485	171	0.3832	55.33	7.00	0.6386	3.0448	3.0891	0.5103
	OSEDiff	25.85	0.7545	**0.2967**	**0.2164**	136	0.4656	64.69	6.44	0.6973	3.9420	3.5458	0.6002
	AdcSR	**25.96**	0.7436	0.3046	0.2200	134	0.4977	**66.26**	6.45	0.7048	4.1134	**3.6511**	0.6528
	TSD-SR	25.63	0.7254	0.2966	0.2136	**135**	0.4742	66.61	5.92	0.7347	4.1294	3.6041	0.6181
	TADSR	26.17	**0.7471**	0.3235	0.2412	148	0.5518	67.01	6.19	0.7397	4.2832	3.7518	0.6763
	Ours	26.45	0.7634	0.2673	0.1973	118	0.4226	61.88	6.75	0.5865	3.4674	3.3795	0.5718
RealSR	StableSR	23.12	0.6798	0.3004	0.2141	128	0.4276	65.88	5.88	0.6233	3.6109	3.2771	0.5748
	DiffBIR	23.33	0.6181	0.3650	0.2399	131	0.5582	69.28	5.84	**0.7050**	4.1008	3.7914	0.6759
	SeeSR	23.60	0.6947	0.3008	0.2225	125	**0.5438**	69.82	**5.40**	0.6706	4.1368	3.7179	0.6892
	ResShift	**23.96**	0.7067	0.3279	0.2475	128	0.3467	56.90	8.06	0.5361	2.8516	3.0242	0.4527
	SinSR	24.51	0.7075	0.3188	0.2353	136	0.3994	60.77	6.29	0.6118	3.1647	3.1589	0.5163
	OSEDiff	23.59	**0.7070**	0.2921	**0.2129**	124	0.4710	69.09	5.65	0.6686	4.0631	3.6937	0.6249
	AdcSR	23.90	0.7037	**0.2885**	**0.2129**	**118**	0.5160	**69.90**	5.35	0.6730	**4.2631**	**3.7755**	**0.6792**
	TSD-SR	23.29	0.6901	0.2743	0.2105	115	0.5113	71.19	5.12	0.7166	4.2650	3.7147	0.6579
	TADSR	23.60	0.6878	0.3171	0.2323	125	0.5733	71.16	5.40	0.7276	4.4573	3.9434	0.7084
	Ours	24.97	0.7499	0.2250	0.1800	97	0.4371	66.50	5.94	0.5290	3.5774	3.4714	0.5911

where $\{a_m\}_{m=1}^{M}$ denotes the indices of the anchor states. The final anchor-guided trajectory alignment loss is therefore shown as follows:

$$\mathcal{L}_{\text{traj}} = \mathcal{L}_{\text{dir}} + \lambda_{\text{mag}} \mathcal{L}_{\text{mag}}, \tag{11}$$

where λ_{mag} controls the weight of the magnitude term.

4.3 Quality-Verified Trajectory Refinement

Although anchor-guided trajectory alignment constrains the transition direction and the magnitude between anchor states, it does not explicitly verify whether every generated latent state exhibits progressive restoration toward the target HR distribution. As a result, intermediate non-anchor states may stagnate or deviate from the desired restoration trajectory. To this end, we introduce a lightweight state judge $\boldsymbol{J}$ that assigns each latent state a restoration progress score (*i.e.*, $s_i = \boldsymbol{J}(\boldsymbol{z}_i)$). A higher score indicates that the corresponding state is closer to the HR domain. We first train $\boldsymbol{J}$ on the pseudo latent states $\{\boldsymbol{z}_i\}_{i=1}^{N}$ by enforcing monotonic ranking with the ranking hinge loss below:

$$\mathcal{L}_{\text{rank}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \max\left(0,\, r_m + s_i - s_{i+1}\right), \tag{12}$$

where r_m denotes the ranking margin. Then the state judge is frozen and applied to the generated trajectory $\{\hat{\boldsymbol{z}}_i\}_{i=1}^{N}$, with all latent states sharing the same state judge. The same ranking loss is used to optimize the generator, encouraging the latent state to make monotonic progress toward HR.

4.4 Loss Function

To ensure faithful content reconstruction in the final output, we decode the latent state and compare the last frame $\boldsymbol{x}_{\text{SR}}$ with clean HR target using pixel and perceptual losses:

$$\mathcal{L}_{\text{rec}} = \|\boldsymbol{x}_{\text{SR}} - \boldsymbol{x}_{\text{HR}}\|_2^2 + \lambda_{\text{lpips}} \,\text{LPIPS}\left(\boldsymbol{x}_{\text{SR}}, \boldsymbol{x}_{\text{HR}}\right) \tag{13}$$

The generator parameters ϕ are optimized by minimizing the following objective:

$$\mathcal{L}_G = \lambda_{\text{DMD}} \mathcal{L}_{\text{DMD}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{traj}} \mathcal{L}_{\text{traj}} + \lambda_{\text{rank}} \mathcal{L}_{\text{rank}}, \tag{14}$$

where λ_{DMD}, λ_{rec}, λ_{traj} and λ_{rank} denote the weight of each loss term.



Figure 3: Qualitative comparisons ($\times 4$) on the RealSR and DIV2K-val datasets. The compared methods typically produce texture deviations (*e.g.*, incorrect text characters of StableSR and TADSR in the first row) and distorted contents (*e.g.*, the twisted faces of DiffBIR and ResShift in the second row) under severe degradation. In contrast, our results benefit from the progressive refinement and strike the balance between perceptual quality and structural fidelity.

5 Experiment

5.1 Experimental settings

Datasets. By convention Kim et al. (2026); Li et al. (2025); Wu et al. (2024a), we use the combination of LSDIR Li et al. (2023) dataset and the first 10K images from FFHQ Karras et al. (2019) dataset for training. During training, the HR ground truth (GT) is randomly cropped to a resolution of 512×512, and the degradation pipeline of Real-ESRGAN Wang et al. (2021) is used to synthesize the 128×128 LR input. Evaluation is conducted under the $\times 4$ setting on synthetic DIV2K-val Agustsson & Timofte (2017), realistic DRealSR Wei et al. (2020) and RealSR Cai et al. (2019) benchmarks. Specifically, following the evaluation protocol of StableSR Wang et al. (2024a), DIV2K-val includes 3,000 degraded LR-HR pairs. DRealSR and RealSR contain 93 and 100 real captured LR HR pairs, respectively.

Metrics. We adopt both full-reference and non-reference-based metrics to evaluate restoration fidelity and perceptual quality. Specifically, PSNR and SSIM are adopted as distortion metrics, while LPIPS and DISTS evaluate the perceptual similarity between super-resolved results and HR GT. We additionally include FID to evaluate the distribution similarity between the results and HR reference. For non-reference-based metrics, we employ MANIQA, MUSIQ, NIQE, CLIP-IQA, LIQE, Q-Align and TOPIQ. Lower values indicate better performance for LPIPS, FID, and NIQE, whereas higher values are preferred for all other metrics.

Baselines. We conduct comprehensive comparisons with representative diffusion-based Real-ISR methods. Multi-step methods include StableSR Wang et al. (2024a), DiffBIR Lin et al. (2024), and SeeSR Wu et al. (2024b). ResShift Yue et al. (2023) is included as a representative few-step method. Recent one-step methods include SinSR Wang et al. (2024b), OSEDiff Wu et al. (2024a), TSDSR Dong et al. (2025), AdcSR Chen et al. (2025), and TADSR Zhang et al. (2025). We use officially released models and recommended inference configurations for all baseline methods.

Implementation details. To balance efficiency and reconstruction performance, both the generated video and pseudo restoration trajectory contain 41 frames, which are temporally encoded into 11 latent states (*i.e.*, $N = 11$). Our method is built on Wan2.2-TI2V-5B Wan et al. (2025), whose teacher is fine-tuned on pseudo restoration trajectories using the flow matching objective to generate LR-to-HR restoration videos, conditioned on an LR input and text prompt. During fine-tuning, only the DiT module is updated, while all remaining modules are frozen. The teacher uses 50 denoising steps and is trained for 250,000 iterations. The state judge $\boldsymbol{J}$ is implemented as a lightweight MLP and pretrained for 10,000 iterations on pseudo latent trajectories using the ranking hinge loss in Eq. equation 12, after which it is frozen. The distillation process is conducted for 100,000 iterations. We include more details in the appendix.

Table 2: Complexity and efficiency analysis. Latency and FLOPs are measured on a single NVIDIA B200 GPU for 128×128 LR inputs and 512×512 HR outputs.

Method	Denoising Steps	Latency (ms)	Params (M)	FLOPs (T)
StableSR	200	3878	1401	199.22
DiffBIR	50	3658	1683	112.76
SeeSR	50	1374	2511	168.88
ResShift	4	103	174	6.51
SinSR	1	45	174	5.29
OSEDiff	1	54	1765	4.83
AdcSR	1	23	456	1.10
TSD-SR	1	42	2207	6.16
TADSR	1	97	1767	4.83
Ours	4	620	5704	315.67

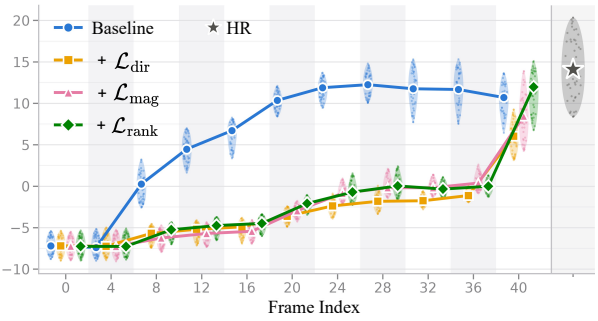

Figure 4: Distribution evolution on the restoration trajectory. VGG features of generated frames are visualized in the clustered feature space.

5.2 Quantitative Results

Quantitative results are reported in Tab. 1. As can be seen, our method consistently achieves superior performance on reference-based metrics (*e.g.*, a PSNR gain of up to 2.4 dB over DiffBIR on the DrealSR dataset), demonstrating its strong ability to reconstruct realistic details while faithfully preserving the LR content. On the non-reference metrics, our method shows suboptimal performance compared to GAN/diffusion-distillation methods (*e.g.*, TADSR and TSD-SR), which results from the focus that these baselines rely on strong generative priors to synthesize texture-rich details, whereas our method is explicitly optimized to faithfully preserve the content of input LR. Meanwhile, it is worth noting that our method still clearly outperforms representative Real-ISR methods such as StableSR, ResShift, and SinSR on most metrics, including a 6.55 point improvement in MUSIQ over SinSR, and a reduction of 1.89 in NIQE over ResShift on DrealSR. Most importantly, we additionally provide an explicit and controllable restoration trajectory that is unavailable in all compared methods.

5.3 Qualitative Results

We include qualitative comparisons on the RealSR and DIV2K-val datasets in Fig. 3. As can be seen, although the compared methods can leverage generative priors and synthesize visually plausible details, they show limited controllability and insufficient fidelity to the input LR content, resulting in texture deviations (*e.g.*, incorrect text characters of StableSR and TADSR in the first row). When the LR input is severely degraded and contains limited information, compared methods either fail in reconstructing realistic results (*e.g.*, the distorted faces in the second row reconstructed by DiffBIR, ResShift, SinSR, and AdcSR), or hallucinate visually plausible but input-inconsistent details (*e.g.*, the realistic but low-fidelity face reconstructed by TADSR). In contrast, our method faithfully preserves the original structure through modeling the latent evolution trajectory, where each restoration state is highly related to the preceding state and therefore provides progressive, input-preserving refinement instead of unconstrained generation. As a result, our method successfully strikes the balance between perceptual quality and structural fidelity. Please find more qualitative comparisons in the appendix.

5.4 Complexity and Efficiency Analysis

We include a detailed complexity and efficiency analysis in Tab. 2, where denoising steps, inference latency, model parameters and FLOPs of each method are reported. Since anchor-guided trajectory alignment and quality-verified trajectory refinement are only included during training, our method introduces no extra computational overhead during inference. As can be seen, despite being built on the video foundation model (*i.e.*, Wan2.2-TI2V-5B), the resulting computational cost remains affordable on commonly used GPUs. Our method requires only 4 denoising steps and achieves an inference latency of 620 ms, which is approximately $6\times$ faster than StableSR (3878 ms) and DiffBIR (3658 ms), and $2.2\times$ faster than SeeSR (1374 ms). Compared with the 4-step ResShift, WANSR achieves substantially superior reconstruction performance across all evaluation metrics(*e.g.*, over 2 dB improvement in PSNR and a 10-point gain in MUSIQ in Tab. 1). It is also worth noting that WANSR provides an explicit and controllable restoration trajectory that is missing in all compared

Table 3: Ablation studies on DIV2K-val and RealSR dataset, where the best results are highlighted with **bold**. The baseline model is trained using $\mathcal{L}_{\mathrm{DMD}}$ and $\mathcal{L}_{\mathrm{rec}}$, after which transition-based direction alignment, anchor-based magnitude alignment, and quality-verified trajectory refinement are progressively incorporated through $\mathcal{L}_{\mathrm{dir}}$, $\mathcal{L}_{\mathrm{mag}}$, and $\mathcal{L}_{\mathrm{rank}}$, respectively.

Dataset	Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	MANIQA↑	MUSIQ↑	NIQE↓	CLIP-IQA↑	LIQE↑	Q-Align↑	TOPIQ↑
DIV2K-val	Baseline	23.17	0.6046	0.2965	0.1990	**24**	0.3069	55.90	5.91	0.4762	2.9603	2.8467	0.4235
	+ $\mathcal{L}_{\mathrm{dir}}$	**23.74**	**0.6239**	0.2762	0.1920	28	0.3779	62.88	5.49	0.5364	3.3892	3.3029	0.5525
	+ $\mathcal{L}_{\mathrm{mag}}$	23.47	0.6182	0.2689	0.1878	26	0.4038	65.48	5.39	0.5556	3.7294	3.5435	0.5830
	+ $\mathcal{L}_{\mathrm{rank}}$ (Full)	23.35	0.6110	**0.2622**	**0.1833**	25	**0.4062**	**66.05**	**5.35**	**0.5735**	**3.7653**	**3.5796**	**0.5940**
RealSR	Baseline	25.13	0.7494	**0.2000**	**0.1676**	98	0.3152	55.20	6.89	0.4254	2.5082	2.7003	0.3982
	+ $\mathcal{L}_{\mathrm{dir}}$	**25.62**	**0.7637**	0.2239	0.1837	102	0.4086	64.11	6.16	0.5021	3.2486	3.2492	0.5484
	+ $\mathcal{L}_{\mathrm{mag}}$	25.27	0.7581	0.2230	0.1826	100	0.4308	66.00	5.99	0.5129	3.4909	3.4234	0.5741
	+ $\mathcal{L}_{\mathrm{rank}}$ (Full)	24.97	0.7499	0.2250	0.1800	**97**	**0.4371**	**66.50**	**5.94**	**0.5290**	**3.5774**	**3.4714**	**0.5911**

baseline methods. Overall, WANSR trades increased model capacity for superior reconstruction quality and controllability, while maintaining a practical inference latency and model complexity.

5.5 Ablation Studies

We begin with a baseline trained with $\mathcal{L}_{\mathrm{DMD}}$ and $\mathcal{L}_{\mathrm{rec}}$. We then progressively equip the baseline with transition-based direction alignment, anchor-based magnitude alignment, and quaflity-verified trajectory refinement via introducing $\mathcal{L}_{\mathrm{dir}}$, $\mathcal{L}_{\mathrm{mag}}$ and $\mathcal{L}_{\mathrm{rank}}$, respectively. Quantitative results on DIV2K-val and RealSR are included in Tab. 3.

Effectiveness of transition-based direction alignment. As shown in the 2nd row of Tab. 3, introducing $\mathcal{L}_{\mathrm{dir}}$ (denoted as + $\mathcal{L}_{\mathrm{dir}}$) consistently improves both reconstruction fidelity and perceptual quality of the baseline model, leading to a PSNR gain of 0.57 dB and 0.49 dB on DIV2K-val and RealSR datasets, respectively. Meanwhile, it also introduces consistent improvement on non-reference metrics. As shown by the yellow trajectory in Fig. 4, transition-based direction alignment reduces oscillations at early states and induces a coherent restoration evolution toward the HR distribution.

Effectiveness of anchor-based magnitude alignment. Compared to the model equipped with transition-based direction alignment, including $\mathcal{L}_{\mathrm{mag}}$ (denoted as + $\mathcal{L}_{\mathrm{mag}}$ in the 3rd row of Tab. 3) consistently improves all non-reference quality metrics, as well as the perceptual metrics LPIPS and DISTS. For instance, MUSIQ increases from 64.11 to 66.00 on RealSR dataset. Despite PSNR dropping moderately, the overall improvement indicates enhanced perceptual quality. As illustrated in Fig. 4, at each frame index, pink trajectory lies closer to HR distribution than the yellow trajectory, demonstrating that magnitude alignment effectively calibrates the restoration extent between anchor states and pushes generated states further toward desired restoration direction.

Effectiveness of quality-verified trajectory refinement. The full model (denoted as + $\mathcal{L}_{\mathrm{rank}}$) further introduces $\mathcal{L}_{\mathrm{rank}}$ to enforce that consecutive states make monotonic progress toward high-quality distribution. As shown in the 4th row of Tab. 3, it consistently improves all non-reference metrics, obtaining MUSIQ gains of 0.85, and NIQE decreases of 0.4 on DRealSR, respectively. Meanwhile, PSNR shows only moderate decreases (*i.e.*, 0.12 dB and 0.30 dB on DIV2K-val and RealSR dataset). Fig. 4 further shows that, with $\mathcal{L}_{\mathrm{mag}}$ (pink curve), the predicted quality scores fluctuate across states and occasionally exhibit quality regression. In contrast, introducing $\mathcal{L}_{\mathrm{rank}}$ produces a steadily increasing score curve (green curve) toward the HR distribution and remains closer to it than the pink trajectory at nearly every frame index, demonstrating that quality-verified refinement consistently advances the entire restoration trajectory.

6 Conclusion

In this paper, we propose WANSR, a trajectory-aware framework that reformulates Real-ISR as a conditional video generation problem. Instead of directly predicting a single HR image, WANSR produces an ordered sequence of restoration states that explicitly captures the progressive evolution from a degraded LR input to a high-quality HR output. Anchor-guided trajectory alignment encourages these states to follow a coherent pseudo-restoration path, while quality-verified trajectory refinement promotes monotonic progress toward the HR distribution. Extensive experiments on synthetic and real-world benchmarks demonstrate the superiority of WANSR in delivering LR-fidelity reconstruction results, while revealing how degradations are progressively removed and image details are gradually recovered. Ablation studies verify the effectiveness of each proposed component.

AI USE STATEMENT

In this work, we used generative AI tools to improve the grammar and readability of author-written drafts and AI coding agents to assist with implementing the proposed methods. We did not use generative AI tools to clean or reformat the dataset. We compared AI-edited text with the original drafts to make sure it aligned with proposed claims. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

WANSR adopts publicly available pretrained models and datasets. The paper specifies the experimental setup, including training iterations, configurations of optimizers, and weights assigned to each loss term, and the prompt used for conditional video generation. We will release the code, configuration files, and checkpoints to facilitate reproduction of our results.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 126–135, 2017.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

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, 2025.

Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23174–23184, 2025.

Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S Ren, Chunle Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18948–18958, 2025.

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

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Bryan Sangwoo Kim, Jeongsol Kim, and Jong Chul Ye. Chain-of-zoom: Extreme super-resolution via scale autoregression and preference alignment. *Advances in Neural Information Processing Systems*, 38:125428–125460, 2026.

Jianze Li, Jiezhang Cao, Yong Guo, Wenbo Li, and Yulun Zhang. One diffusion step to real-world super-resolution via flow trajectory distillation. *arXiv preprint arXiv:2502.01993*, 2025.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1775–1787, 2023.

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, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. *arXiv preprint arXiv:2301.11699*, 2023.

Fanxu Meng, Haotong Yang, Yiding Wang, and Muhan Zhang. Chain of images for intuitively reasoning. *arXiv preprint arXiv:2311.09241*, 2023.

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.

Ying Tai, Rui Xie, Chen Zhao, Kai Zhang, Zhenyu Zhang, Jun Zhou, and Jian Yang. Addsr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. *Pattern Recognition*, pp. 113012, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. 2024a.

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.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. 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, 2024b.

Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *European conference on computer vision*, pp. 101–117. Springer, 2020.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37: 92529–92553, 2024a.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. 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, 2024b.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *European conference on computer vision*, pp. 74–91. Springer, 2024.

Chang-Han Yeh, Hau-Shiang Shiu, Chin-Yang Lin, Zhixiang Wang, Chi-Wei Hsiao, Ting-Hsuan Chen, and Yu-Lun Liu. Diffir2vr-zero: Zero-shot video restoration with diffusion-based image restoration models. *arXiv preprint arXiv:2407.01519*, 2024.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b.

Zijin Yin, Tiankai Hang, Yiji Cheng, Shiyi Zhang, Runze He, Yu Xu, Chunyu Wang, Bing Li, Zheng Chang, Kongming Liang, et al. Generative visual chain-of-thought for image editing. *arXiv preprint arXiv:2603.01893*, 2, 2026.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in neural information processing systems*, 36: 13294–13307, 2023.

Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23153–23163, 2025.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4791–4800, 2021.

Tianyi Zhang, Zheng-Peng Duan, Peng-Tao Jiang, Bo Li, Ming-Ming Cheng, Chun-Le Guo, and Chongyi Li. Time-aware one step diffusion network for real-world image super-resolution. *arXiv preprint arXiv:2508.16557*, 2025.

Shenghe Zheng, Junpeng Jiang, and Wenbo Li. V-bridge: Bridging video generative priors to versatile few-shot image restoration. *arXiv preprint arXiv:2603.13089*, 2026.

Zhiyu Zhu, Jinhui Hou, Hui Liu, Huanqiang Zeng, and Junhui Hou. Learning efficient and effective trajectories for differential equation-based image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

This appendix is structured as follows:

– Section A provides a theoretical analysis and ablation studies for our design in anchor-guided trajectory alignment, providing theoretical insights into its formulation.

– Section B provides a detailed process for preparing the pseudo restoration trajectory generation.

– Section C presents more qualitative comparisons to illustrate the high-fidelity and realistic results.

– Section D provides implementation details on the training, including training stages and objective weights.

A Anchor-Guided Trajectory Alignment

During training, we adopt the anchor-based magnitude and transition-based direction as trajectory constraints. We analyze the rationality of such a design below.

Anchor-based magnitude constraint. Consider two adjacent anchor images $\boldsymbol{x}_m^{anc}$ and $\boldsymbol{x}_{m+1}^{anc}$. Their interpolated image trajectory is parameterized by:

$$\boldsymbol{x}(\tau) = (1-\tau)\boldsymbol{x}_m^{anc} + \tau \boldsymbol{x}_{m+1}^{anc}, \qquad \tau \in [0,1], \tag{15}$$

and the corresponding latent trajectory is:

$$\boldsymbol{z}(\tau) = \boldsymbol{E}\big(\boldsymbol{x}(\tau)\big). \tag{16}$$

For two neighboring samples separated by $\Delta\tau$, their transition vector is:

$$\boldsymbol{v}_i = \boldsymbol{z}(\tau_i + \Delta\tau) - \boldsymbol{z}(\tau_i). \tag{17}$$

Assuming that the frozen encoder is locally smooth, a first-order expansion gives:

$$\boldsymbol{v}_i = \boldsymbol{z}'(\tau_i)\Delta\tau + \mathcal{O}(\Delta\tau^2). \tag{18}$$

Consequently, its normalized direction satisfies:

$$\frac{\boldsymbol{v}_i}{\|\boldsymbol{v}_i\|_2} = \frac{\boldsymbol{z}'(\tau_i)}{\|\boldsymbol{z}'(\tau_i)\|_2} + \mathcal{O}(\Delta\tau), \tag{19}$$

whereas its magnitude satisfies:

$$\|\boldsymbol{v}_i\|_2 = \|\boldsymbol{z}'(\tau_i)\|_2\Delta\tau + \mathcal{O}(\Delta\tau^2). \tag{20}$$

Equation 19 and Equation 20 reveal different properties of direction and magnitude: the normalized direction mainly describes the local geometry of the restoration path and is approximately invariant to the interpolation interval. In contrast, the local magnitude scales directly with $\Delta\tau$ and therefore depends on the arbitrary number of intermediate states. Increasing the interpolation density reduces the magnitude of every local transition even though the underlying restoration process remains unchanged. Therefore, directly matching such magnitudes enforces the generated trajectory to imitate the artificial temporal discretizations of the pseudo trajectory, while anchor displacement can avoid this ambiguity. For two adjacent anchor states, we define:

$$\boldsymbol{u}_m = \boldsymbol{z}_{a_{m+1}}^{anc} - \boldsymbol{z}_{a_m}^{anc}, \tag{21}$$

where the magnitude depends only on the two meaningful restoration anchors and remains unchanged regardless of how many intermediate states are inserted between them, thus representing the net restoration extent of an anchor interval without imposing an artificial transition speed.

Transition-based direction constraint. For an anchor interval, the displacement between two anchors can be written as

$$\boldsymbol{u}_m = \boldsymbol{z}(1) - \boldsymbol{z}(0) = \int_0^1 \boldsymbol{z}'(\tau)\, d\tau. \tag{22}$$

Accordingly, the normalized anchor direction is

$$\frac{\boldsymbol{u}_m}{\|\boldsymbol{u}_m\|_2} = \frac{\int_0^1 \boldsymbol{z}'(\tau)\, d\tau}{\left\|\int_0^1 \boldsymbol{z}'(\tau)\, d\tau\right\|_2}, \tag{23}$$

Table 4: Ablation studies on DIV2K-val and RealSR dataset, where AM, TD and AD denote anchor-based magnitude constraint, transition-based direction constraint and anchor-based direction constraint, respectively.

Dataset	Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑
DIV2K-val	$AM + TD$	**23.47**	**0.6182**	0.2689	0.1878	65.48
	$AM + AD$	23.32	0.6028	**0.2616**	**0.1862**	**65.67**
RealSR	$AM + TD$	**25.27**	**0.7581**	**0.2230**	**0.1826**	**66.00**
	$AM + AD$	25.17	0.7434	0.2295	0.1837	65.68

which characterizes only the net displacement between the two anchors. In contrast, as shown in Equation 19, the normalized transition direction approximates the local tangent $\boldsymbol{z}'(\tau_i)/\|\boldsymbol{z}'(\tau_i)\|_2$. For a nonlinear latent trajectory, this local tangent generally varies along the restoration path and cannot be recovered from the anchor displacement alone. Therefore, constraining only the anchor direction specifies the coarse evolution between two anchors while leaving the intermediate path underconstrained, allowing trajectories with different curvatures or local oscillations. Dense transition directions instead provide local supervision along the trajectory and thus more explicitly constrain its geometry. This property is further reflected by the weighted direction objective. With $w_i = \|\boldsymbol{v}_i\|_2$, the discrete objective in Eq. 9[1] approaches:

$$\mathcal{L}_{\text{dir}} \approx \frac{\int \|\boldsymbol{z}'(\tau)\|_2 \left[1 - \cos\left(\hat{\boldsymbol{z}}'(\tau), \boldsymbol{z}'(\tau)\right)\right] d\tau}{\int \|\boldsymbol{z}'(\tau)\|_2 \, d\tau}, \tag{24}$$

which can be interpreted as an arc length weighted directional alignment along the pseudo trajectory. Since the contribution of each transition is weighted by its local displacement, the resulting objective is less sensitive to the particular temporal discretization and better reflects the geometry of the underlying restoration path. We further conduct an ablation study on the replacing the transition-based direction constraint with anchor-based direction constraint, and report the results in Table 4. As can be seen, the combination of $AM + TD$ achieves an overall best performance.

B Pseudo Restoration Trajectory Generation

We summarizes the generation of the pseudo restoration trajectory in Algorithm 1. Given an HR image $\boldsymbol{x}_{\text{HR}}$, we first construct a sequence of degradation anchors by progressively applying the degradation operations $\mathcal{D} = \{\mathcal{D}_k\}_{k=1}^{K}$ that are generated with the Real-ESRGAN pipeline Wang et al. (2021). Specifically, starting from the clean HR image, each degradation operation is sequentially applied to the output of the preceding operation, producing $\tilde{\boldsymbol{X}}^{\text{anc}} = \{\tilde{\boldsymbol{x}}_k\}_{k=0}^{K}$, which describes the progressive degradation from the HR image toward its degraded LR. We then reverse this sequence to obtain the restoration anchor frames $\boldsymbol{X}^{\text{anc}} = \{\boldsymbol{x}_m^{\text{anc}}\}_{m=1}^{M}$, ordering from the most degraded state to the clean HR state. To construct a dense pseudo restoration trajectory, we distribute $P-1$ transitions (P indicates the target video frames) across the $M-1$ adjacent anchor intervals. For each pair of neighboring anchors $\boldsymbol{x}_m^{\text{anc}}$ and $\boldsymbol{x}_{m+1}^{\text{anc}}$, intermediate states are generated through uniform interpolation:

$$\boldsymbol{x} = (1-\alpha)\boldsymbol{x}_m^{\text{anc}} + \alpha \boldsymbol{x}_{m+1}^{\text{anc}}, \qquad \alpha = \frac{j}{p_m}, \tag{25}$$

where p_m denotes the number of transitions assigned to the m-th anchor interval. Concatenating the interpolated states from all intervals produces the pseudo video trajectory $\boldsymbol{X} = \{\boldsymbol{x}_p\}_{p=1}^{P}$, which progressively evolves from the degraded LR state toward the clean HR state while preserving the predefined degradation anchors. Finally, the pseudo video sequence is encoded by the frozen video VAE encoder $\boldsymbol{E}$ to obtain the latent restoration trajectory:

$$\boldsymbol{Z} = \boldsymbol{E}(\boldsymbol{X}) = \{\boldsymbol{z}_i\}_{i=1}^{N}. \tag{26}$$

Due to the temporal compression of the video VAE, the latent trajectory contains fewer temporal states than the pixel trajectory, with $N = (P-1)/r + 1$, where r denotes the temporal compression ratio. The pixel space anchor positions are accordingly mapped to their corresponding latent

[1]To differentiate from this appendix, we use abbreviations to denote sections, tables, and figures in the main texts (*i.e.*., "Sec." for sections, "Tab." for tables, and "Fig." for figures)

Algorithm 1 Pseudo Video Trajectory Generation

Input: HR image $\boldsymbol{x}_{\mathrm{HR}}$, degradation operations $\mathcal{D} = \{\mathcal{D}_k\}_{k=1}^{K}$, target video frame number P, frozen video VAE encoder $\boldsymbol{E}$, temporal compression ratio r
Output: Pseudo pixel trajectory $\boldsymbol{X} = \{\boldsymbol{x}_p\}_{p=1}^{P}$, pseudo latent trajectory $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i=1}^{N}$, latent anchor indices $\mathcal{A}$

$\tilde{\boldsymbol{x}}_0 \leftarrow \boldsymbol{x}_{\mathrm{HR}}$; */* Pseudo Anchor Generation */*
$\tilde{\boldsymbol{X}}^{\mathrm{anc}} \leftarrow \{\tilde{\boldsymbol{x}}_0\}$;
for $k = 1$ to K **do**
  Sample the degradation parameters of $\mathcal{D}_k$;
  $\tilde{\boldsymbol{x}}_k \leftarrow \mathcal{D}_k\left(\tilde{\boldsymbol{x}}_{k-1}\right)$;
  Append $\tilde{\boldsymbol{x}}_k$ to $\tilde{\boldsymbol{X}}^{\mathrm{anc}}$;
end for
$\boldsymbol{X}^{\mathrm{anc}} \leftarrow \mathrm{Reverse}\left(\tilde{\boldsymbol{X}}^{\mathrm{anc}}\right) = \{\boldsymbol{x}_m^{\mathrm{anc}}\}_{m=1}^{M}$; */* Reverse to LR → HR */*
/ Pseudo Restoration Trajectory Construction */*
Allocate $P-1$ transitions over the $M-1$ adjacent anchor intervals, yielding $\{p_m\}_{m=1}^{M-1}$ such that $\sum_{m=1}^{M-1} p_m = P-1$;
$\boldsymbol{X} \leftarrow \varnothing$, $\mathcal{A}_{\mathrm{pix}} \leftarrow \{1\}$;
for $m = 1$ to $M-1$ **do**
  for $j = 0$ to $p_m - 1$ **do**
    $\alpha \leftarrow j/p_m$;
    $\boldsymbol{x} \leftarrow (1-\alpha)\boldsymbol{x}_m^{\mathrm{anc}} + \alpha \boldsymbol{x}_{m+1}^{\mathrm{anc}}$; */* Uniform interpolation */*
    Append $\boldsymbol{x}$ to $\boldsymbol{X}$;
  end for
  Append $|\boldsymbol{X}| + 1$ to $\mathcal{A}_{\mathrm{pix}}$;
end for
Append $\boldsymbol{x}_M^{\mathrm{anc}}$ to $\boldsymbol{X}$; */* Final HR GT */*
/ Latent Trajectory Encoding */*
$N \leftarrow (P-1)/r + 1$; */* Temporal compression */*
$\boldsymbol{Z} \leftarrow \boldsymbol{E}(\boldsymbol{X}) = \{\boldsymbol{z}_i\}_{i=1}^{N}$; */* Video VAE encoding */*
$\mathcal{A} \leftarrow \mathrm{TemporalIndexMapping}\left(\mathcal{A}_{\mathrm{pix}}, r\right)$; */* Pixel to latent anchors */*
return $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\mathcal{A}$;

indices to obtain the latent anchor set $\mathcal{A}$. The obtained $\boldsymbol{Z}$ and $\mathcal{A}$ then serve as the pseudo restoration trajectory and anchor indices for the trajectory alignment objectives.

C More Qualitative Results

We provide more qualitative comparisons on the DRealSR Wei et al. (2020), RealSR Cai et al. (2019), and DIV2K-val Agustsson & Timofte (2017) datasets in Figure 5. As can be seen, existing methods often struggle to recover faithful details from severely degraded LR inputs. For instance, OSEDiff Wu et al. (2024a) introduces noise-like patterns that are unsupported by the input in the first leaf example, while others tend to oversmooth local structures and lose fine details. For structured regions, compared methods may hallucinate unreliable contents (*e.g.*, distorted patterns of DiffBIR Lin et al. (2024) and TADSR Zhang et al. (2025) in the second example). Similar issues can also be observed in the rope and leaf cases, where the recovered textures either deviate from the underlying structures, or become excessively smooth. The limitation becomes more obvious for challenging facial regions, where several methods alter facial geometry and generate identity-inconsistent details (*e.g.*, SeeSR Wu et al. (2024b) and AdcSR Chen et al. (2025)), resulting in noticeable deviations from the GT. In contrast, our method consistently preserves the underlying structures of the LR input while recovering natural and coherent details. By explicitly modeling the progressive latent evolution trajectory, each restoration state is constrained by its preceding state, which eliminates uncontrolled hallucination and achieves content-preserving refinement. Consequently, our method produces more faithful structures and details, achieving a better balance between perceptual quality and reconstruction fidelity.



Figure 5: Qualitative comparisons ($\times 4$) on the DreaSR, RealSR and DIV2K-val datasets. As can be seen, the compared methods confront problems such as hallucinating details that are unfaithful to the input LR (*e.g.*, noise-like details of OSEDiff in the first leaf case), and distorted contents (*e.g.*, unrealistic faces of ResShift and SinSR in the fourth case). In contrast, our method achieves high-fidelity and realistic results.

D Implementation Details

Teacher pre-training. We build the teacher on Wan2.2-TI2V-5B Wan et al. (2025). The Wan2.2 VAE compresses a clip by $4\times$ temporally and $16\times$ spatially into 48 latent channels. The VAE and the text encoder are kept frozen throughout the fine-tuning process. During fine-tuning, the full transformer is trained with the learning rate of 2×10^{-5}, and optimized by the AdamW optimizer Loshchilov & Hutter (2017). The teacher takes 50 denoising steps and is trained for 250K iterations.

Few-step distillation. The student generator, the fake score network and the frozen real score network are all initialized from the fine-tuned teacher. The student conducts 4 denoising steps. Genera-

tor and fake score network use separate AdamW optimizers with learning rates 5×10^{-7} and 1×10^{-7}, $\beta = (0.0, 0.999)$, weight decay 0.01 and constant schedules.

Text condition. All samples share one prompt: *"Generate a temporally static sequence where all frames share identical layout and geometry. Over time, progressively improve quality by removing artifacts from unknown degradations (e.g., noise/blur/compression), and restore content faithful to the first frame. No motion, no scene change."* The text embedding is computed once and reused for every iteration.

Objective weights. The objective weights λ_{DMD}, λ_{rec}, λ_{traj} λ_{rank} in Eq. 14 are set to 0.9, 1, 0.15, 7, respectively.