FREQUENCY-GUIDED TEMPORAL CORRECTION FOR HALLUCINATION SUPPRESSION IN DIFFUSION-BASED REAL-WORLD IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models perform strongly in real-world image super-resolution (Real-ISR), yet their generative priors can introduce structural distortions and spurious details. Without ground truth (GT) at inference, distinguishing useful refinement from hallucinations remains challenging. Using GT only for offline analysis, we examine how deviations from a selected early prediction and recent prediction history relate to reconstruction errors within each frequency band. We find a frequency-dependent reference preference: the selected early prediction provides stronger error-related cues in the low-frequency band, whereas recent prediction history is more informative in the high-frequency band. Based on this finding, we propose Frequency-Guided Temporal Correction (FGTC), a training-free and plug-and-play framework for multi-step diffusion-based Real-ISR. FGTC selects a structurally reliable early low-frequency reference and maintains recent high-frequency prediction history, then performs frequency-decoupled correction with deviation-adaptive strength before feeding the corrected prediction back into the original reverse diffusion process. Experiments on three benchmarks with StableSR and SeeSR show consistent improvements in reconstruction fidelity, including PSNR gains of 0.92–1.31 dB, while reducing structural distortions and spurious textures without backbone retraining or GT at inference.

1 INTRODUCTION

Real-world image super-resolution (Real-ISR) aims to recover high-resolution (HR) images from low-resolution (LR) observations affected by unknown degradations. Because the input does not uniquely determine the missing image content, restoring visually convincing details does not necessarily recover the underlying scene. Practical degradation modeling and generative adversarial network (GAN) models have substantially advanced this task (Zhang et al., 2021; Wang et al., 2021; Chen et al., 2022; Liang et al., 2022). More recently, pretrained diffusion models have provided strong generative priors, enabling realistic texture synthesis under complex degradations (Wang et al., 2024a; Lin et al., 2024; Wu et al., 2024b). However, these priors can also introduce plausible yet incorrect structures and details during iterative refinement. Figure 1(a) illustrates this problem through hallucinated window boundaries that deviate from the ground truth (GT). Suppressing such hallucinations requires constraining erroneous generation while preserving useful detail refinement.

Existing approaches improve content consistency by separating structure refinement from detail enhancement (Sun et al., 2025a), guide subsequent denoising with an early prediction selected for its structural agreement with the LR input (Li et al., 2025), or stabilize sampling through frequency decomposition and moving averages of historical predictions (Qian et al., 2024). These strategies motivate using predictions from the sampling trajectory as references. We refer to a previous clean prediction or an aggregation of such predictions as a *temporal reference*. However, it remains unclear which historical predictions provide suitable references for correcting different frequency components. Early predictions may provide useful structural guidance while lacking developed details, whereas recent predictions contain richer textures but may also include accumulated distortions. Without GT during inference, agreement with a historical reference alone does not establish reconstruction accuracy, and deviations from it may reflect either useful refinement or erroneous generation. Since low-frequency components tend to capture coarse spatial variations, while high-

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.

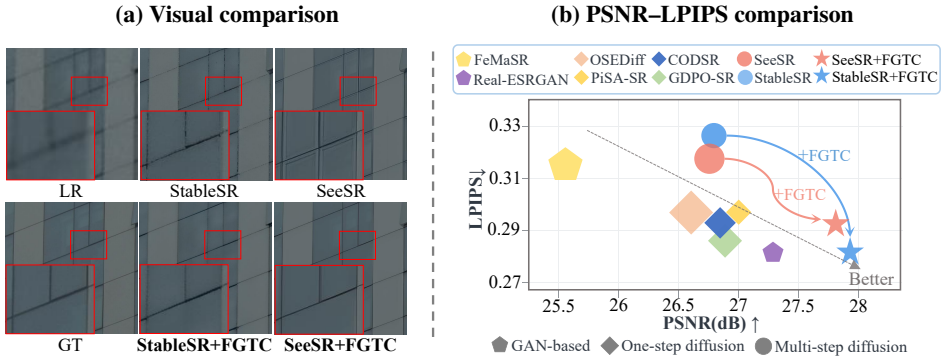


Figure 1: FGTC suppresses hallucinations in diffusion-based Real-ISR. (a) Diffusion SR hallucinates window edges; FGTC reduces these artifacts and corrects the boundaries. (b) PSNR–LPIPS comparison with state-of-the-art Real-ISR methods on DRealSR. FGTC consistently improves both StableSR and SeeSR in the PSNR–LPIPS space and achieves competitive performance against the compared methods. LR and GT denote the low-resolution input and ground truth, respectively.

frequency components emphasize local changes and textures, we ask: *which temporal reference provides more informative error cues within each frequency band?*

To investigate this question, we use GT only for offline analysis of decoded clean predictions along the diffusion sampling trajectory. Within each frequency band, we compare the current prediction with a selected early prediction and recent prediction history, and examine how the resulting deviations correlate with reconstruction errors across sampling steps. This analysis reveals a frequency-dependent reference preference: deviations from the selected early prediction correlate more strongly with low-frequency reconstruction errors, whereas deviations from recent prediction history correlate more strongly with high-frequency reconstruction errors (Figure 2).The bands thus favor distinct temporal references for correction during sampling.

Motivated by this frequency-dependent reference preference, we propose Frequency-Guided Temporal Correction (FGTC), a training-free, plug-and-play framework for hallucination suppression in multi-step diffusion-based Real-ISR. Its Low-frequency Reference Selection (LRS) module selects a structurally reliable clean prediction during the early sampling stage based on its low-frequency structural consistency with the LR input. Its Frequency-Decoupled Correction (FDC) module then uses the selected low-frequency component and recent high-frequency prediction history to correct subsequent clean predictions, with reference deviations adaptively controlling the correction strength. The corrected prediction is fed back into the original reverse diffusion process. FGTC thus applies frequency-specific temporal constraints without requiring GT during inference or retraining the backbone. Experiments on three benchmarks demonstrate improved reconstruction fidelity over both StableSR and SeeSR. Figure 1(b) shows consistent PSNR–LPIPS improvements for both backbones on DRealSR compared with competing methods.

Our main contributions are summarized as follows:

- We reveal a frequency-dependent preference for temporal references by analyzing, within each frequency band, how deviations from different temporal references correlate with reconstruction errors along the diffusion trajectory. The analysis shows that the selected early reference provides stronger error-related cues in the low-frequency band, whereas recent prediction history is more informative in the high-frequency band.
- We propose Frequency-Guided Temporal Correction (FGTC), a training-free and plug-and-play framework that translates this finding into frequency-specific temporal correction. FGTC selects a reliable early low-frequency reference for structural correction and exploits recent high-frequency prediction history for detail correction, with reference deviations adaptively controlling the correction strength during multi-step diffusion sampling.
- Extensive experiments on three benchmarks and two diffusion-based Real-ISR backbones demonstrate consistent improvements in reconstruction fidelity and hallucination suppression. Comprehensive ablations further validate the effectiveness of the correction components, frequency-specific reference assignment, and round-trip control.

2 Related Work

Diffusion-based Real-ISR. Pretrained diffusion models provide strong generative priors for restoring images with complex, unknown degradations. StableSR (Wang et al., 2024a) adapts latent diffusion priors to Real-ISR, while DiffBIR (Lin et al., 2024) separates degradation removal from generative reconstruction. Conditioning mechanisms further connect these priors to the degraded observation: PASD (Yang et al., 2024) introduces pixel-aware guidance, and SeeSR (Wu et al., 2024b) incorporates degradation-aware semantic information. SUPIR (Yu et al., 2024) explores large-scale generative priors for photo-realistic restoration, and DiT4SR (Duan et al., 2025) extends the diffusion transformer (DiT) architecture to Real-ISR. ResShift (Yue et al., 2023) shortens diffusion sampling through residual shifting. For efficient inference, OSEDiff (Wu et al., 2024a) and PiSA-SR (Sun et al., 2025b) use one-step restoration; AdcSR (Chen et al., 2025a) compresses a one-step diffusion model, and InvSR (Yue et al., 2025) supports a flexible number of steps through diffusion inversion. These advances improve restoration quality and sampling efficiency. However, strong generative priors may also introduce plausible content that is insufficiently supported by the degraded observation, raising concerns about restoration fidelity.

Fidelity Control in Diffusion SR. Existing approaches improve fidelity through input alignment, artifact-aware guidance, and structural constraints. FaithDiff (Chen et al., 2025b) aligns degraded-image features with the diffusion process and jointly fine-tunes the encoder and diffusion model for faithful reconstruction. SARGD (Zheng et al., 2024) detects implausible regions in latent space and applies adaptive reality guidance to mitigate artifacts during sampling. StructSR (Li et al., 2025) instead exploits an early prediction from the sampling trajectory as structural guidance for subsequent denoising. Recent one-step methods address the fidelity–realism balance through region-adaptive generation in CODSR (Chen et al., 2026), image-space control in IFCSR (Back et al., 2026), and controllable diffusion-transformer generation in ODTSR (Fang et al., 2026). FiDeSR (Kim et al., 2026) additionally uses frequency-adaptive enhancement to preserve reconstruction details. These studies show that restoration fidelity can be improved by constraining diffusion generation with information from the degraded input or intermediate predictions along the sampling trajectory, motivating further investigation into how structural and detail information evolves during sampling.

Frequency- and Time-Aware Restoration. A complementary line of research exploits the frequency and temporal characteristics of diffusion restoration. FDDiff (Wang et al., 2024b) uses frequency-domain refinement to progressively recover missing high-frequency information, while HDW-SR (Yang et al., 2026) incorporates wavelet decomposition and explicit high-frequency guidance. DTPSR (Jiang et al., 2026) separates low- and high-frequency semantic priors for frequency-aware guidance. Jointly exploiting frequency and temporal information, MASF (Qian et al., 2024) maps intermediate predictions to clean data space and aggregates them separately across frequency components to stabilize sampling. IAFS (Zhang et al., 2025) combines iterative refinement with frequency-aware particle fusion to balance structural fidelity and perceptual detail. TADSR (Zhang et al., 2026) uses different diffusion timesteps as controllable generative priors in a one-step model. These studies exploit frequency and temporal information for enhancement, aggregation, and refinement. We further investigate how the choice of temporal reference affects the correlation between prediction deviations and reconstruction errors within each frequency band, informing separate structural and detail correction.

3 Method

3.1 Frequency-Dependent Temporal Reference Analysis

Early predictions provide structural guidance (Li et al., 2025), but lack later-emerging details, so deviations from them may reflect useful refinement rather than error. Since low- and high-frequency components emphasize different aspects of image content, we compare early and recent-history references within each band, using GT only offline to assess deviation–error correlations.

We use StableSR for Real-ISR with $T = 200$ sampling steps, indexed by $t = T, \dots, 1$. At each step, we decode the uncorrected clean latent prediction $Z_{0|t}$ as $x_t = \mathcal{D}(Z_{0|t})$, where $\mathcal{D}$ is the frozen decoder. To compare frequency components, we apply a one-level 2D Haar discrete wavelet transform (DWT), obtaining $(x_t^{LL}, x_t^{HF}) = \mathrm{DWT}(x_t)$. Here, x_t^{LL} is the low-frequency (LL) subband, and $x_t^{HF} = \{x_t^b\}_{b\in\mathcal{B}}$ contains the high-frequency (HF) subbands, with $\mathcal{B} = \{\mathrm{LH}, \mathrm{HL}, \mathrm{HH}\}$. L and

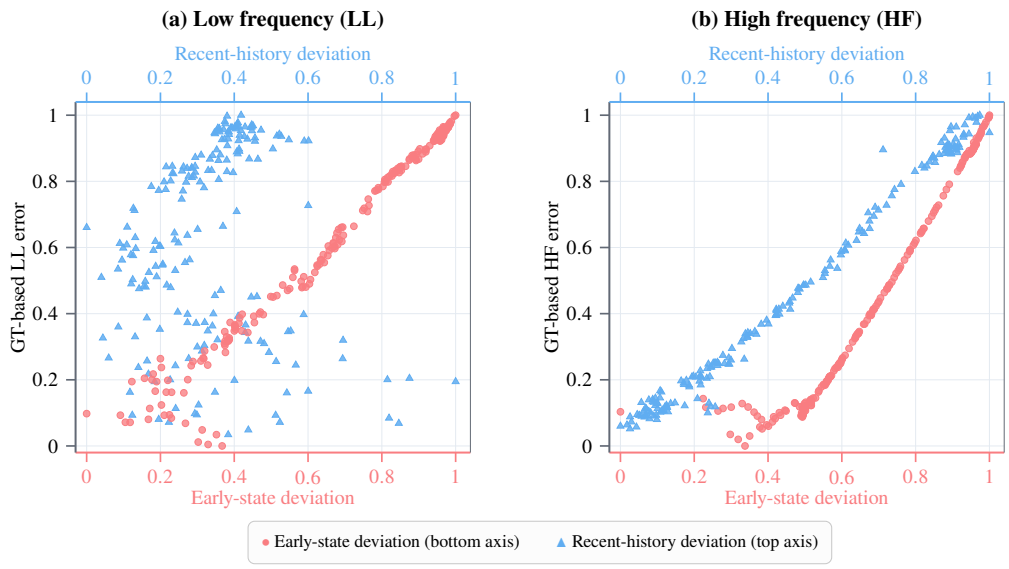


Figure 2: Representative scatter plots of reference deviations and GT-based reconstruction errors for (a) LL and (b) HF components. Red circles denote deviations from the selected early prediction, and blue triangles denote deviations from recent history. Each point represents a sampling step. Early-reference deviations show a stronger positive monotonic association in LL, whereas recent-history deviations do so in HF.

H denote low- and high-pass filtering along each spatial dimension. We decompose GT and both temporal references identically for comparison.

The early reference is selected by LL structural consistency with LR; recent history averages the latest K preceding clean predictions. For $q \in \{\text{early}, \text{recent}\}$, reference coefficients are denoted by $x^{LL}_{t,q}$ and $x^{b}_{t,q}$. For coefficient arrays u and v, we define their relative deviation as

$$\delta(u, v) = \frac{\|u - v\|_1}{\|v\|_1 + \varepsilon}, \tag{1}$$

where $\|\cdot\|_1$ sums absolute coefficient values and $\varepsilon > 0$ prevents division by zero. With GT coefficients denoted by x^{LL}_{GT} and x^{b}_{GT}, reconstruction errors e and reference deviations d are

$$\begin{aligned} e^{LL}_t &= \delta(x^{LL}_t, x^{LL}_{\text{GT}}), \quad e^{HF}_t = \frac{1}{3}\sum_{b\in\mathcal{B}} \delta(x^{b}_t, x^{b}_{\text{GT}}), \\ d^{LL}_{t,q} &= \delta(x^{LL}_t, x^{LL}_{t,q}), \quad d^{HF}_{t,q} = \frac{1}{3}\sum_{b\in\mathcal{B}} \delta(x^{b}_t, x^{b}_{t,q}). \end{aligned} \tag{2}$$

Using these error and deviation sequences, we quantify how well each reference deviation tracks reconstruction error. For each image and reference q, we compute Spearman rank correlations across sampling steps between $d^{LL}_{t,q}$ and e^{LL}_t, and between $d^{HF}_{t,q}$ and e^{HF}_t, then average the per-image correlations across images. Within each band, both references are evaluated against corresponding GT-based reconstruction error. Supplementary image-averaged statistics appear in Appendix A.

Our analysis reveals a frequency-dependent preference for temporal references: the selected early prediction shows a stronger correlation with LL reconstruction errors, whereas recent prediction history is more correlated with HF errors. This pattern is consistent with the distinct evolution of low- and high-frequency content during sampling. Low-frequency structures tend to stabilize early, allowing the selected early prediction to provide a stable anchor for structural correction. High-frequency details continue to develop, making recent prediction history a reference that follows the current stage of detail refinement. Based on these findings, we design Frequency-Guided Temporal Correction (FGTC) to use different temporal references for structural and detail correction. Its

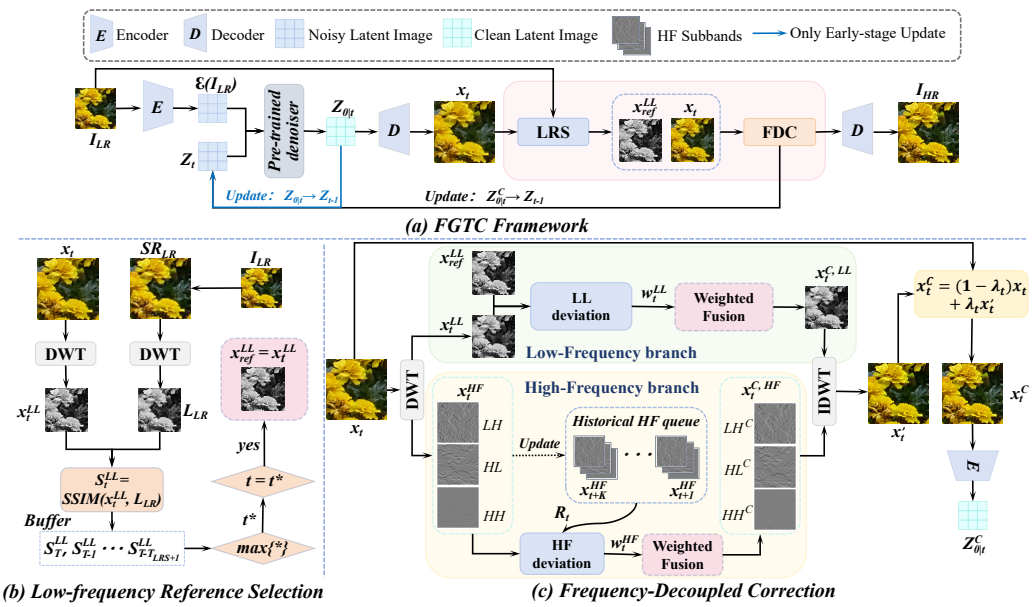


Figure 3: Overview of FGTC. LRS selects an early clean prediction based on its low-frequency structural consistency with the LR input. FDC uses the selected LL component to correct structural distortion and recent HF history to suppress spurious details while preserving useful refinement. The corrected components are reconstructed and blended with the current prediction, then encoded for the sampling update.

Low-frequency Reference Selection (LRS) module selects an early LL reference based on its structural consistency with the LR input. Its Frequency-Decoupled Correction (FDC) module then uses the selected LL reference and recent HF prediction history to correct subsequent predictions, with frequency-specific deviations adapting the correction strength. FGTC operates within the original sampling process without GT at inference or backbone retraining.

3.2 Method Overview

FGTC is a training-free, plug-and-play framework that translates the frequency-dependent temporal-reference preference identified in Sec. 3.1 into frequency-specific correction. It consists of Low-frequency Reference Selection (LRS) and Frequency-Decoupled Correction (FDC). LRS selects a reliable early clean prediction based on its low-frequency structural consistency with the LR input. FDC then decomposes subsequent clean predictions into wavelet components, using the selected early LL reference for structural correction and recent HF prediction history for detail correction. Frequency-specific reference deviations adaptively control the correction strength. The corrected components are reconstructed, blended with the current clean prediction, and re-encoded for the original sampling update. FGTC requires only the LR input and information from the sampling trajectory, without backbone retraining or GT at inference.

3.3 Low-frequency Reference Selection

Early clean predictions provide useful structural information, but their quality varies across images and sampling timesteps. Since high-frequency details remain under development during early sampling, directly evaluating the full prediction may be affected by incomplete textures. LRS therefore evaluates the LL components of early predictions and selects the one with the highest structural consistency with the LR observation as the reference for subsequent correction.

We bicubically upsample the LR input I_{LR} to the target resolution, denoting the result by $SR_{\mathrm{LR}} = \mathrm{Bicubic}(I_{\mathrm{LR}})$ and its Haar LL component by $L_{\mathrm{LR}} = \mathrm{DWT}_{LL}(SR_{\mathrm{LR}})$. During the first T_{LRS} reverse sampling steps, LRS scores each current LL component using $S_t^{LL} = \mathrm{SSIM}(x_t^{LL}, L_{\mathrm{LR}})$.

The current score is first added to a buffer, and $S_{\max}$ denotes the largest score in that buffer, including the current step. The selected low-frequency reference x_{ref}^{LL} stores the LL component associated with the highest score observed so far and is updated as

$$x_{\text{ref}}^{LL} \underset{t\in(T-T_{\text{LRS}},\,T]}{\longleftarrow} \begin{cases} x_t^{LL}, & S_t^{LL} = S_{\max}, \\ x_{\text{ref}}^{LL}, & \text{otherwise.} \end{cases} \tag{3}$$

Here, t takes integer sampling-step indices, so the interval contains exactly T_{LRS} steps. The first candidate initializes x_{ref}^{LL}; a later candidate replaces it when its score equals the buffer maximum, including ties. The score buffer contains only scalars, while LRS retains the selected LL array. Selection favors low-frequency structural consistency with LR without requiring GT. After the search interval, x_{ref}^{LL} remains fixed for subsequent correction.

3.4 Frequency-Decoupled Correction

For $1 \le t \le T - T_{\text{LRS}}$, FDC corrects LL using the fixed x_{ref}^{LL} and HF using recent prediction history.

Historical HF reference. To follow detail evolution without tying the correction to the incomplete early prediction, FDC maintains a first-in-first-out buffer of up to K preceding uncorrected HF predictions. At step t, for each subband $b \in \mathcal{B}$, the reference R_t^b is the arithmetic mean of the available entries in that subband. Because the reverse trajectory is traversed with decreasing t, the stored index $t + i$ denotes a previously visited step. When the buffer is not full, we average all available entries; when it is empty, we set $R_t^b = x_t^b$, so the first correction step is well-defined. The current uncorrected HF subbands are appended only after correction, and the oldest entries are discarded once the buffer exceeds K. This update rule smooths isolated fluctuations while retaining the recent direction of detail refinement.

Adaptive correction. The offline analysis identifies the reference assignment, while the observable deviation at the current step determines how strongly that reference should be used. Using δ from Equation 1, we compute the LL and HF deviations and map them to retention weights:

$$\begin{aligned} d_t^{LL} &= \delta(x_t^{LL}, x_{\text{ref}}^{LL}), \qquad w_t^{LL} = \exp(-d_t^{LL}), \\ d_t^{HF} &= \frac{1}{3}\sum_{b\in\mathcal{B}} \delta(x_t^b, R_t^b), \quad w_t^{HF} = \exp(-d_t^{HF}). \end{aligned} \tag{4}$$

Because $d_t^{LL}, d_t^{HF} \ge 0$, the exponential map gives $w_t^{LL}, w_t^{HF} \in (0, 1]$. Thus, a small deviation keeps the current prediction, whereas a large deviation lowers the retention weight and increases the contribution from the corresponding reference. With superscript C denoting corrected quantities, we write

$$\begin{aligned} x_t^{C,LL} &= w_t^{LL} x_t^{LL} + (1 - w_t^{LL}) x_{\text{ref}}^{LL}, \\ x_t^{C,b} &= w_t^{HF} x_t^b + (1 - w_t^{HF}) R_t^b, \qquad b \in \mathcal{B}. \end{aligned} \tag{5}$$

The same HF weight is shared by LH, HL, and HH to keep the correction strength consistent across high-frequency orientations, while the three subbands are blended independently with their corresponding references. This design preserves the directional characteristics of each subband and avoids unintended information mixing across orientations.

Sampling update. After frequency-specific correction, we collect the HF subbands as $x_t^{C,HF} = \{x_t^{C,b}\}_{b\in\mathcal{B}}$ and apply the inverse discrete wavelet transform (IDWT) to reconstruct

$$x'_t = \text{IDWT}(x_t^{C,LL}, x_t^{C,HF}). \tag{6}$$

The reconstructed image is then blended with the original clean prediction using a timestep-dependent strength:

$$x_t^C = (1 - \lambda_t) x_t + \lambda_t x'_t, \qquad \lambda_t = \frac{t}{T - T_{\text{LRS}}}. \tag{7}$$

Since t decreases from $T - T_{\text{LRS}}$ toward 1, λ_t gradually decreases from 1, applying stronger correction earlier and relaxing it during late refinement. We encode the blended image with the frozen encoder to obtain the corrected clean latent prediction $Z_{0|t}^C = \mathcal{E}(x_t^C)$, using its posterior mean with the backbone's normalization and latent scaling. Given the current noisy latent Z_t, the original sampler uses $Z_{0|t}^C$ to compute Z_{t-1} without changing its update rule or noise schedule; LRS steps use the uncorrected $Z_{0|t}$.

Table 1: Quantitative comparison with state-of-the-art Real-ISR methods on DRealSR, RealSR, and DIV2K-Val. The best and second-best results for each dataset and metric are marked in red and blue, respectively. Arrows indicate whether higher or lower values are better. Light-gray columns denote StableSR and SeeSR with FGTC.

Dataset	Metric	GAN-based		One-step diffusion				Multi-step diffusion					
		Real-ESRGAN	FeMaSR	OSEDiff	PiSA-SR	CODSR	GDPO-SR	SeeSR	StableSR	DiffBIR	DiT4SR	SeeSR +FGTC	StableSR +FGTC
DRealSR (real-world)	PSNR ↑	27.29	25.55	26.60	27.00	26.85	26.89	26.75	26.79	25.36	24.48	27.82	27.93
	SSIM ↑	0.7766	0.7235	0.7537	0.7506	0.7461	0.7536	0.7373	0.7190	0.6085	0.6442	0.7832	0.7844
	MS-SSIM ↑	0.8711	0.8451	0.8645	0.8680	0.8666	0.8604	0.8625	0.8561	0.8073	0.8068	0.8738	0.8788
	LPIPS ↓	0.2819	0.3157	0.2967	0.2960	0.2924	0.2858	0.3174	0.3268	0.4516	0.3659	0.2924	0.2816
	DISTS ↓	0.2089	0.2239	0.2163	0.2169	0.2111	0.2116	0.2315	0.2248	0.2703	0.2438	0.2073	0.2005
RealSR (real-world)	PSNR ↑	24.33	23.74	23.83	24.18	24.03	24.17	23.83	23.32	23.36	22.08	24.75	24.41
	SSIM ↑	0.7337	0.7068	0.7037	0.7122	0.6977	0.7024	0.6904	0.6751	0.6214	0.6305	0.7257	0.7351
	MS-SSIM ↑	0.8704	0.8685	0.8579	0.8671	0.8603	0.8520	0.8610	0.8523	0.8295	0.8262	0.8663	0.8729
	LPIPS ↓	0.2726	0.2937	0.2920	0.2672	0.2751	0.2672	0.3007	0.3056	0.3610	0.3194	0.2656	0.2601
	DISTS ↓	0.2066	0.2286	0.2128	0.2044	0.2003	0.1977	0.2224	0.2164	0.2320	0.2237	0.2001	0.1966
DIV2K-Val (synthetic)	PSNR ↑	22.97	21.74	22.40	22.54	22.46	22.60	22.36	21.94	22.33	20.49	23.38	23.25
	SSIM ↑	0.5932	0.5449	0.5664	0.5597	0.5614	0.5661	0.5592	0.5243	0.5200	0.4984	0.6056	0.5991
	MS-SSIM ↑	0.8192	0.7920	0.8011	0.8045	0.8005	0.7924	0.7972	0.7828	0.7794	0.7472	0.8119	0.8154
	LPIPS ↓	0.3113	0.3126	0.2942	0.2823	0.2878	0.2896	0.3194	0.3119	0.3512	0.3449	0.3321	0.2801
	DISTS ↓	0.2141	0.2057	0.1975	0.1934	0.1929	0.1963	0.1968	0.2052	0.2120	0.2107	0.2111	0.1911

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets and Metrics. We evaluate on synthetic DIV2K-Val and real-world RealSR and DRealSR (Agustsson & Timofte, 2017; Cai et al., 2019; Wei et al., 2020). For DIV2K-Val, 3,000 random 512×512 HR validation crops are degraded with StableSR's pipeline (Wang et al., 2024a). For RealSR and DRealSR, we follow StableSR, using center-cropped 128×128 LR patches for ×4 SR. We report PSNR, SSIM, MS-SSIM, LPIPS, and DISTS (Wang et al., 2003; 2004; Zhang et al., 2018; Ding et al., 2020). PSNR and SSIM use the Y channel in YCbCr space; higher is better for PSNR, SSIM, and MS-SSIM, whereas lower is better for LPIPS and DISTS.

Compared Methods. We organize the compared methods into three groups. The GAN-based group contains Real-ESRGAN (Wang et al., 2021) and FeMaSR (Chen et al., 2022). The one-step diffusion group contains OSEDiff (Wu et al., 2024a), PiSA-SR (Sun et al., 2025b), CODSR (Chen et al., 2026), and GDPO-SR (Yi et al., 2026). The multi-step diffusion group contains StableSR (Wang et al., 2024a), SeeSR (Wu et al., 2024b), DiffBIR (Lin et al., 2024), and the DiT-based DiT4SR (Duan et al., 2025). To test the plug-and-play ability of FGTC, we further apply it to StableSR and SeeSR and report both the original and enhanced results. We use official implementations and pretrained models when available, run each method with its recommended inference settings, and process all outputs using the same evaluation protocol.

Implementation Details. We integrate FGTC into frozen StableSR and SeeSR backbones, retaining their sampler update rules and noise schedules. LRS searches the first 30% of the reverse trajectory, i.e., $T_{\mathrm{LRS}} = \lfloor 0.3T \rfloor$. For each decoded clean prediction in this interval, it computes SSIM between its Haar LL component and that of the bicubically upsampled LR input, retaining the highest-scoring LL component as the structural reference. FDC applies a one-level 2D Haar DWT to decoded clean predictions and constructs each historical HF estimate from the latest $K = 5$ predictions before correction. The backbone retains its original conditioning. Unless stated otherwise, input preprocessing and sampling settings follow the original configuration of each backbone.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Quantitative Results. Table 1 compares FGTC with representative GAN-based, one-step, and multi-step diffusion Real-ISR methods on DRealSR, RealSR, and DIV2K-Val. FGTC consis-



Figure 4: Qualitative comparisons on Real-ISR. FGTC reduces spurious textures and improves structural fidelity over the original StableSR and SeeSR outputs.

tently improves PSNR, SSIM, and MS-SSIM for StableSR and SeeSR across all three benchmarks, with PSNR gains of 0.92–1.31 dB. An FGTC-enhanced backbone achieves the best PSNR, SSIM, LPIPS, and DISTS on each dataset and remains competitive in MS-SSIM. StableSR+FGTC performs strongly on DRealSR and RealSR, while SeeSR+FGTC leads PSNR on RealSR and DIV2K-Val. Perceptual metrics improve in most settings, so fidelity gains generally coincide with better agreement with GT rather than over-smoothing. These results demonstrate FGTC's plug-and-play applicability across backbones and synthetic and real-world degradations.

Qualitative Results. Figure 4 presents qualitative comparisons between FGTC and representative Real-ISR methods. Across the three examples, FGTC produces reconstructions that are more consistent with GT in both global structure and local details. In the first example, FGTC yields a cleaner map region with a more coherent boundary. In the second example, it preserves more regular character contours and clearer separation between adjacent strokes. In the third example, FGTC suppresses spurious surface textures while maintaining a continuous object boundary. These visual comparisons show that FGTC effectively reduces structural distortions and hallucinated textures while preserving useful image details. The improvements are consistent with the frequency-specific design of FGTC, where the selected early LL reference provides structural guidance and recent HF history supports detail refinement.

4.3 Ablation Study

We use SeeSR on DRealSR to examine component contributions, temporal-reference assignment, and round-trip operations under a unified evaluation setting. Base denotes the original SeeSR; Full

includes LRS and both FDC branches with the proposed correction strategy. Best results are bold, and the complete FGTC configuration is shaded light gray.

Table 2: Component ablation with SeeSR on DRealSR. Best results are bold.

Variant	LRS	FDC		PSNR↑	SSIM↑	MS-SSIM↑	LPIPS↓	DISTS↓
		LL	HF					
Base				26.75	0.7373	0.8625	0.3174	0.2315
S1	✓	✓		27.27	0.7584	0.8658	0.3057	0.2245
S2			✓	27.45	0.7733	0.8673	0.2940	0.2163
Full	✓	✓	✓	**27.82**	**0.7832**	**0.8738**	**0.2924**	**0.2073**

Table 3: Reference ablation with SeeSR on DRealSR. Hist. denotes recent history.

Variant	Reference		PSNR↑	SSIM↑	MS-SSIM↑	LPIPS↓	DISTS↓
	LL	HF					
V1	Early	Early	27.70	0.7801	0.8731	0.2964	0.2109
V2	Hist.	Early	27.53	0.7754	0.8706	**0.2901**	0.2094
V3	Hist.	Hist.	27.29	0.7678	0.8643	0.3016	0.2169
Full	Early	Hist.	**27.82**	**0.7832**	**0.8738**	0.2924	**0.2073**

Table 4: Ablation on round-trip operations with SeeSR on DRealSR. Decode–Encode and DWT–IDWT denote the two round-trip operations.

Variant	Decode–Encode	DWT–IDWT	FGTC correction	PSNR↑	SSIM↑	MS-SSIM↑	LPIPS↓	DISTS↓
SeeSR (no added operations)				26.7503	0.7373	0.862470	0.3174	0.2315
DE only	✓			26.9299	0.7474	0.857994	0.3037	0.2227
DID only		✓		26.7503	0.7373	0.862473	0.3174	0.2315
DE + DID	✓	✓		26.9291	0.7473	0.857954	0.3037	0.2227
FGTC (full)	✓	✓	✓	**27.8167**	**0.7832**	**0.873811**	**0.2924**	**0.2073**

Component contributions. To assess the contribution of each correction branch, we separately remove the low-frequency and high-frequency components of FGTC. As shown in Table 2, both LL correction with LRS and recent-history HF correction improve over the original SeeSR, indicating that structural and detail correction are individually effective. Combining the two branches further improves all metrics, demonstrating that they provide complementary benefits for reconstruction fidelity and perceptual quality.

Frequency-specific reference assignment. To verify whether different frequency bands benefit from different temporal references, we evaluate all combinations of early and recent-history references for LL and HF correction. As shown in Table 3, the proposed Early/Hist. assignment achieves the best PSNR, SSIM, MS-SSIM, and DISTS among the tested variants. Although the reversed Hist./Early setting obtains a slightly lower LPIPS, it performs worse on the fidelity-oriented metrics. These results support the frequency-dependent reference preference revealed by the offline analysis.

Round-trip operations. Since FGTC performs Decode–Encode and DWT–IDWT during correction, we further examine whether the improvements arise from these operations themselves. As shown in Table 4, DWT–IDWT alone produces almost no change, while Decode–Encode brings only limited gains and slightly reduces MS-SSIM. Applying both operations without correction yields nearly identical results to Decode–Encode alone. In contrast, enabling FGTC correction substantially improves all metrics, showing that the performance gains mainly come from the proposed correction rather than the round-trip transformations.

5 Conclusion

We introduced FGTC, a training-free, plug-and-play method for hallucination suppression in diffusion-based Real-ISR. Offline analysis shows that selected early predictions provide stronger low-frequency error cues, while recent history provides stronger high-frequency cues. FGTC therefore uses an early LL reference for structural correction and recent HF history for detail correction. Reference deviations adapt the strength of each frequency-specific correction, while the original backbone sampler remains unchanged. Component and reference-assignment ablations support the complementary contributions of the two branches. Across three benchmarks, FGTC improves StableSR and SeeSR reconstruction fidelity without GT at inference or backbone retraining. Qualitative examples show fewer structural distortions and spurious textures while preserving coherent boundaries and useful details across varied real-world scenes.

AI USE STATEMENT

Generative AI tools were used only for auxiliary tasks, including code debugging, literature retrieval, language polishing, figure and LaTeX preparation, and assistance in summarizing and interpreting experimental results. All core scientific activities, including problem formulation, method design, experimental planning, scientific interpretation of results, and manuscript conclusions, were conducted with direct and continuous involvement of the authors. The authors carefully reviewed and verified all AI-assisted content and take full responsibility for the final manuscript, results, claims, citations, and conclusions.

ETHICS STATEMENT

This work does not involve human-subject experiments or the collection of private or sensitive personal data. All experiments use publicly available benchmark datasets and pretrained models for real-world image super-resolution. The proposed method is intended for general image restoration research and does not introduce additional ethical concerns beyond those commonly associated with image generation and restoration technologies.

REPRODUCIBILITY STATEMENT

We provide the details needed to reproduce our experiments throughout the main paper and appendix. Section 4.1 describes the datasets, evaluation metrics, compared methods, and implementation settings. Section 3 presents the complete FGTC formulation, including reference selection, frequency-decoupled correction, and sampling updates. Appendix C provides the full inference algorithm and additional implementation details, while Appendix B reports the sensitivity analysis of the HF history length. All experiments are conducted using publicly available datasets and official pretrained models when available, under the same evaluation protocol described in the paper.

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.

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 (CVPR)*, pp. 38187–38197, June 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, 2025a.

Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 1329–1338, 2022.

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.

Junyang Chen, Jinshan Pan, and Jiangxin Dong. Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28188–28197. IEEE, 2025b.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18948–18958. 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.

Lei Jiang, Xin Liu, Xinze Tong, Zhiliang Li, Jie Liu, Jie Tang, and Gangshan Wu. Disentangled textual priors for diffusion-based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38228–38237, June 2026.

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 (CVPR)*, pp. 38270–38280, June 2026.

Yachao Li, Dong Liang, Tianyu Ding, and Sheng-Jun Huang. Structsr: refuse spurious details in real-world image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5022–5030, 2025.

Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5647–5656. IEEE, 2022.

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.

Yurui Qian, Qi Cai, Yingwei Pan, Yehao Li, Ting Yao, Qibin Sun, and Tao Mei. Boosting diffusion models with moving average sampling in frequency domain. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8911–8920. IEEE, 2024.

Lingchen Sun, Rongyuan Wu, Jie Liang, Zhengqiang Zhang, Hongwei Yong, and Lei Zhang. Improving the stability and efficiency of diffusion models for content consistent super-resolution. *IEEE Transactions on Image Processing*, 34:8421–8434, 2025a. doi: 10.1109/TIP.2025.3640863.

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, 2025b.

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.

Xingjian Wang, Li Chai, and Jiming Chen. Frequency-domain refinement with multiscale diffusion for super resolution. *arXiv preprint arXiv:2405.10014*, 2024b.

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.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

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.

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.

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.

Qiaosi Yi, Shuai Li, Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, and Lei Zhang. Gdpo-sr: Group direct preference optimization for one-step generative image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2177–2187, June 2026.

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.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23153–23163. IEEE, 2025.

Hexin Zhang, Dong Li, Jie Huang, Bingzhou Wang, Xueyang Fu, and Zhengjun Zha. Iterative inference-time scaling with adaptive frequency steering for image super-resolution. *arXiv preprint arXiv:2512.23532*, 2025.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4771–4780. IEEE, 2021.

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*, 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.

Qingping Zheng, Ling Zheng, Yuanfan Guo, Ying Li, Songcen Xu, Jiankang Deng, and Hang Xu. Self-adaptive reality-guided diffusion for artifact-free super-resolution. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25806–25816. IEEE, 2024.

APPENDIX FOR FGTC

This appendix provides the inference algorithm, additional ablations, and visual results for Frequency-Guided Temporal Correction (FGTC). It is organized as follows:

- A: supplementary statistics for temporal-reference analysis.
- B: sensitivity study of the HF history length.
- C: FGTC inference algorithm and implementation details.
- D: additional qualitative restoration examples.

A SUPPLEMENTARY TEMPORAL-REFERENCE STATISTICS

We examine whether a larger deviation from a temporal reference indicates a larger reconstruction error. For a single image, we first consider one frequency band and one reference, such as LL with the early reference. At each sampling step, we pair the current prediction's reference-deviation score with its GT-based error score in the same band. These pairs form two sequences over sampling steps. Spearman's rank correlation (ρ) compares the rankings of the two sequences: a higher positive value means that steps with larger reference deviations also tend to have larger GT-based errors.

We repeat this calculation for both references (early and recent history) in both bands (LL and HF), obtaining four correlation coefficients per image. Each Spearman entry in Table 5 reports the mean and standard deviation of the corresponding coefficient across images.

We also evaluate high-error detection using the area under the receiver operating characteristic curve (ROC-AUC). For each image and frequency band, we label sampling steps whose GT-based error scores exceed the image-specific 80th percentile as positive (high-error) steps and the remaining steps as negative. The reference-deviation scores serve as prediction scores, with larger deviations used to rank steps as more likely to have high errors. Both references use the same error-based labeling rule. ROC-AUC measures how well these scores rank high-error steps above other steps; a value of 0.5 corresponds to chance-level ranking and 1 to perfect separation. We compute ROC-AUC separately for each image, band, and reference, then report its mean and standard deviation across images.

Table 5 summarizes the per-image statistics. In LL, the early reference achieves higher mean Spearman correlation and mean high-error ROC-AUC than recent history. In HF, recent history achieves higher mean values on both metrics. The $\pm$ values denote standard deviations across images.

Table 5: Summary of per-image statistics for temporal-reference analysis. Entries are mean $\pm$ standard deviation across images. Higher is better for both metrics.

Metric	Early reference	Recent history
LL Spearman ρ	$\mathbf{0.796 \pm 0.260}$	0.206 ± 0.240
HF Spearman ρ	0.776 ± 0.291	$\mathbf{0.825 \pm 0.198}$
LL high-error AUC	$\mathbf{0.820 \pm 0.256}$	0.717 ± 0.223
HF high-error AUC	0.909 ± 0.179	$\mathbf{0.943 \pm 0.093}$

B SENSITIVITY TO HF HISTORY LENGTH

We evaluate the sensitivity to HF history length using SeeSR on DRealSR, varying K from 2 to 8 while keeping all other settings fixed. Here, K denotes the maximum number of preceding predictions used to construct the recent-history HF reference.

As shown in Table 6, all five metrics improve as K increases from 2 to 5. The default $K = 5$ achieves the highest PSNR, SSIM, and MS-SSIM. Increasing K to 6 reduces LPIPS and DISTS by only 0.0003 and 0.0005, respectively, while decreasing PSNR by 0.14 dB. Further increases to 7 and 8 worsen all five metrics relative to $K = 6$. We therefore retain $K = 5$ for its stronger reconstruction fidelity and competitive perceptual similarity.

Table 6: Effect of HF history length K with SeeSR on DRealSR. Best results are bold; the default $K = 5$ is shaded.

K	2	3	4	5 (Ours)	6	7	8
PSNR↑	27.37	27.49	27.57	**27.82**	27.68	27.56	27.46
SSIM↑	0.7767	0.7785	0.7811	**0.7832**	0.7825	0.7818	0.7810
MS-SSIM↑	0.8688	0.8701	0.8725	**0.8738**	0.8731	0.8719	0.8704
LPIPS↓	0.2966	0.2956	0.2946	0.2924	**0.2921**	0.2926	0.2937
DISTS↓	0.2154	0.2123	0.2099	0.2073	**0.2068**	0.2084	0.2101

C Algorithm and Implementation Details

Algorithm 1 summarizes early structural reference selection and subsequent frequency-decoupled correction. The HF history starts empty and is updated only during correction, using uncorrected predictions after their deviations have been evaluated.

Algorithm 1 Inference procedure for Frequency-Guided Temporal Correction (FGTC)

Require: I_{LR}; frozen encoder $\mathcal{E}$, decoder $\mathcal{D}$, sampler; T steps, T_{LRS} selection steps, K HF entries

Initialize Z_T, c; set $Q_H = \emptyset$, $x_{\mathrm{ref}}^{LL} = \emptyset$, $S_{\max} = -\infty$
$\mathcal{B} \leftarrow \{\mathrm{LH}, \mathrm{HL}, \mathrm{HH}\}$
$L_{\mathrm{LR}} \leftarrow \mathrm{DWT}_{LL}(\mathrm{Bicubic}(I_{\mathrm{LR}}))$
for $t = T, T-1, \ldots, 1$ **do**
  $Z_{0|t} \leftarrow \mathrm{PredictClean}(Z_t, t, c)$; $x_t \leftarrow \mathcal{D}(Z_{0|t})$
  $(x_t^{LL}, \{x_t^b\}_{b\in\mathcal{B}}) \leftarrow \mathrm{DWT}(x_t)$
  if $t > T - T_{\mathrm{LRS}}$ **then**
    $S_t^{LL} \leftarrow \mathrm{SSIM}(x_t^{LL}, L_{\mathrm{LR}})$
    if $S_t^{LL} \geq S_{\max}$ **then**
      $(S_{\max}, x_{\mathrm{ref}}^{LL}) \leftarrow (S_t^{LL}, x_t^{LL})$
    end if
    $Z_{t-1} \leftarrow \mathrm{Sampler}(Z_t, Z_{0|t}, t, c)$
  else
    $R_t^b \leftarrow x_t^b$ if $Q_H = \emptyset$, otherwise $R_t^b \leftarrow \mathrm{Mean}(Q_H^b)$
    Compute w_t^{LL}, w_t^{HF} via Eq. 4
    Correct LL and HF via Eq. 5
    Reconstruct x_t' via Eq. 6; blend x_t^C via Eq. 7
    $Z_{0|t}^C \leftarrow \mathrm{Mean}(\mathcal{E}(x_t^C))$
    $Q_H \leftarrow \mathrm{FIFO}(Q_H, \{x_t^b\}_{b\in\mathcal{B}}, K)$
    $Z_{t-1} \leftarrow \mathrm{Sampler}(Z_t, Z_{0|t}^C, t, c)$
  end if
end for
return $I_{\mathrm{HR}} = \mathcal{D}(Z_0)$

D Additional Qualitative Restoration Examples

Figures 5–7 present eight additional qualitative comparisons across different scenes. Each example includes the LR input, GT, competing Real-ISR methods, and the original and FGTC-enhanced StableSR and SeeSR outputs. Red boxes mark regions enlarged for closer comparison of structure and texture. These examples complement the quantitative results in the main paper.

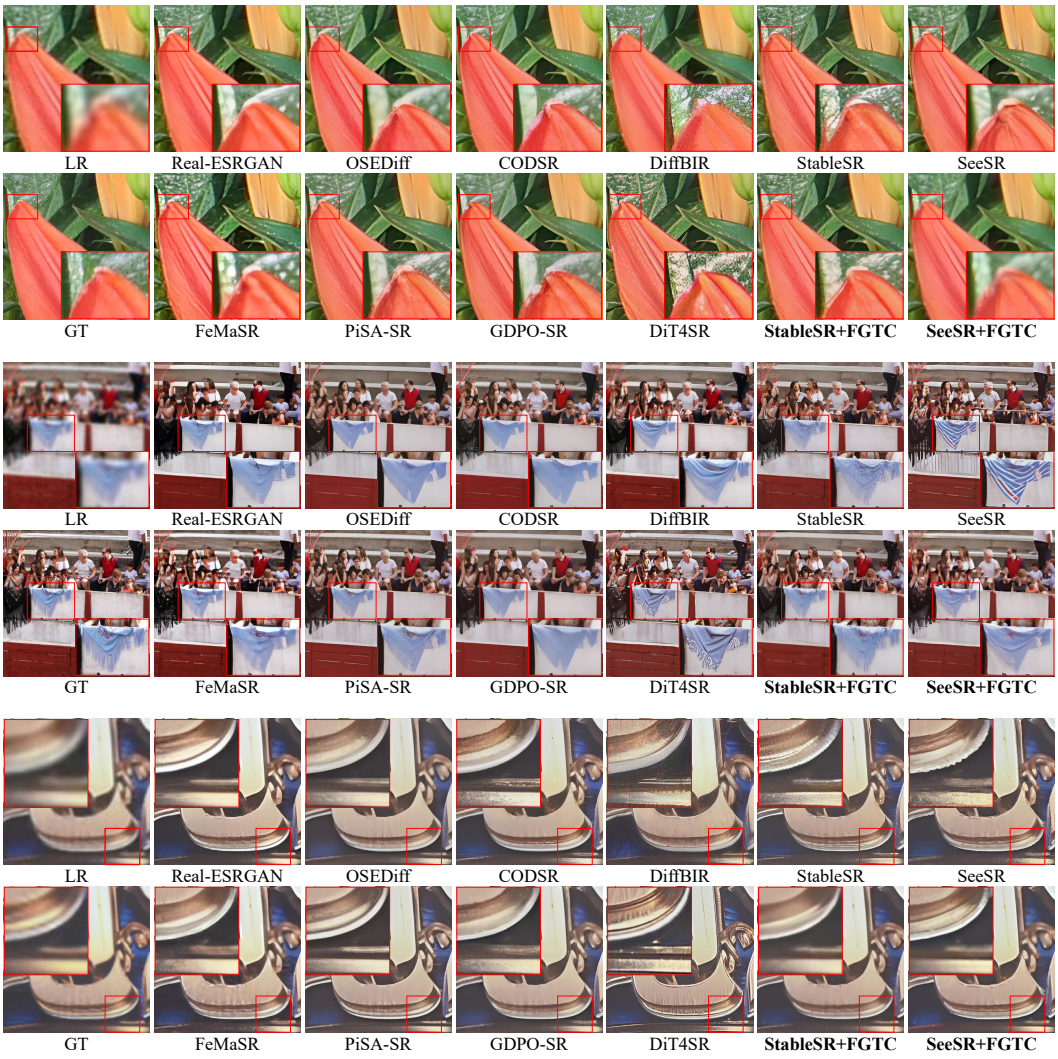


Figure 5: Additional qualitative comparisons, examples 1–3 from top to bottom.

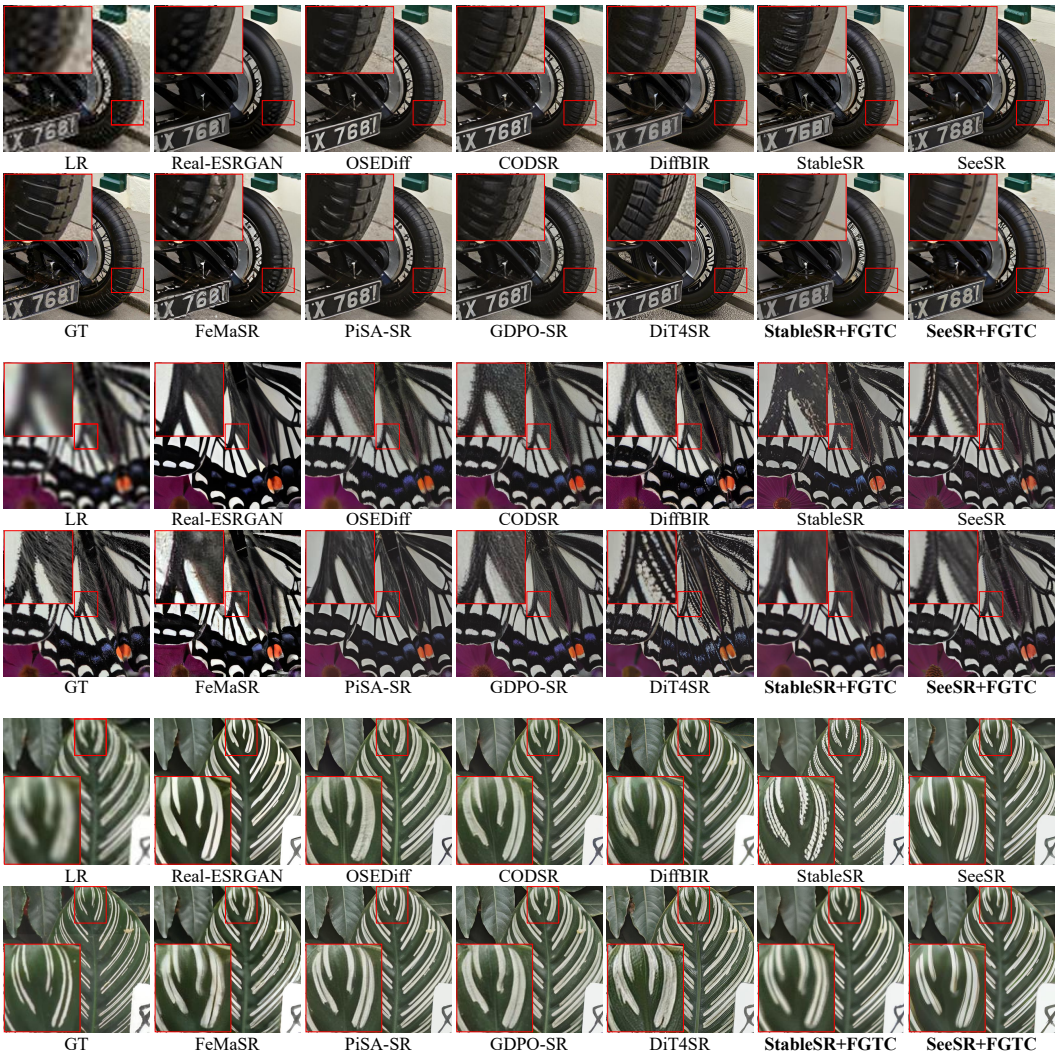


Figure 6: Additional qualitative comparisons, examples 4–6 from top to bottom.

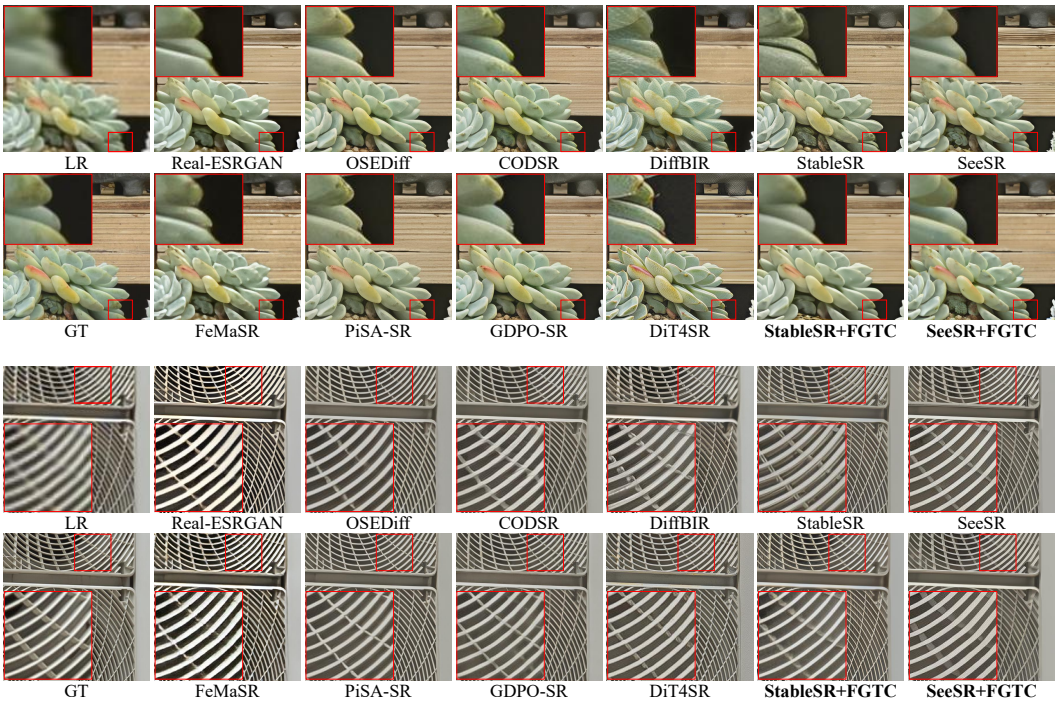


Figure 7: Additional qualitative comparisons, examples 7–8 from top to bottom.