PReS-SR: Prior-Preserving Residual Specialization for Efficient Real-World Image Super-Resolution

Anonymous authors
Paper under double-blind review

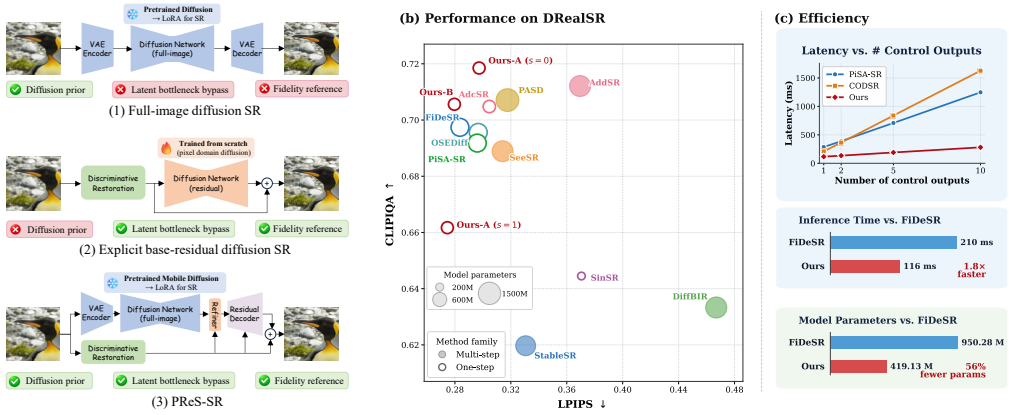

Figure 1: PReS-SR achieves a competitive quality-efficiency trade-off for Real-ISR. (a) Comparison of the pipelines of full-image diffusion SR, explicit base-residual diffusion SR, and our PReS-SR. (b) Performance comparison on DRealSR. (c) Comparison of efficiency.

Abstract

Real-world image super-resolution (Real-ISR) requires faithful recovery of observation-supported structures and plausible reconstruction of under-determined details. Pretrained full-image diffusion models provide rich natural-image priors, but their VAE reconstruction bottleneck may degrade spatially precise content. Existing generative SR methods mitigate this issue with LR-derived guidance during denoising or decoding, while recent VAE-free approaches move diffusion directly into pixel space. Despite these advances, the diffusion pathway still reconstructs the entire HQ image, coupling structural recovery with perceptual detail synthesis. Residual diffusion SR methods avoid such redundancy by predicting only a residual over a deterministic pixel-space SR reconstruction, but retarget the diffusion process itself from full-image generation to residual prediction. To retain the pretrained generative prior while leveraging the advantages of residual diffusion SR, we propose PReS-SR, a prior-preserving residual specialization framework that moves the residual-specialization boundary after diffusion prediction. Specifically, a discriminative restoration is carried directly to the output, while the pretrained diffusion backbone retains its full-image representation. A base-injected latent residual refinement block and a feature-injected residual decoder convert this representation into base-relative residual completion. Moreover, fidelity-perception control is localized to the lightweight post-backbone residual decoding, allowing multiple operating points to share the same diffusion computation. Experiments show that PReS-SR is particularly effective under constrained generative capacity, delivering competitive restoration quality with 56% fewer parameters and 1.8× faster inference than FiDeSR, while producing five controlled outputs up to 3.7× faster than PiSA-SR.

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.

1 INTRODUCTION

Image super-resolution aims to recover high-quality images from degraded low-quality inputs. While existing methods perform well under simplified synthetic degradations, real-world degradations remain substantially more complex. With the rapid development of image generation models, pretrained latent diffusion models have become an important source of natural-image priors for Real-ISR (Wang et al., 2024a; Yang et al., 2024; Lin et al., 2024; Wu et al., 2024b; Yu et al., 2024), enabling realistic recovery of under-determined details. Recent one-step diffusion SR methods further improve efficiency by eliminating costly iterative sampling (Wu et al., 2024a; Wang et al., 2024b; Kim et al., 2026; Dong et al., 2025; Arora et al., 2025; Wang et al., 2025).

Nevertheless, Real-ISR requires not only plausible detail synthesis but also faithful recovery of observation-constrained structures. The VAE used in latent diffusion introduces a reconstruction bottleneck that can degrade spatially precise structures and fine details (Wang et al., 2026; Yu et al., 2026). To address this issue, recent generative SR methods increasingly incorporate LR-derived guidance along the denoising trajectory or at the decoding stage (Wang et al., 2024a; Yang et al., 2024; Lin et al., 2024; Wu et al., 2024b; Arora et al., 2025), while VAE-pruned methods retain latent-space diffusion but reduce the cost of VAE reconstruction (Chen et al., 2025), whereas VAE-free methods eliminate this bottleneck by moving diffusion directly into pixel space (Wang et al., 2026). However, the diffusion pathway is still tasked with reconstructing the entire HQ image, unnecessarily coupling observation-constrained structure recovery and perceptually realistic detail synthesis into a single fidelity–perception trade-off. This coupling becomes more restrictive for lightweight deployment. Pruning or distilling the diffusion U-Net reduces the capacity available to simultaneously preserve structures and synthesize details, while simplifying the VAE further aggravates the reconstruction bottleneck.

Explicit base–residual restoration offers a natural alternative: reliably recovered content is carried directly through a pixel-space base, leaving the generative pathway to model only the missing details. However, existing residual diffusion methods retarget diffusion to base-relative residual prediction, often with task-specific architectures or conditioning (Li et al., 2022; Whang et al., 2022; Zhang et al., 2023; Shang et al., 2024; Yang et al., 2026), limiting direct reuse of pretrained T2I diffusion models in their native full-image formulation. This raises our central question: can base–residual restoration be combined with a pretrained full-image diffusion prior without retargeting diffusion itself? We introduce PReS-SR to achieve this by moving the residual-specialization boundary after diffusion prediction: the upstream representation remains full-image, while only downstream decoding is specialized for base-relative residual completion.

This design bridges the pixel-space restoration base and the full-image diffusion representation, while also enabling efficient post-backbone fidelity–perception control. Our contributions are:

- We propose PReS-SR with Post-Backbone Residual Specialization (PBRS) to combine a pretrained full-image T2I diffusion prior with explicit base–residual restoration. PBRS retains the upstream full-image diffusion formulation and specializes only downstream decoding for base-relative residual completion. This provides a fidelity reference that bypasses the VAE bottleneck and enables detail generation using the pretrained prior, without retargeting pretrained diffusion to a residual-prediction task.
- We introduce Cross-Space Restoration Injection (CSRI), comprising a base-aware latent refinement block (BLRB) and feature refinement blocks (FRBs), to adapt full-image diffusion representations for base-relative residual completion. The BLRB conditions latent refinement on the encoded restoration base, while the FRBs inject multi-scale restoration features into residual decoding.
- We systematically study residual decoding, topology–capacity interaction, and post-backbone fidelity–perception adjustment with shared upstream computation, enabling efficient one-step residual diffusion SR and controllable fidelity–perception trade-offs.

2 RELATED WORK

Full-Image SR with Diffusion Models. Diffusion-based Real-ISR methods adapt pretrained generative priors to reconstruct HQ images. StableSR and PASD inject time-aware or pixel-aware

conditions, while DiffBIR, SeeSR, SUPIR, and GuideSR use deterministic, semantic, multimodal, or full-resolution guidance (Wang et al., 2024a; Yang et al., 2024; Lin et al., 2024; Wu et al., 2024b; Yu et al., 2024; Arora et al., 2025). One-step variants improve efficiency through variational or target score distillation (Wu et al., 2024a; Dong et al., 2025); AdcSR prunes the pipeline, GenDR-Pix removes the VAE for pixel-space restoration, and FiDeSR introduces latent residual refinement and frequency-aware injection (Chen et al., 2025; Wang et al., 2026; Kim et al., 2026). Nevertheless, these methods still reconstruct the complete HQ image through the generative pathway, coupling structural recovery with detail synthesis in a way that can become increasingly restrictive under model compression.

Explicit Base–Residual Diffusion Restoration. Explicit residual learning has long been explored in super-resolution, as in VDSR, DRRN, and LapSRN (Kim et al., 2016; Tai et al., 2017; Lai et al., 2017). This idea was later extended to diffusion-based restoration. SRDiff (Li et al., 2022) performs diffusion on the pixel-space residual between the HR target and an upsampled LR image, while DvSR (Whang et al., 2022) replaces the interpolated base with a deterministic prediction followed by stochastic residual refinement. UCDIR (Zhang et al., 2023) further strengthens this paradigm by injecting restoration and degradation guidance. Following the trend, ResDiff (Shang et al., 2024) introduces a CNN base with frequency-guided residual diffusion, and HDW-SR (Yang et al., 2026) exploits wavelet-derived high-frequency guidance for residual modeling. These methods formulate diffusion in a base-relative residual space, with a prediction target different from that of pretrained full-image generative models.

Controllable Diffusion Restoration. Controllable diffusion restoration allows users to adjust the fidelity–perception trade-off at inference time. StableSR (Wang et al., 2024a) modulates a frozen diffusion prior with time-aware features, while PiSA-SR (Sun et al., 2025) adjusts pixel- and semantic-oriented LoRA guidance scales. CODSR (Chen et al., 2026) selectively activates generative priors in texture-rich regions, while LQ-guided modulation preserves structural fidelity. ODTSR (Fang et al., 2026) regulates fidelity through adjustable control noise, whereas TADSR (Zhang et al., 2026) uses the diffusion timestep as an explicit fidelity control. IFCSR (Back et al., 2026) instead moves the control to image space by interpolating between fidelity- and realism-oriented outputs, enabling new operating points without additional network inference. However, most existing approaches place fidelity–perception control within the generative pathway, requiring repeated diffusion inference for different preferences.

3 METHOD

3.1 POST-BACKBONE RESIDUAL SPECIALIZATION

As Figure 2 shows, our PReS-SR framework aims to obtain the modeling economy of explicit base–residual restoration together with a pretrained full-image diffusion prior, while protecting reliably recovered pixel-space content from VAE-induced reconstruction degradation. The key idea is to move the residual specialization boundary from the diffusion process to downstream decoding. We first formalize the distinctions among conventional full-image diffusion SR, explicit base–residual diffusion SR, and PReS-SR.

Full-image diffusion SR. Given an LQ input x_L, conventional latent diffusion SR encodes it into a latent representation z_L, predicts a reconstructed full-image latent $\hat{z}_H$ with G_θ, and decodes it into the SR image:

$$z_L = E(x_L), \qquad \hat{z}_H = G_\theta(z_L), \qquad x_{\mathrm{SR}} = D(\hat{z}_H). \tag{1}$$

Here, E and D denote the pretrained VAE encoder and decoder, respectively. For simplicity, G_θ abstracts different diffusion prediction formulations (e.g., noise, velocity or flow) and their reconstruction rules.

Explicit base–residual diffusion SR. Explicit base–residual restoration instead first obtains a deterministic pixel-space reconstruction

$$b = B(x_L), \tag{2}$$

where B denotes a discriminative SR model. Existing residual diffusion formulations then specialize the diffusion process itself to predict a residual relative to b, yielding a distinct diffusion formulation rather than a downstream variant of pretrained full-image diffusion:

$$\hat{r} = G^r_\theta(x_L, b), \qquad x_{\mathrm{SR}} = b + \hat{r}. \tag{3}$$

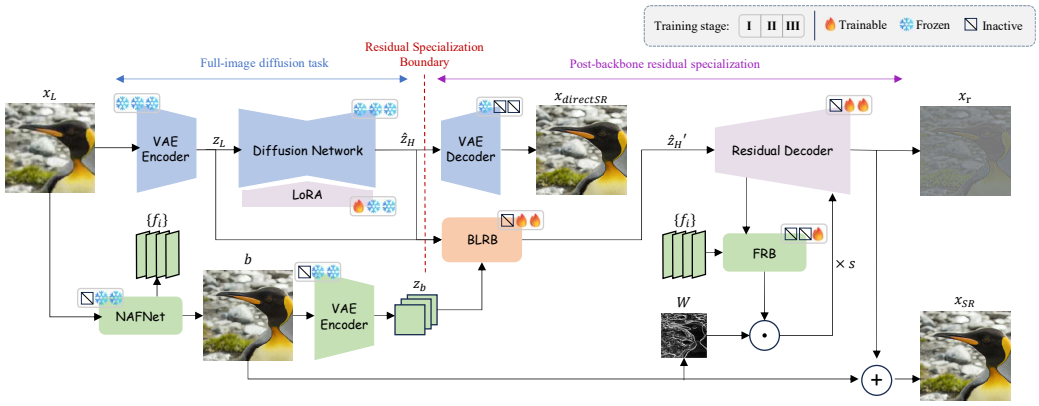

Figure 2: Overview of PReS-SR. A NAFNet base B produces a faithful base b, while the pretrained diffusion backbone with LRRB predicts a full-image latent $\hat{z}_H$. Post-Backbone Residual Specialization replaces the VAE decoder with a residual decoder D_r that outputs x_r, giving $x_{\mathrm{SR}} = b + x_r$. Cross-Space Restoration Injection injects z_b and multi-scale NAFNet features into the residual pathway, with s controlling the perception–fidelity trade-off.

Post-Backbone Residual Specialization. To preserve the full-image diffusion formulation while retaining the modeling economy of explicit base–residual restoration, we introduce Post-Backbone Residual Specialization (PBRS). Instead of specializing the diffusion process for residual prediction, PBRS retains the upstream full-image latent prediction, and moves residual specialization after diffusion prediction. Given the reconstructed full-image latent representation $\hat{z}_H$, PBRS replaces the full-image decoder D with a residual decoder D_r, which learns to extract information complementary to the discriminative restoration base b and decode it into a pixel-space residual:

$$x_r = D_r(\hat{z}_H), \qquad x_{\mathrm{SR}} = b + x_r. \tag{4}$$

In our implementation, B is instantiated with NAFNet (Chen et al., 2022) to provide a structurally reliable restoration base b. PBRS directly carries b to the output while retaining full-image latent prediction in the diffusion backbone, specializing only the downstream decoder for base-relative residual completion. This formulation allows observation-supported content in b to bypass the VAE decoder, reducing redundant full-image reconstruction while retaining the upstream full-image diffusion representation. Thus, PBRS decouples output residualization from diffusion-task residualization.

3.2 Cross-Space Restoration Injection

PBRS introduces a representation gap between the full-image diffusion latent and the pixel residual. We address this gap through Cross-Space Restoration Injection (CSRI), which exposes discriminative restoration conditions to the generative pathway across latent and feature spaces.

CSRI consists of two components: latent injection and feature injection. Given the fidelity-oriented base b produced by NAFNet, we encode it with the pretrained VAE encoder to obtain the base latent z_b. Meanwhile, we extract features from four spatially matched decoder stages of the NAFNet, yielding a set of multi-scale fidelity features $\{f_i\}$. Latent injection introduces z_b into the base-aware latent refinement block (BLRB) to refine the residual z_r predicted by the diffusion U-Net and facilitate the subsequent residual decoding. Feature injection further injects $\{f_i\}$ into the corresponding stages of the residual decoder, directly incorporating multi-scale restoration features during the decoding of x_r.

Latent Injection. Latent injection employs a BLRB to refine the latent residual z_r. The BLRB is built upon the Latent Residual Refinement Block (LRRB) (Kim et al., 2026), but is additionally conditioned on the base latent z_b. In our one-step implementation, the diffusion U-Net predicts a latent residual z_r, such that the full-image latent is obtained as $\hat{z}_H = z_L - z_r$. The BLRB predicts a base-aware residual correction Δz_r from $[z_L, z_r, z_b]$. The overall refinement can be written as

$$\Delta z_r = \mathrm{BLRB}([z_L, z_r, z_b]), \qquad z_r' = z_r + \Delta z_r, \tag{5}$$

Following the one-step diffusion formulation, the refined residual z'_r is subtracted from z_L to obtain the reconstructed full-image latent representation:

$$\hat{z}'_H = z_L - z'_r. \tag{6}$$

In this way, the base latent z_b provides additional fidelity information for refining the full-image latent representation before residual decoding.

Feature Injection. Feature injection adopts a similar dense-block like BLRB. The multi-scale NAFNet features $\{f_i\}$ are first projected and resized to match the corresponding decoder stages, yielding $\{\hat{f}_i\}$. At each stage, the decoder feature h_i is concatenated with $\hat{f}_i$ and fed into a Feature Refinement Block (FRB), which can be viewed as a dimension-adapted dense refinement block, to predict a feature correction:

$$\Delta_i = \mathrm{FRB}_i([\hat{f}_i, h_i]). \tag{7}$$

For inference-time fidelity–perception control, we consider global and spatially adaptive feature injection. Mode A performs global injection, where a scalar s controls the injection strength over the entire image. Mode B performs spatially adaptive injection, where a spatial weight map W_i is derived from the Sobel gradient map of the discriminative restoration base b. For global injection, $W_i = \mathbf{1}$. This assigns stronger fidelity injection to edge- and texture-rich regions while applying weaker injection elsewhere. The decoder feature is updated as

$$h'_i = h_i + s\,(W_i \odot \Delta_i)\,, \tag{8}$$

The introduction of s further enables an efficient controllable fidelity–perception trade-off. As the modulation is applied entirely after the diffusion backbone, SR outputs with different preferences can share the same U-Net inference, avoiding repeated diffusion computation.

CSRI strengthens the reconstruction quality of PBRS by fully exploiting the information provided by the discriminative SR branch. The latent and feature injections respectively introduce learnable corrections Δz_r and Δ_i, enabling progressive refinement from the full-image latent representation to the pixel-space residual across different resolutions and representation spaces. Such cross-space injection allows the diffusion representation and the VAE-bypassed base to complement each other more effectively, while also providing an efficient mechanism for controllable fidelity–perception adjustment.

3.3 Training Strategy

PBRS and CSRI substantially modify the original SR pathway, and directly optimizing the entire framework with residual specialization may disturb the generative prior inherited from the pretrained diffusion model. To mitigate this issue, we adopt a three-stage training strategy that progressively introduces SR adaptation, residual specialization, and controllable feature injection.

Stage I: Direct SR Adaptation. We first adapt the pretrained DreamLite model to the SR task while preserving its generative prior as much as possible. The DreamLite backbone is kept frozen, while LoRA parameters and the trainable convolution layers are optimized together with an LRRB to adapt the model to full-image SR. We initially optimize only the reconstruction loss $\mathcal{L}_{\mathrm{rec}}$ to stabilize SR reconstruction, and subsequently introduce the adversarial loss $\mathcal{L}_{\mathrm{adv}}$ to improve perceptual quality.

Stage II: Residual Specialization. We then introduce PBRS together with latent injection. All parameters optimized in Stage I are frozen, and only the BLRB and residual decoder are trained. The residual decoder is initialized from the decoder weights obtained in Stage I. This stage adapts the downstream reconstruction pathway from full-image decoding to base-relative residual completion without further modifying the diffusion U-Net, thereby reducing the risk of degrading its pretrained generative capability.

Stage III: Controllable Feature Injection. Finally, we introduce feature injection on top of the Stage-II model and jointly optimize the newly added injection modules. During training, the injection strength s is randomly sampled to expose the model to different fidelity–perception preferences. We further use an s-aware adversarial objective,

$$\mathcal{L}^s_{\mathrm{adv}} = (1 - s)\mathcal{L}_{\mathrm{adv}}, \tag{9}$$

such that stronger perceptual supervision is applied toward the perception-oriented operating point. This enables controllable fidelity–perception adjustment without requiring repeated diffusion inference.

Training Objectives. To balance reconstruction fidelity and perceptual quality, our training objective consists of a reconstruction loss $\mathcal{L}_{\text{rec}}$ and an adversarial loss $\mathcal{L}_{\text{adv}}$. We further adopt detail-aware weighting (DAW) (Kim et al., 2026) to adaptively emphasize regions containing rich structural details and perceptually important textures. The reconstruction objective combines weighted pixel-wise and perceptual losses (Zhang et al., 2018):

$$\mathcal{L}_{\text{rec}} = \lambda_{\text{mse}}\mathbb{E}\left[W_{\text{DAW}} \odot (x_r - (x_{\text{GT}} - b))^2\right] + \lambda_{\text{lpips}}\mathbb{E}\left[W'_{\text{DAW}} \odot \mathcal{L}_{\text{LPIPS_map}}(x_{\text{SR}}, x_{\text{GT}})\right], \quad (10)$$

where W_{DAW} and W'_{DAW} denote the detail-aware weights for the pixel and perceptual losses, respectively.

For adversarial training, we employ a Real-ESRGAN-style (Wang et al., 2021) spectrally normalized U-Net discriminator. The discriminator and generator adversarial objectives are

$$\mathcal{L}_D = \frac{1}{2}\mathbb{E}\left[\ell_{\text{BCE}}(D(x_{\text{GT}}), 1) + \ell_{\text{BCE}}(D(\text{sg}(x_{\text{SR}})), 0)\right], \quad (11)$$

and

$$\mathcal{L}_{\text{adv}} = \mathbb{E}\left[\ell_{\text{BCE}}(D(x_{\text{SR}}), 1)\right], \quad (12)$$

respectively. The overall generator objective is

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}, \quad (13)$$

with $\mathcal{L}_{\text{adv}}$ replaced by $\mathcal{L}^s_{\text{adv}}$ in Stage III.

4 Experiments

4.1 Experimental Setup

Datasets. Following FiDeSR (Kim et al., 2026), we adopt LSDIR (Li et al., 2023), DIV2K (Agustsson & Timofte, 2017), Flickr2K (Timofte et al., 2017), and the first 10K images from FFHQ (Karras et al., 2019) as training data. We use the Real-ESRGAN (Wang et al., 2021) degradation pipeline to generate LQ-HQ pairs for Real-ISR by degrading HQ images. For evaluation, we use the real-world paired benchmarks DRealSR (Wei et al., 2020) and RealSR (Cai et al., 2019), using their provided LQ-HQ pairs and center-cropped 128×128 LQ and 512×512 HQ image pairs. We additionally report results on the synthetic DIV2K validation protocol in Appendix A.1.

Evaluation Metrics. We employ both full-reference and no-reference metrics. For full-reference metrics, we use PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022), and FID (Heusel et al., 2017). For no-reference metrics, we use CLIPIQA (Wang et al., 2023), NIQE (Mittal et al., 2012), MUSIQ (Ke et al., 2021), and MANIQA (Yang et al., 2022).

Implementation Details. PReS-SR is implemented based on DreamLite-mobile (Feng et al., 2026), an unified on-device diffusion model that supports both image generation and image editing. The pretrained VAE and U-Net are kept frozen, and we add trainable LoRA layers to the U-Net and VAE decoder for fine-tuning, setting the LoRA rank to 8. All models are trained on NVIDIA RTX 3090 GPUs with a batch size of 8. We use the AdamW (Loshchilov & Hutter, 2019) optimizer with a learning rate of 1e-5. We use RAM (Zhang et al., 2024) to extract text prompts. When enabled, the corresponding loss weights are set to $\lambda_{\text{MSE}} = 1$, $\lambda_{\text{LPIPS}} = 2$, and $\lambda_{\text{adv}} = 0.1$, respectively.

4.2 Comparison with existing methods

To examine the restoration and inference performance of the proposed PReS-SR, we compare it with multiple Real-ISR methods, including 1) multi-step diffusion: StableSR (Wang et al., 2024a), SeeSR (Wu et al., 2024b), DiffBIR (Lin et al., 2024), PASD (Yang et al., 2024), AddSR (Tai et al., 2026), and HDW-SR (Yang et al., 2026); and 2) one-step diffusion: OSEDiff (Wu et al., 2024a), SinSR (Wang et al., 2024b), PiSA-SR (Sun et al., 2025), AdcSR (Chen et al., 2025), and FiDeSR (Kim et al., 2026). During inference, we use A mode with $s \in \{0, 1\}$ as the perception- and fidelity-oriented operating points, respectively, and the B mode with $s = 1$ as a balanced operating point, to show the fidelity–perception spectrum.

Table 1: Quantitative comparison with state-of-the-art methods on the real-world paired DRealSR and RealSR benchmarks. The best results for all methods are highlighted in bold and underlined, while the best one-step diffusion methods are reported in red and blue.

Dataset	Method	Metrics								
		PSNR↑	SSIM↑	LPIPS↓	DISTS↓	CLIPIQA↑	NIQE↓	MUSIQ↑	MANIQA↑	FID↓
DRealSR	StableSR-200s	27.93	0.7491	0.3306	0.2290	0.6197	6.3865	58.55	0.5569	147.48
	SeeSR-50s	28.14	0.7713	0.3141	0.2298	0.6889	6.4633	64.73	0.6016	146.98
	DiffBIR-50s	25.90	0.6245	0.4669	0.2298	0.6333	6.3275	66.13	0.6221	180.34
	PASD-20s	**29.17**	**0.7954**	0.3176	0.2220	0.7071	7.7138	50.44	0.4683	143.08
	AddSR-4s	26.62	0.7384	0.3695	0.2651	0.7121	7.8536	65.06	0.5993	167.69
	HDW-SR	28.43	0.7824	0.2960	0.2153	0.6970	6.2000	**66.36**	–	–
	OSEDiff-1s	27.92	0.7835	0.2967	0.2162	0.6956	6.4389	64.70	0.5898	135.45
	SinSR-1s	28.34	0.7490	0.3705	0.2487	0.6445	7.0300	55.37	0.4912	175.96
	PiSA-SR-1s	28.32	0.7804	0.2960	0.2169	0.6918	6.1803	66.11	0.6161	130.48
	AdcSR-1s	28.10	0.7606	0.3046	0.2200	0.7048	6.4444	66.27	0.5915	134.05
	FiDeSR-1s	28.90	0.7907	0.2836	0.2112	0.6974	6.2014	65.78	**0.6239**	127.97
	PReS-SR-A (s=0)	28.37	0.7557	0.2973	0.2100	**0.7185**	**5.4528**	63.54	0.5631	128.51
	PReS-SR-B	28.84	0.7638	0.2796	0.2020	0.7056	5.5554	62.55	0.5648	125.17
	PReS-SR-A (s=1)	28.89	0.7690	**0.2746**	**0.1985**	0.6617	5.6600	62.73	0.5624	**124.73**
RealSR	StableSR-200s	24.73	0.7084	0.3050	0.2156	0.6379	5.6251	65.44	0.6259	130.43
	SeeSR-50s	25.21	0.7216	0.3004	0.2218	0.6674	5.5929	69.69	0.6435	125.09
	DiffBIR-50s	24.83	0.6501	0.3650	0.2399	0.7053	5.8404	69.28	0.6502	130.52
	PASD-20s	**26.63**	**0.7660**	0.2869	0.2032	0.4857	5.9815	60.03	0.5390	124.52
	AddSR-4s	22.54	0.6419	0.3867	0.2718	**0.7298**	6.5640	**71.44**	**0.6751**	156.68
	HDW-SR	25.71	0.7428	0.2672	0.2044	0.6702	5.3900	70.10	–	–
	OSEDiff-1s	25.15	0.7341	0.3194	0.2127	0.6686	5.6364	69.08	0.6335	123.49
	SinSR-1s	26.29	0.7350	0.3194	0.2344	0.6181	6.2854	60.75	0.5420	134.27
	PiSA-SR-1s	25.50	0.7418	0.2672	0.2044	0.6697	5.5054	70.15	0.6551	124.18
	AdcSR-1s	25.47	0.7220	0.2885	0.2129	0.6731	5.3480	69.90	0.6353	118.42
	FiDeSR-1s	26.02	0.7457	0.2626	0.1965	0.6896	5.3194	69.82	0.6681	**109.68**
	PReS-SR-A (s=0)	25.72	0.7317	0.2700	0.2030	0.7054	**4.6897**	67.70	0.6039	117.00
	PReS-SR-B	26.28	0.7419	0.2516	0.1939	0.6936	4.8077	66.67	0.6008	110.29
	PReS-SR-A (s=1)	26.39	0.7490	**0.2451**	**0.1900**	0.6153	4.9151	66.44	0.5972	110.12

Table 2: Efficiency on an RTX 3090 at 512×512. (a) Parameters and median end-to-end latency. (b) Cumulative latency for five controlled outputs.

(a) Single output.

Method	Params (M)	E2E (ms)
AdcSR-1s	456.07	80.46
FiDeSR-1s	950.28	210.18
PReS-SR	419.13	115.55

(b) Five control outputs.

Method	Latency (ms)
PiSA-SR	708.66
CODSR	839.20
PReS-SR	190.30

Quantitative Comparison. In Table 1 and Table 2, we compare the proposed PReS-SR with existing SOTA methods in terms of restoration quality and inference efficiency. PReS-SR is more efficient than most existing one-step diffusion methods as shown in Table 2 and 7, with 1.8× faster inference and 56% fewer parameters than FiDeSR (Kim et al., 2026). The post-backbone modulation design also benefits efficiency for multiple control outputs as shown in Table 2 (b), because the upstream diffusion computation is shared. Moreover, PReS-SR with $s = 1$ achieves strong full-reference image quality assessment (FR-IQA) performance, while PReS-SR with $s = 0$ achieves strong no-reference image quality assessment (NR-IQA) performance (e.g., the best or second-best CLIPIQA and NIQE). The B mode with $s = 1$ provides a balanced operating point along the fidelity–perception trade-off, maintaining competitive performance across both FR- and NR-IQA metrics.

Qualitative Comparison. Figure 3 provides a visual comparison of PReS-SR against state-of-the-art diffusion-based super-resolution methods on challenging text and natural-texture scenes. Compared with competing methods, PReS-SR better preserves the appearance and local structures supported by the ground truth while maintaining realistic fine details. In the upper example, our method faithfully recovers the Chinese characters, whereas several baselines tend to generate distorted text or introduce artifacts. In the lower example, PReS-SR better preserves the fine, hair-like texture of the dark wing and the delicate venation of the green wing membrane, without introducing conspicuous hallucinations. The qualitative comparison in Figure 5 further supports these results: the base preserves global structure, while the predicted residual remains aligned with the character strokes and supplies complementary details. In contrast, direct full-image reconstruction exhibits distortions similar to those in the VAE-reconstructed GT, indicating the constraint imposed by the VAE bottleneck.

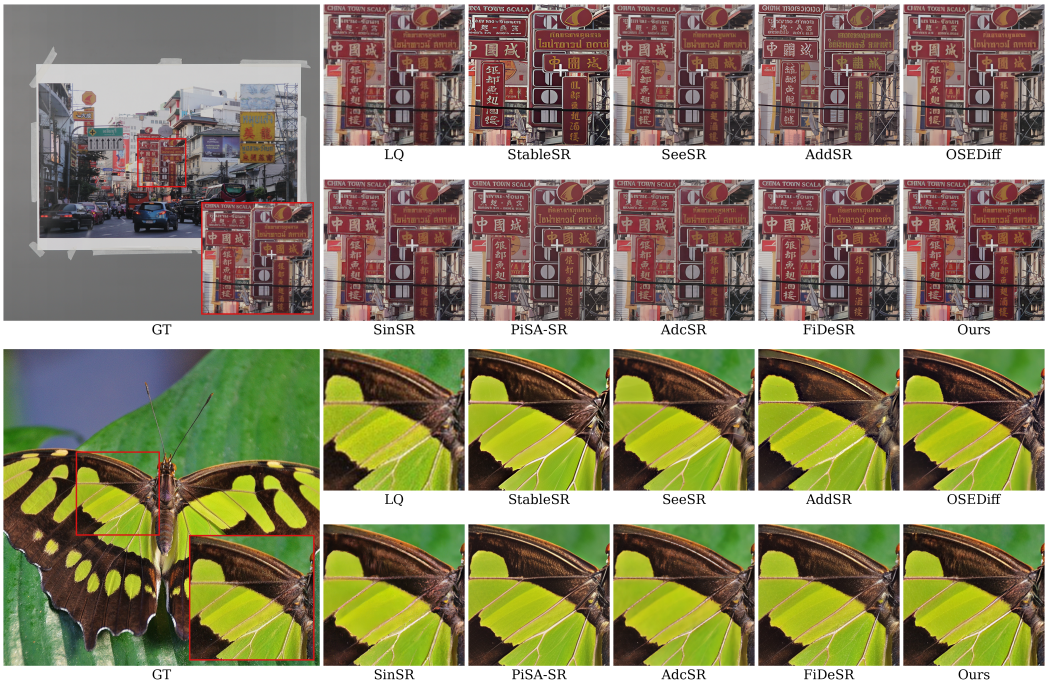

Figure 3: Qualitative comparisons with state-of-the-art SR methods.

Table 3: Ablation study on RealSR. We evaluate residual specialization, CSRI, and pretrained T2I prior. Best and second-best results among residual variants are bolded and underlined, respectively.

Method	Residual specialization		CSRI		T2I pretrain	PSNR↑	LPIPS↓	CLIPIQA↑	MANIQA↑
	Backbone	Post-backbone	Latent	Feature					
NAFNet base						26.82	0.2849	0.6469	0.5575
Full image					✓	25.94	0.2717	0.5948	0.6119
Base + residual	✓				✓	24.79	0.3623	0.6351	0.5090
		✓			✓	26.07	<u>0.2655</u>	0.6798	<u>0.6133</u>
		✓	✓		✓	26.09	0.2667	<u>0.6839</u>	**0.6153**
		✓		✓	✓	<u>26.28</u>	0.2657	0.6822	0.5943
		✓	✓	✓	✓	<u>26.28</u>	**0.2516**	**0.6936**	0.6008
		✓	✓	✓		**26.65**	0.2887	0.6696	0.5276

4.3 Ablation and Analysis

Table 3 evaluates the effects of residual specialization, cross-space restoration injection, and the pretrained T2I prior. More ablation results and analyses can be found in Appendix A.4.

Effect of Residual Specialization. Relative to the full-image topology, PBRS alone consistently improves all metrics, showing that PBRS improves performance by preserving base structures while adapting the generative pathway to residual completion. We also evaluate the variant with backbone-level residual specialization, which uses the same backbone but finetunes the U-Net to directly predict the residual through the generative pathway. Its marked performance degradation reflects the mismatch between the residual latent distribution and the pretrained VAE decoder, supporting the importance of placing base–residual adaptation after the backbone.

Effect of Cross-Space Restoration Injection. On top of PBRS, latent injection further improves CLIPIQA and MANIQA, whereas feature injection supplies multi-scale restoration cues during decoding and improves PSNR. Combining both achieves the best PSNR, LPIPS, and CLIPIQA among the pretrained residual variants. These results demonstrate their complementary roles: latent injection uses z_b to adaptively exploit the generative prior, while feature injection reinforces structural preservation through multi-scale decoding corrections.

Table 4: Topology–capacity interaction on RealSR. PReS-SR uses the default Mode-B setting. Bold denotes the better result within the same configuration.

Configuration	Mobile U-Net	TinyVAE	Method	PSNR↑	LPIPS↓	CLIPIQA↑	MUSIQ↑	Params↓
Full U-Net + Full VAE			Full image baseline	**26.25**	0.2604	0.6766	69.3646	950.3 M
			PReS-SR	26.11	**0.2560**	**0.6882**	**69.4299**	1060.4 M
Mobile U-Net + Full VAE	✓		Full image baseline	25.96	0.2794	0.6850	67.2562	474.3 M
	✓		PReS-SR	**25.96**	**0.2720**	**0.6949**	**67.3033**	584.4 M
Mobile U-Net + Tiny VAE	✓	✓	Full image baseline	25.94	0.2717	0.5948	**68.0334**	393.1 M
	✓	✓	PReS-SR	**26.28**	**0.2516**	**0.6936**	66.6688	419.13 M

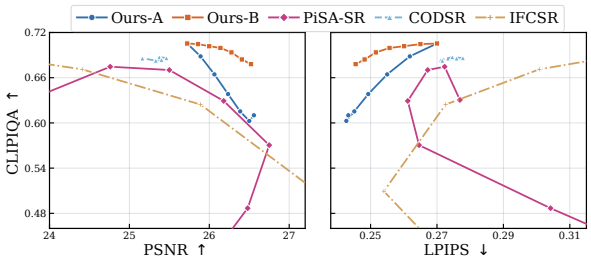

Figure 4: Comparison of controllable ISR methods on the RealSR dataset. PReS-SR, PiSA-SR, CODSR, and IFCSR vary s, λ_{pix}, t_s, and γ, respectively.

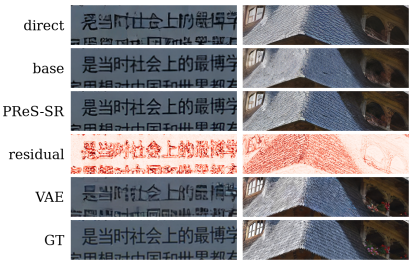

Figure 5: Qualitative results of Post-Backbone Residual Specialization with CSRI.

Effect of Pretrained T2I Prior. Under the same architecture and training configuration, randomly initializing the U-Net produces a metric profile close to the deterministic NAFNet base, suggesting that the generative pathway contributes limited complementary information without pretraining. In contrast, retaining the pretrained T2I prior substantially improves perceptual quality, indicating that the model can leverage the broad natural-image statistics learned from large-scale T2I pretraining to produce more natural and perceptually plausible reconstructions.

Topology–Capacity Interaction. Table 4 examines how reconstruction topology interacts with generative capacity by progressively reducing the U-Net and VAE capacity. With the full SD2.1 setting, PReS-SR improves LPIPS, CLIPIQA, and MUSIQ with a small PSNR trade-off. As the generative pathway is compressed, its advantage becomes more pronounced, extending to both fidelity and perceptual metrics in the most compact setting, except a slight decline for MUSIQ. This trend suggests that PBRS improves the quality attainable under constrained generative capacity by carrying reliably restored content through the pixel-space base and reserving the generative pathway for complementary details, thereby enabling substantially lighter backbone configurations and a favorable system quality–efficiency trade-off.

Experiments on Controllable SR. We evaluate PReS-SR under global (Mode A) and spatially adaptive (Mode B) control. Figure 4 reports the fidelity–perception trajectories on RealSR. Increasing s in Mode A shifts the operating point from perception toward fidelity. Mode B improves the trade-off over part of this range, attaining higher perceptual quality at comparable fidelity. Spatial weighting emphasizes base guidance in structure-rich regions while preserving generative flexibility in less constrained regions. Overall, PReS-SR enables competitive and continuous fidelity–perception adjustment. Detailed settings and additional analysis are provided in Appendix A.5.

5 Conclusion

We present PReS-SR, an efficient one-step Real-ISR framework that combines a discriminative base with pretrained full-image diffusion priors. PBRS retains full-image diffusion while specializing only post-backbone decoding for base-relative residual completion, and CSRI injects restoration cues across latent and feature spaces. Experiments demonstrate a strong efficiency–quality trade-off across real-world degradations, especially for lightweight models, and enable efficient fidelity–perception control through shared upstream diffusion computation.

AI USE STATEMENT

Generative AI tools were used in this work for three purposes: (1) to assist with drafting and polishing portions of the manuscript for clarity, grammar, and organization; (2) to support literature retrieval and discovery, including identifying potentially relevant prior work; and (3) to support research execution by assisting with experimental implementation and evaluation, as well as the analysis and presentation of experimental results. The authors checked all AI-assisted literature suggestions and factual claims against original sources, independently reviewed and validated all AI-assisted technical suggestions and outputs, and verified all reported experiments, numerical results, and conclusions. The authors take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 1122–1131. IEEE, 2017.

Aditya Arora, Zhengzhong Tu, Yufei Wang, Ruizheng Bai, Jian Wang, and Sizhuo Ma. Guidesr: Rethinking guidance for one-step high-fidelity diffusion-based super-resolution. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3973–3981. IEEE, 2025.

Jonghee Back, Jongju Kim, Jeong-Uk Kim, Eunjin Kim, and Minyong Jeon. Ifcsr: Inference-free fidelity-realism control for one-step diffusion-based real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38187–38197, 2026.

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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3086–3095. IEEE, 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28208–28220. IEEE, 2025.

Hao Chen, Junyang Chen, Jinshan Pan, and Jiangxin Dong. Bridging fidelity-reality with controllable one-step diffusion for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30584–30594, 2026.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pp. 17–33. Springer, 2022.

Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4621–4630, 2021.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(5):2567–2581, 2022.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23174–23184. IEEE, 2025.

Yushun Fang, Yuxiang Chen, Shibo Yin, Qiang Hu, Jiangchao Yao, Ya Zhang, Xiaoyun Zhang, and Yanfeng Wang. One-step diffusion transformer for controllable real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23440–23450, 2026.

Kailai Feng, Yuxiang Wei, Bo Chen, Yang Pan, Hu Ye, Songwei Liu, Chenqian Yan, and Yuan Gao. DreamLite: A lightweight on-device unified model for image generation and editing, 2026.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4396–4405. IEEE, 2019.

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

Aro Kim, Myeongjin Jang, Chaewon Moon, Youngjin Shin, Jinwoo Jeong, and Sang-hyo Park. Fidesr: High-fidelity and detail-preserving one-step diffusion super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38270–38280, 2026.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1646–1654, 2016.

Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5835–5843. IEEE, 2017.

Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1775–1787. IEEE, 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. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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

Shuyao Shang, Zhengyang Shan, Guangxing Liu, LunQian Wang, XingHua Wang, Zekai Zhang, and Jinglin Zhang. Resdiff: Combining cnn and diffusion model for image super-resolution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 8975–8983, 2024.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2333–2343. IEEE, 2025.

Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2790–2798. IEEE, 2017.

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.

Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 1110–1121. IEEE, 2017.

Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2555–2563, 2023.

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

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

Yan Wang, Shijie Zhao, Kai Chen, Kexin Zhang, Junlin Li, and Li Zhang. Gendr: Lightning generative detail restorator. *arXiv e-prints*, pp. arXiv–2503, 2025.

Yan Wang, Shijie Zhao, Junlin Li, et al. Eliminating vae for fast and high-resolution generative detail restoration. In *International Conference on Learning Representations*, volume 2026, pp. 105906–105924, 2026.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25796–25805. IEEE, 2024b.

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

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.

Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16272–16282. IEEE, 2022.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25456–25467. IEEE, 2024b.

Chao Yang, Boqian Zhang, Jinghao Xu, and Guang Jiang. Hdw-sr: High-frequency guided diffusion model based on wavelet decomposition for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23462–23472, 2026.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1190–1199, 2022.

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.

Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25669–25680. IEEE, 2024.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. Pixeldit: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14273–14282, 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 586–595. IEEE, 2018.

Tianyi Zhang, Zheng-Peng Duan, Chun-Le Guo, Peng-Tao Jiang, Bo Li, Ming-Ming Cheng, and Chongyi Li. Time-aware one step diffusion network for real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30543–30552, 2026.

Yi Zhang, Xiaoyu Shi, Dasong Li, Xiaogang Wang, Jian Wang, and Hongsheng Li. A unified conditional framework for diffusion-based image restoration. *Advances in Neural Information Processing Systems*, 36:49703–49714, 2023.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1724–1732, 2024.

A Appendix

A.1 Additional Quantitative Results

Results on Synthetic DIV2K-Val. For this synthetic protocol, LQ images are generated by applying the Real-ESRGAN degradation pipeline (Wang et al., 2021) to 3,000 cropped 512×512 DIV2K (Agustsson & Timofte, 2017) HQ images, yielding corresponding 128×128 LQ inputs.

Additional Comparison under Full-Resolution RealSR Evaluation. In addition to the 512×512 crop-based evaluation used in the main paper, we further evaluate PReS-SR on the 100 full-resolution RealSR $\times 4$ test pairs, providing a complementary comparison with recent Real-ISR methods that report results under this full-resolution setting rather than the crop-based protocol. Table 6 presents the quantitative comparison under this setting. PReS-SR maintains a strong fidelity–perception trade-off: PReS-SR-A ($s = 1$) achieves 26.57 dB PSNR and 0.7644 SSIM, exceeding GenDR-Pix by 1.54 dB and 0.0131, respectively, while also obtaining slightly better LPIPS and NIQE. At the perception-oriented endpoint, PReS-SR-A ($s = 0$) achieves higher CLIPIQA and MUSIQ, while Mode B provides a balanced operating point with strong fidelity and perceptual quality.

A.2 Additional Implementation Details

NAFNet backbone. We employ NAFNet (Chen et al., 2022) as the deterministic base-restoration backbone. It follows a four-level encoder–decoder architecture with additive skip connections. Each NAFBlock contains two residual branches: the first uses LayerNorm, 1×1 channel expansion, a 3×3 depthwise convolution, SimpleGate, and simplified channel attention, while the second uses a gated feed-forward transformation. Learnable residual scales stabilize both branches without conventional nonlinear activations. A stride-2 2×2 convolution downsamples each encoder level, whereas a 1×1 convolution followed by PixelShuffle upsamples each decoder level. In our configuration, the initial width is 32; the four encoder levels contain $(1, 1, 1, 28)$ NAFBlocks with channel widths $(32, 64, 128, 256)$, the 512-channel bottleneck contains one NAFBlock, and the decoder contains $(1, 1, 1, 1)$ NAFBlocks with widths $(256, 128, 64, 32)$. The three-channel output is added to the input through a global residual connection. This pretrained backbone is frozen, and its decoder features are detached before being supplied to the FRBs.

Latent Residual Refinement Block (LRRB). Both BLRB and FRB use the LRRB architecture (Kim et al., 2026). Given an input $q \in \mathbb{R}^{c_{\text{in}} \times H \times W}$, where c_{in} is its channel number and $H \times W$ is its spatial resolution, an input 1×1 convolution $\mathcal{C}_{\text{in}}$ maps it to the internal width m:

$$q_0 = \mathcal{C}_{\text{in}}(q), \qquad q_0 \in \mathbb{R}^{m \times H \times W}. \tag{14}$$

LRRB then applies N densely connected residual blocks. Let q_{k-1} denote the input to block k. Each block contains five 3×3 convolutions. The first four each produce g channels and use the SiLU activation ϕ:

$$d_{k,j} = \phi(\mathcal{C}_{k,j}([q_{k-1}, d_{k,1}, \ldots, d_{k,j-1}])), \quad j = 1, \ldots, 4, \quad k = 1, \ldots, N, \tag{15}$$

where the feature list preceding $d_{k,1}$ is empty and $[\cdot]$ denotes channel-wise concatenation. Here, g is the growth width and $d_{k,j}$ denotes the output of convolution j in block k. The fifth convolution maps the resulting $m + 4g$ channels back to m channels, and its output is added to the block input with residual scale α:

$$\mathcal{D}_k(q_{k-1}) = \mathcal{C}_{k,5}([q_{k-1}, d_{k,1}, d_{k,2}, d_{k,3}, d_{k,4}]), \qquad q_k = q_{k-1} + \alpha \mathcal{D}_k(q_{k-1}). \tag{16}$$

An output 1×1 convolution $\mathcal{C}_{\text{out}}$ maps the final internal feature to c_{out} channels:

$$\text{LRRB}(q) = \mathcal{C}_{\text{out}}(q_N) \in \mathbb{R}^{c_{\text{out}} \times H \times W}. \tag{17}$$

The fifth convolution $\mathcal{C}_{k,5}$ in every block and $\mathcal{C}_{\text{out}}$ are zero-initialized. Consequently, LRRB initially predicts a zero correction and leaves its host pathway unchanged.

Base-aware latent refinement block (BLRB). BLRB operates in the latent space after the backbone. It receives the four-channel LQ latent z_L, the four-channel residual latent z_r predicted by the diffusion U-Net, and the four-channel base latent z_b obtained by encoding the NAFNet base

Table 5: Quantitative comparison on 3,000 synthetically degraded DIV2K validation crops. LQ images are generated using the Real-ESRGAN degradation pipeline. The best results for all methods are highlighted in bold and underlined, while the best one-step diffusion methods are reported in red and blue.

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	CLIPIQA↑	NIQE↓	MUSIQ↑	MANIQA↑	FID↓
StableSR-200s	23.27	0.5733	0.3106	0.2045	0.6773	4.7621	65.83	0.6178	24.43
SeeSR-50s	23.73	0.6057	0.3193	0.1966	0.6864	4.7949	68.40	0.6200	25.80
DiffBIR-50s	23.14	0.5441	0.3669	0.2209	0.7299	4.9920	**69.87**	**0.6440**	32.71
PASD-20s	24.41	**0.6252**	0.3794	0.2218	0.5567	5.4273	61.27	0.5362	31.53
AddSR-4s	22.38	0.5554	0.3815	0.2340	**0.7538**	5.8168	69.16	0.6300	35.07
HDW-SR	**24.52**	0.6162	0.2823	0.1934	0.6937	4.4300	69.68	–	–
OSEDiff-1s	23.72	0.6109	0.2942	0.1975	0.6681	4.7104	67.97	0.6131	26.33
SinSR-1s	24.40	0.6018	0.3244	0.2067	0.6492	6.0002	62.87	0.5393	35.59
PiSA-SR-1s	23.87	0.6058	0.2823	0.1934	0.6928	4.5563	69.68	0.6375	25.09
AdcSR-1s	23.74	0.5921	0.2853	0.1899	0.6763	4.3577	68.00	0.6073	25.52
FiDeSR-1s	24.33	0.6250	**0.2678**	**0.1845**	0.6873	4.6644	68.87	0.6384	**23.30**
PReS-SR-A (s=0)	23.87	0.5753	0.2879	0.1982	0.7188	4.0677	67.48	0.5655	26.81
PReS-SR-B	24.18	0.5836	0.2775	0.1905	0.6984	**4.0533**	66.78	0.5722	25.21
PReS-SR-A (s=1)	24.16	0.5860	0.2759	0.1851	0.6758	4.0775	67.62	0.5770	24.25

Table 6: Comparison on full-resolution RealSR $\times 4$. Best results are highlighted in bold.

Method	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	CLIPIQA↑	MUSIQ↑
SinSR-1	25.9900	0.7072	0.4022	6.2412	0.6670	59.2200
OSEDiff-1	24.5700	0.7202	0.3036	4.3408	0.6829	67.3000
AdcSR-1	25.6900	0.7362	0.2825	4.1846	0.6695	67.4700
InvSR-1	24.5000	0.7262	0.2872	4.2218	0.6919	67.4700
GenDR-1	23.1800	0.7135	0.2859	4.1588	0.7014	**68.3600**
GenDR-Pix	25.0300	0.7513	0.2702	4.0751	0.6816	66.2900
PReS-SR-A ($s = 0$)	25.8733	0.7449	0.2882	**3.8739**	**0.7263**	66.9302
PReS-SR-B	26.4111	0.7549	0.2715	3.9059	0.7092	66.7717
PReS-SR-A ($s = 1$)	**26.5704**	**0.7644**	**0.2679**	4.0397	0.6980	66.2323

reconstruction. These tensors have the same spatial resolution and form a 12-channel input. BLRB predicts a four-channel correction Δz_r:

$$\Delta z_r = \mathrm{BLRB}([z_L, z_r, z_b]), \qquad z'_r = z_r + \Delta z_r, \qquad \hat{z}'_H = z_L - z'_r. \tag{18}$$

Here, z'_r is the refined residual latent and $\hat{z}'_H$ is the corresponding full-image latent. In our standard configuration, BLRB uses internal width $m = 128$, growth width $g = 64$, $N = 8$ dense blocks, and residual scale $\alpha = 1.0$; its channel mapping is $12 \to 128 \to 4$. The residual decoder decodes $\hat{z}'_H$ into an image residual, which is added to the NAFNet base reconstruction. This specializes the generative pathway for complementary residual recovery rather than complete-image reconstruction.

Feature Refinement Block (FRB). FRBs inject four NAFNet decoder features into the residual decoder. We index the levels by $i = 1, \ldots, 4$ from coarse to fine. For a 512×512 output, their channel numbers are $(c_i) = (256, 128, 64, 32)$ and their spatial resolutions are $(64 \times 64, 128 \times 128, 256 \times 256, 512 \times 512)$ in the same order. At level i, a 1×1 projection $\mathcal{P}_i$ maps f_i from c_i to 64 channels, and adaptive average pooling aligns it with the spatial size of the 64-channel residual-decoder feature h_i:

$$\hat{f}_i = \mathrm{Pool}_i(\mathcal{P}_i(f_i)), \qquad \Delta_i = \mathrm{FRB}_i([h_i, \hat{f}_i]), \qquad h'_i = h_i + s(W_i \odot \Delta_i). \tag{19}$$

Here, $\hat{f}_i$ is the aligned NAFNet feature, Δ_i is the 64-channel FRB correction, s is the global injection strength, and W_i is the spatial control map resized to level i and broadcast along the channel dimension; $\odot$ denotes element-wise multiplication, and h'_i is the updated decoder feature. Thus, every FRB maps the 128-channel concatenation $[h_i, \hat{f}_i]$ to 64 channels. From coarse to fine, their internal widths are $(m_i) = (64, 64, 32, 32)$, growth widths are $(g_i) = (32, 32, 16, 16)$, and dense-block numbers are $(N_i) = (2, 2, 2, 1)$; all use residual scale $\alpha_i = 0.2$. They are attached to zero-indexed residual-decoder layers $(0, 5, 10, 15)$.

Spatially adaptive feature injection. For the B mode, we construct the spatial control map from the NAFNet base $b \in [0, 1]$. We average its RGB channels, compute the Sobel gradient magnitude

Table 7: Extended single-output efficiency comparison on an RTX 3090 at 512×512 resolution. E2E denotes median end-to-end latency per image.

Method	Params (M)	E2E (ms)
OSEDiff-1s	949.56	216.28
SinSR-1s	173.92	353.90
PiSA-SR-1s	949.56	215.18
PReS-SR	419.13	115.55

Table 8: Component-wise complexity of PReS-SR at 512×512. Module latencies are isolated FP32 measurements, whereas total latencies are measured end to end. The loaded-model total additionally includes the non-executed TinyVAE decoder for the direct path.

Component	Params (M)	Single output			n controlled outputs		
		Calls	MACs (G)	Latency (ms)	Calls	MACs (G)	Latency (ms)
TinyVAE encoder ($\times 2$ forward)	1.22	2	122.325	14.77	2	122.325	14.77
NAFNet	17.11	1	63.817	30.37	1	63.817	30.37
Mobile U-Net	389.97	1	160.252	59.66	1	160.252	59.66
LRRB	0.72	1	2.948	1.37	1	2.948	1.37
BLRB	7.67	1	31.415	3.33	1	31.415	3.33
Residual decoder	1.22	1	70.675	8.53	n	$70.675n$	$8.53n$
FRBs (4 levels)	1.21	1 set	36.641	9.75	n sets	$36.641n$	$9.75n$
Total	**419.13**	–	**488.074**	**115.55 E2E**	–	$380.757 + 107.317n$	**190.30** ($n = 5$)

G, and normalize it by its image-wise 95th percentile τ_{95}. The full-resolution control map is

$$W = \mathcal{G}_\sigma(\mathrm{clip}(G/(\tau_{95} + \epsilon), 0, 1)^\gamma), \tag{20}$$

where $\mathcal{G}_\sigma$ denotes Gaussian smoothing. We use $\gamma = 1$, $\sigma = 1$, and $s = 1$ by default. Adaptive average pooling resizes W to each injection resolution to obtain W_i, which is broadcast across channels in $h_i' = h_i + s(W_i \odot \Delta_i)$. Consequently, well-restored edge-rich regions receive stronger structural guidance, whereas under-restored texture regions retain greater influence from the generative pathway. For the A mode, we set $W_i = 1$.

A.3 Extended Efficiency and Complexity Analysis

Extended Efficiency Comparison. Table 7 reports the one-step baselines omitted from the compact efficiency table in the main paper. PReS-SR is substantially smaller than OSEDiff and PiSA-SR and faster than all three baselines, while SinSR retains the lowest parameter count. IFCSR is omitted from the runtime comparison because no official implementation or pretrained model is publicly available at the time of submission.

Component-wise Complexity Analysis. Table 8 decomposes the computational cost of the major components in PReS-SR. Even with the lightweight Mobile U-Net adopted in our implementation, the diffusion backbone remains the largest computational component, requiring 59.66 ms and accounting for approximately 46.7% of the summed component-wise latency. This is expected, as it still performs the main generative modeling and contains most of the model capacity. In contrast, NAFNet contains only 17.11M parameters but requires 30.37 ms, reflecting its role in dense full-resolution structure recovery rather than high-capacity generative modeling. The lightweight VAE further keeps the latent-space interface inexpensive: although the encoder is invoked twice for the LQ input and the deterministic base, the two encodings together require only 14.77 ms. Finally, the modules introduced for post-backbone residual specialization remain comparatively lightweight, with BLRB, the residual decoder, and the four FRBs requiring 3.33 ms, 8.53 ms, and 9.75 ms, respectively. Overall, the breakdown reflects the intended division of labor in PReS-SR: structural reconstruction is handled by a compact but computation-intensive pixel-space branch, generative modeling remains concentrated in the diffusion backbone, and residual specialization is confined to a lightweight downstream pathway.

Efficiency of Controllable Inference. Table 8 isolates the cost of post-backbone control. The shared prefix—the two TinyVAE encodings, NAFNet, Mobile U-Net, LRRB, and BLRB—accounts for 109.50 ms, or 85.7% of the summed isolated module latency. Only the residual decoder and

Table 9: Additional ablation of cross-space restoration injection and base-aware latent refinement on RealSR.

Setting	Output	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	NIQE↓	MANIQA↑
NAFNet base	Base	26.82	0.7546	0.2849	0.6469	4.9740	0.5575
full-image baseline	Full image	25.94	0.7331	0.2717	0.5948	5.3716	0.6119
+ Latent injection	Full image	26.17	0.7349	0.2590	0.5702	5.6042	0.6077
+ Latent + Feature injection	Full image	26.72	0.7390	0.2809	0.6337	5.1616	0.5899
PBRS w/o BLRB	Base + residual	26.53	0.7450	0.2711	0.6729	5.0638	0.5943
PBRS	Base + residual	26.07	0.7361	0.2655	0.6798	4.8573	0.6133
PReS-SR	Base + residual	26.28	0.7419	0.2516	0.6936	4.8077	0.6008

Table 10: Ablation of the fidelity base model on RealSR. Base MACs are measured for a 512×512 input. The PReS-SR variants use the corresponding base model and inference in mode A (s=1).

Setting	Fidelity base	Base MACs (G)	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	CLIPIQA↑	NIQE↓	MUSIQ↑	MANIQA↑
Base only	NAFNet (4-level)	63.817	26.82	0.7546	0.2849	0.2140	142.6935	0.6469	4.9740	65.0803	0.5575
	NAFNet (5-level)	20.356	26.69	0.7543	0.2731	0.2099	138.0431	0.5846	5.0996	66.4029	0.5503
	MIMO	30.332	26.92	0.7371	0.3532	0.2761	212.7387	0.3546	6.2870	57.6690	0.4328
PReS-SR	NAFNet (4-level)	63.817	26.39	0.7490	0.2451	0.1900	110.12	0.6153	4.9151	66.44	0.5972
	NAFNet (5-level)	20.356	26.38	0.7435	0.2627	0.1977	112.8367	0.6517	4.8833	67.2579	0.5903
	MIMO	30.332	26.31	0.7361	0.2784	0.2088	111.7822	0.7015	4.9864	66.7678	0.5873

FRBs vary across operating points, contributing 18.28 ms (14.3%) per component-wise pass. Consequently, a single adjustable output requires only one diffusion-backbone evaluation, unlike methods that combine multiple endpoint predictions to construct an operating point (Sun et al., 2025; Back et al., 2026). For five outputs, sharing the upstream computation limits the measured E2E latency to 190.30 ms, only $1.65\times$ the single-output latency of 115.55 ms rather than a five-fold increase. Overall, the post-backbone design confines controllable adjustment to a lightweight downstream pathway while sharing the dominant generative computation across operating points.

A.4 Additional Mechanism Ablations

Table 9 complements the main ablation with SSIM and NIQE, as well as additional controls for the residual interface.

Ablation protocol. The NAFNet-only baseline uses the same training data and degradation pipeline as PReS-SR. Under backbone-level residual specialization, the U-Net itself is adapted to predict a residual, whereas PBRS freezes the full-image SR U-Net and moves residual adaptation after the backbone. The "PBRS w/o BLRB" variant removes latent refinement entirely and inputs the unrefined backbone representation $\hat{z}_H = z_L - z_r$ to the residual decoder. In contrast, the PBRS baseline retains BLRB but sets z_b to zero, preserving its residual-adaptation capacity without base-aware conditioning. Thus, the former comparison isolates the contribution of BLRB, whereas adding latent injection isolates the benefit of base-latent guidance under the same architecture. Feature injection is controlled similarly by retaining the FRBs but setting their predicted corrections to zero. The T2I-prior control instead randomly initializes the U-Net and otherwise follows the same losses and training configuration as PReS-SR.

Effect of Restoration Injection under Full-Image Reconstruction. Applying NAFNet cues to the full-image topology does not lead to consistent improvements. With both latent and feature injection, the model becomes substantially closer to the NAFNet base in its metric profile: fidelity improves markedly, while perceptual performance also shifts toward that of the deterministic reconstruction. This suggests that, when the decoder is still responsible for reconstructing the entire image, strong restoration injection tends to drive the decoder toward the already well-restored NAFNet solution, rather than helping the generative pathway contribute complementary details. In contrast, PBRS explicitly separates reliable reconstruction from generative refinement, providing a topology in which restoration cues can guide, rather than dominate, the pretrained generative pathway.

Effect of Base-aware Latent Refinement. Bypassing BLRB while retaining the residual decoder produces a similarly base-dominated tendency: all reported metrics shift toward those of the NAFNet

Table 11: Quantitative controllability sweep of PReS-SR on RealSR. Both modes use the same trained model and vary only the inference-time injection strength s.

Mode	s	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	CLIPIQA↑	NIQE↓	MUSIQ↑	MANIQA↑
A	0	25.72	0.7317	0.2700	0.2030	117.00	0.7054	4.69	67.70	0.6039
	0.25	25.89	0.7379	0.2618	0.1978	114.97	0.6883	4.76	67.51	0.6088
	0.50	26.06	0.7431	0.2550	0.1940	112.95	0.6644	4.83	67.32	0.6094
	0.75	26.23	0.7469	0.2493	0.1915	111.79	0.6383	4.87	67.01	0.6057
	1.00	26.39	0.7490	0.2451	0.1900	110.12	0.6153	4.92	66.44	0.5972
	1.25	26.50	0.7491	0.2427	0.1891	107.78	0.6026	4.91	65.67	0.5841
	1.50	26.56	0.7467	0.2433	0.1897	105.38	0.6101	4.85	64.68	0.5659
B	0	25.72	0.7317	0.2700	0.2030	117.00	0.7054	4.69	67.70	0.6039
	0.25	25.86	0.7344	0.2650	0.2003	115.51	0.7046	4.73	67.49	0.6052
	0.50	26.00	0.7371	0.2601	0.1978	113.63	0.7018	4.76	67.22	0.6051
	0.75	26.14	0.7396	0.2557	0.1957	111.93	0.6995	4.80	66.95	0.6039
	1.00	26.28	0.7419	0.2516	0.1939	110.29	0.6936	4.81	66.67	0.6008
	1.25	26.41	0.7440	0.2482	0.1925	108.70	0.6842	4.82	66.37	0.5957
	1.50	26.52	0.7457	0.2455	0.1915	107.43	0.6780	4.84	66.01	0.5888

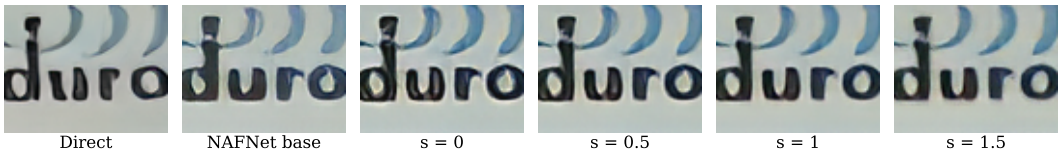

Figure 6: Visual comparison of PReS-SR under different feature-injection strengths s.

base, with higher PSNR and SSIM but weaker perceptual quality than PBRS. This suggests that, without latent refinement, the residual branch tends to make more conservative corrections, leaving the final reconstruction more strongly determined by the NAFNet base. BLRB refines the pretrained full-image latent representation for residual decoding, allowing the generative pathway to contribute complementary perceptual details beyond the deterministic reconstruction. Since removing BLRB also slightly changes model capacity, this comparison should be viewed as supporting evidence for the role of latent refinement rather than a strictly parameter-matched control.

Ablation on Fidelity Base Model. To examine whether PReS-SR depends on a specific fidelity backbone, we replace the default 4-level NAFNet with two structurally different alternatives. The compact 5-level NAFNet adds an extra 32-channel stage and retains the original channel widths $[32, 64, 128, 256]$ for stages 2–5, thereby moving these stages to half of their original spatial resolutions. The second alternative adapts MIMO-UNet (Cho et al., 2021), whose single encoder–decoder processes multi-scale inputs and produces multi-scale outputs, as the fidelity base. As shown in Table 10, PReS-SR improves LPIPS, DISTS, and FID over the corresponding base-only result for all three backbones, despite their markedly different standalone characteristics. This consistent improvement supports that PReS-SR is not tied to a particular fidelity backbone. The 5-level variant further reduces base computation from 63.817G to 20.356G MACs while retaining similar reconstruction fidelity, whereas the MIMO-UNet variant provides a higher-CLIPIQA operating point. We use the 4-level NAFNet in the main paper because it offers the strongest overall balance, with the best PSNR, SSIM, LPIPS, DISTS, and FID among the three PReS-SR variants.

A.5 Additional Fidelity–Perception Control Analysis

Experimental Setup and Comparison Protocol. We conduct the fidelity–perception control experiments on RealSR using the same trained PReS-SR model. Modes A and B are defined in Appendix A.2. In Figure 4, we sweep $s = 0, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5$ for both modes, with $s = 1.5$ testing extrapolation beyond the training range. We compare against PiSA-SR (Sun et al., 2025), CODSR (Chen et al., 2026), and IFCSR (Back et al., 2026) by varying their native control variables λ_{pix}, t_s, and γ, respectively. Operating points follow the respective papers under the same RealSR evaluation protocol; since IFCSR is not yet open-sourced, we use its reported results.

Quantitative and Qualitative Results on Controllability. Table 11 reports the complete control sweep. As s increases, both modes progressively improve fidelity-oriented metrics, including

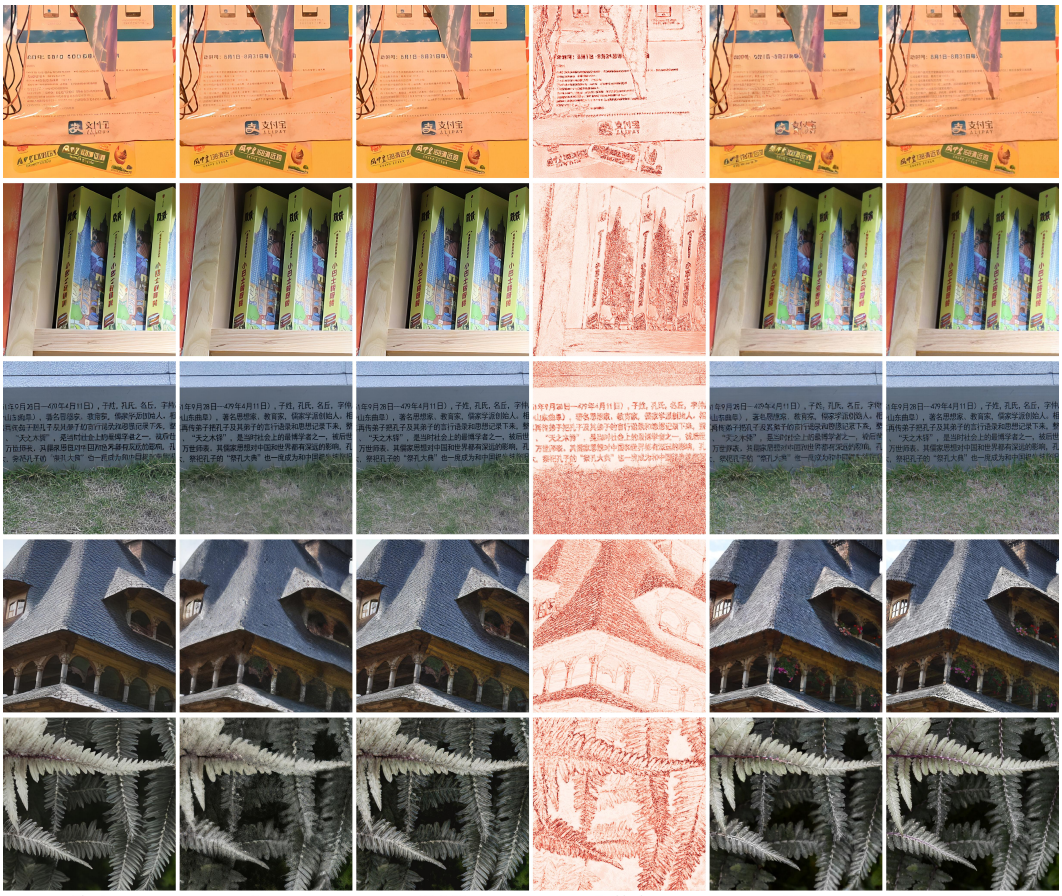

Figure 7: Additional qualitative results of Post-Backbone Residual Specialization with CSRI.

PSNR, LPIPS, DISTS, and FID. Mode B preserves CLIPIQA more effectively: at $s = 1$, it achieves 0.6936 CLIPIQA versus 0.6153 for Mode A at comparable PSNR (26.28 versus 26.39), supporting the benefit of spatially selective injection. At larger s, Mode B also better preserves no-reference perceptual quality, with lower NIQE and higher MUSIQ and MANIQA than Mode A. Qualitatively, Figure 6 visualizes Mode A as the global injection strength s varies from 0 to 1.5. Increasing s gradually modulates local details around the character strokes and radio arcs while largely preserving the global layout. Compared with the Direct pathway in Stage I, the controlled outputs better maintain character geometry and avoid visible structural distortions. These results complement the quantitative trajectories in Figure 4, demonstrating a smooth and visually interpretable fidelity–perception transition. Beyond the Sobel-derived training-free default used here, the explicit spatial weight W_i provides an interface for application-specific control. It could incorporate task-specific priors, such as text, face, or object masks, to emphasize fidelity where structural accuracy is critical.

A.6 Additional Qualitative Results

Figures 7–8 provide complementary visual evidence for residual specialization and restoration quality. In Figure 7, the NAFNet base preserves global geometry but lacks fine detail, while the predicted residual concentrates on missing high-frequency structures. Their combination restores text and texture more faithfully than direct full-image reconstruction, which can distort local structures. The similar distortions in the Direct and VAE results further indicate that direct full-image reconstruction is constrained by the VAE bottleneck. On synthetic DIV2K examples (Figure 8), PReS-SR preserves fine structures such as pufferfish spines and mushroom-gill structures while avoiding conspicuous over-smoothing or hallucinated patterns.

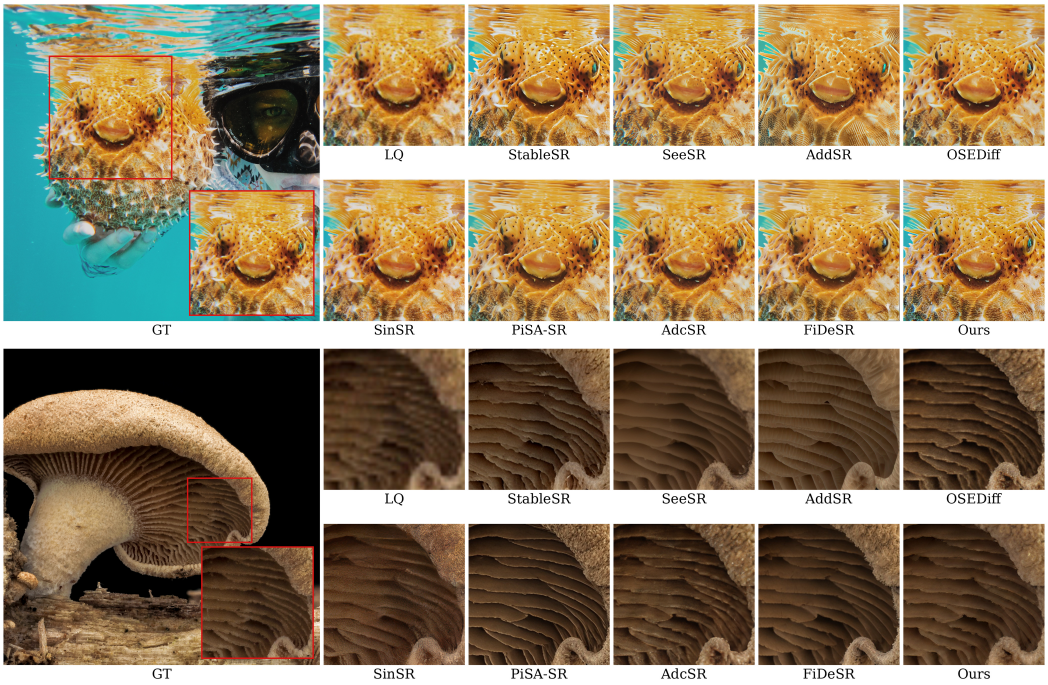

Figure 8: Qualitative comparison on synthetically degraded DIV2K images.



Figure 9: Failure cases under severe degradation. Each group shows LR, NAFNet, PReS-SR, and GT from left to right.

A.7 Limitations and Failure Analysis

Figure 9 shows representative failures under severe input degradation. In the first example, NAFNet cannot reliably recover the regular glass structure of the distant building, and PReS-SR consequently retains local geometric distortions. In the second, heavily degraded small characters remain blurred or malformed. PReS-SR carries the discriminative base directly to the output, avoiding VAE reconstruction for structures already recovered by NAFNet. However, when such fine structures are absent from the base, their recovery still depends on the VAE-based generative residual pathway and therefore remains limited by its representation capacity.