SPECTRAL PROFILING FOR COMPLEMENTARY DUAL-BRANCH CAPACITY ALLOCATION IN ONE-STEP DIFFUSION SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

One-step diffusion super-resolution enables efficient generative restoration by replacing iterative sampling with a single denoising pass. A fixed U-Net exhibits heterogeneous spectral responses across layers, yet existing dual-LoRA methods typically retain uniform ranks across adapted layers. We propose Frequency-Response-Guided Diffusion Super-Resolution (FreqDiSR), which profiles a frozen U-Net by measuring the Fourier high-frequency response of each layer and converts the resulting profile into a static adaptation capacity map. Under a fixed total rank budget for the two branches at each profiled module, decreasing responses increase the capacity of the pixel branch, whereas increasing responses increase the capacity of the semantic branch, coupling the measured spectral ordering with their respective fidelity and perception objectives. Haar-based decoder modulation and low-frequency supervision extend the same separation of structure and detail to feature processing and training objectives without introducing additional denoising steps. Allocation controls with matched budgets favor the measured correspondence over inverse mappings, depth-based mappings, and permutations on perceptual metrics, while subset and degradation analyses show that the spectral profile is stable. On DRealSR, FreqDiSR improves PSNR by $0.62\,\mathrm{dB}$ and reduces LPIPS by 0.0140 over PiSA-SR. Across three benchmarks, it achieves the lowest LPIPS and the lowest DISTS on RealSR and DRealSR.

1 INTRODUCTION

Diffusion-based approaches have advanced single-image SR, especially for real-world degradations that are difficult to model explicitly Saharia et al. (2023); Wang et al. (2024a); Lin et al. (2024); Wu et al. (2024b). Priors learned from large-scale image distributions enable the recovery of plausible textures that are often suppressed by regression-based models Ledig et al. (2017); Sajjadi et al. (2017); Wang et al. (2024a). This capability, however, exposes a persistent *fidelity–perception trade-off*: perceptual synthesis favors rich high-frequency details, whereas faithful reconstruction requires consistency with LR-supported structure, geometry, and color Blau & Michaeli (2018). The DRealSR example in Figure 1 illustrates this tension. This tension becomes particularly acute in one-step diffusion SR, where LR-supported structure and plausible missing details must be recovered within a single U-Net evaluation. Unlike iterative diffusion models, a one-step model has no subsequent denoising states to progressively refine structural deviations or implausible textures Wang et al. (2024b); Wu et al. (2024a); Sun et al. (2025).

Existing approaches address this challenge through controllable feature guidance and pixel-aware conditioning Wang et al. (2024a); Yang et al. (2024), semantic guidance Wu et al. (2024b), shortened diffusion trajectories Yue et al. (2023), and distillation into accelerated or single-step inference Wang et al. (2024b); Wu et al. (2024a); Tai et al. (2026). These methods improve input conditioning, feature control, or sampling efficiency, but do not explicitly determine how to distribute limited adaptation capacity across one-step U-Net layers. Representative dual-LoRA SR designs separate pixel fidelity and perceptual enhancement through objective-specific adapters while retaining uniform ranks across adapted U-Net layers Sun et al. (2025). This separation determines what each branch learns, whereas the layer-wise placement of their finite capacity remains uniform. The result-

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.

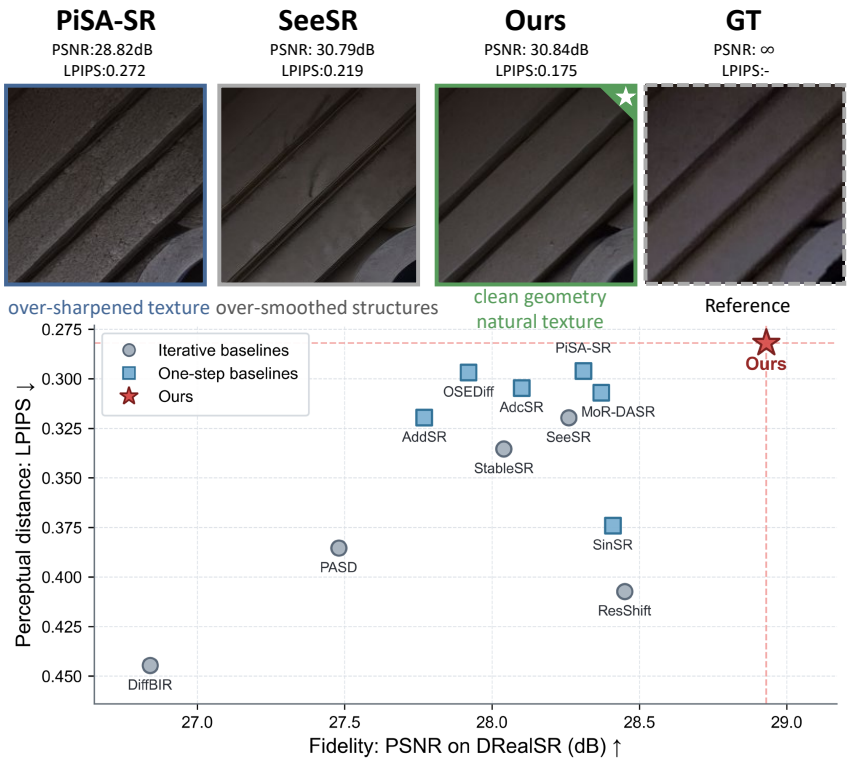

Figure 1: Representative fidelity–perception comparison on DRealSR. In this example, FreqDiSR achieves higher PSNR and lower LPIPS than the displayed diffusion-based baselines while preserving the dominant geometric pattern.

ing question is how to distribute a fixed two-branch rank budget across layers with heterogeneous measured responses.

Low-frequency content mainly carries LR-constrained geometry and color, whereas high-frequency textures admit multiple plausible reconstructions. Since U-Net layers exhibit heterogeneous spectral behavior and adapter placement affects adaptation Si et al. (2024); Xiang et al. (2023), layer-wise high-frequency response provides an observable signal for organizing adaptation capacity. We convert this signal into a complementary allocation rule: decreasing responses increase pixel-branch rank, whereas increasing responses increase semantic-branch rank. The rule couples the measured spectral ordering with the existing fidelity- and perception-oriented branch objectives rather than a predefined depth schedule.

We propose a dual-LoRA framework for one-step SR, termed Frequency-Response-Guided Diffusion Super-Resolution (**FreqDiSR**). FreqDiSR measures layer-wise Fourier high-frequency responses and converts them into complementary pixel- and semantic-branch ranks. The resulting map redistributes a fixed per-module budget and remains fixed throughout adaptation. Profiling is performed once on the frozen backbone, preserving the same single-pass inference path. Fourier responses determine where adaptation capacity is placed; Haar decomposition determines which frequency components are refined or constrained. We modulate detail subbands in selected late-decoder features and apply a training-only low-frequency anchor to preserve LR-supported structure.

The measured profiles show pronounced heterogeneity within and across U-Net blocks and remain stable across disjoint image subsets and degraded inputs. Among budget-matched nonuniform controls, the measured allocation improves LPIPS and DISTS over inverse, depth-based, and permuted mappings. It also outperforms the all-module rank-4 baseline on both metrics. Across three benchmarks, FreqDiSR achieves the lowest LPIPS and the lowest DISTS on both real-world benchmarks while retaining one-step inference. On DRealSR, it improves PSNR by 0.62 dB and reduces LPIPS by 0.0140 relative to PiSA-SR.

The main contributions of this work are summarized as follows:

- To allocate a fixed dual-LoRA budget across spectrally heterogeneous U-Net modules, we introduce response-guided allocation that converts frozen-backbone Fourier profiles into complementary pixel- and semantic-branch ranks.
- To balance LR-supported structure and perceptual detail, we introduce Haar-based decoder modulation and training-only low-frequency supervision while preserving one-step inference.
- Budget-matched allocation controls favor the measured correspondence over inverse, depth-based, and permuted alternatives, while profile analyses show stable extraction across image subsets and degradations. FreqDiSR achieves the lowest LPIPS on all three benchmarks and the lowest DISTS on both real-world benchmarks.

2 Related Work

2.1 Diffusion-based Image Super-Resolution

Diffusion priors shift super-resolution from deterministic reconstruction toward sampling realistic textures conditioned on degraded observations Saharia et al. (2023); Wang et al. (2024a). This formulation requires generated detail to remain anchored to input-supported structure. Feature guidance and pixel-aware conditioning reinforce that connection Wang et al. (2024a); Yang et al. (2024), while two-stage restoration–generation pipelines Lin et al. (2024) and semantic guidance Wu et al. (2024b) separate reconstruction from detail synthesis. Joint adaptation and diffusion-pretrained initialization further strengthen the restoration prior Chen et al. (2025b); Lin et al. (2025).

Sampling cost becomes central when this prior is deployed for SR. ResShift shortens the diffusion trajectory Yue et al. (2023); SinSR Wang et al. (2024b), OSEDiff Wu et al. (2024a), and PiSA-SR Sun et al. (2025) compress restoration into one step. Target-score distillation Dong et al. (2025) and adversarial compression Chen et al. (2025a) transfer iterative generation into a learned single-pass mapping. Single-pass SR places structure recovery and detail synthesis in the same computation and motivates explicit organization of the adaptation capacity within that pass.

2.2 Parameter-Efficient Adaptation for Diffusion Models

Parameter-efficient adaptation expresses this problem as a finite trainable budget. Low-rank adaptation (LoRA) freezes pretrained weights and inserts compact updates into selected layers Hu et al. (2022), while diffusion-specific approaches tune scaling factors or singular values Xie et al. (2023); Han et al. (2023). Dual-LoRA SR separates fidelity and perceptual enhancement into objective-specific branches while retaining uniform layer ranks Sun et al. (2025). AdaLoRA reallocates rank during optimization according to parameter importance Zhang et al. (2023); AIRA uses activation outliers for offline layer-wise rank assignment Li et al. (2025); and SR-LoRA uses the stable rank of pretrained weights as a static allocation prior Zhang et al. (2025). MoR-DASR instead routes rank-level experts from input degradation cues He et al. (2026). These methods vary the total capacity of matrices, layers, or routed experts. FreqDiSR preserves a fixed local budget and uses a dataset-level Fourier profile of the frozen SR backbone to change its composition between two objective-specific branches. The resulting static map requires no inference-time rank prediction or expert routing.

2.3 Frequency Modeling in Image Super-Resolution

Frequency modeling provides signal decompositions that distinguish coarse structure from orientation-sensitive detail. Wavelet analysis separates low-frequency structure and color from orientation-sensitive high-frequency edges and textures Mallat (1989). Wavelet-based CNNs and SR models exploit this multiscale separation for reconstruction Liu et al. (2018); Huang et al. (2017), while wavelet-domain diffusion carries it into generative restoration Huang et al. (2024). Prior methods use spectral decomposition for features or reconstruction objectives. FreqDiSR instead uses layer-wise Fourier responses to allocate a fixed LoRA rank budget across the U-Net, while Haar supports decoder detail modulation and low-frequency supervision. Fourier determines where capacity is placed; Haar determines which components are refined or constrained.

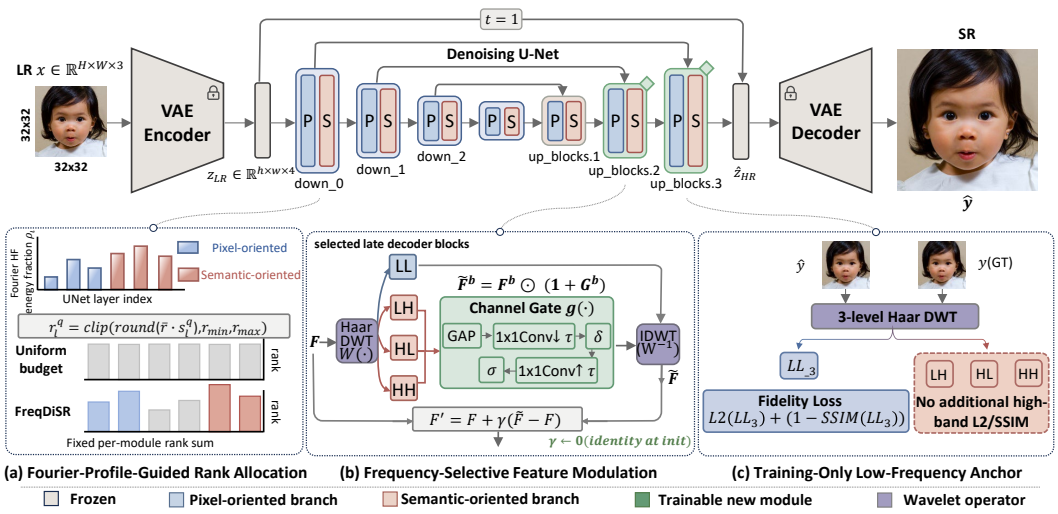

Figure 2: Overview of FreqDiSR. Measured layer-wise frequency responses guide complementary ranks for the pixel and semantic LoRA branches. Decoder feature modulation refines high-frequency subbands, while a training-only level-3 low-frequency anchor supplements the inherited full-image objectives.

3 METHOD

3.1 OVERVIEW

Given a low-resolution image x, FreqDiSR performs one-step restoration with a pretrained latent-diffusion dual-LoRA SR backbone. The frozen VAE encoder encodes x into a latent representation z_{LR}. A precomputed Fourier energy profile assigns complementary ranks to the pixel- and semantic-oriented LoRA branches of the U-Net. At the fixed timestep $t = 1$, the rank-adapted U-Net predicts a residual $R(z_{\mathrm{LR}}, c)$, where c denotes the text condition. Frequency-selective Haar modulation refines the high-frequency subbands of selected late-decoder features during the same forward pass. The resulting high-resolution latent is $\hat{z}_{\mathrm{HR}} = z_{\mathrm{LR}} - R(z_{\mathrm{LR}}, c)$, which the frozen VAE decoder decodes into the SR output $\hat{y}$.

During training, the level-3 low-frequency component of $\hat{y}$ is anchored to that of the target alongside the inherited full-image objectives. At inference, the allocated branch ranks and decoder modulation remain active in the same one-step path, while the low-frequency structural anchor is removed.

3.2 FREQUENCY DECOMPOSITION AND ALLOCATION PRIOR

For a target y and prediction $\hat{y}$, Haar decomposition provides a spatially local multiresolution separation into approximation and detail components Mallat (1989). For any image u, we write

$$\begin{aligned} \mathcal{W}(u) &= \{u^{\mathrm{LL}}, u^{\mathrm{LH}}, u^{\mathrm{HL}}, u^{\mathrm{HH}}\}, \\ \mathcal{H}(u) &= \{u^{\mathrm{LH}}, u^{\mathrm{HL}}, u^{\mathrm{HH}}\}, \end{aligned} \tag{1}$$

where $\mathcal{W}$ is the single-level Haar transform, $\mathcal{H}$ collects its high-frequency components, u^{LL} is the low-frequency approximation, and u^{LH}, u^{HL}, and u^{HH} are orientation-specific detail subbands.

We use Haar because its separable two-tap filters form an orthogonal perfect-reconstruction transform with minimal spatial support and low overhead Mallat (1989). The LL band represents LR-constrained structure and color, whereas the LH/HL/HH bands represent orientation-sensitive edges and textures with multiple plausible realizations Liu et al. (2018); Huang et al. (2017); Blau & Michaeli (2018). Applying this basis once for feature modulation and recursively to level 3 for supervision maintains a common structure–detail definition across feature and objective spaces. Fourier response separately governs cross-layer capacity allocation.

Layer-wise response provides an observable ordering for complementary rank assignment: decreasing response increases pixel-branch rank, whereas increasing response increases semantic-branch rank. Because the two branches are optimized for fidelity and perceptual enhancement, respectively, the rule couples the measured ordering with their existing objective specialization.

3.3 Frequency-Response-Guided Dual-LoRA Allocation

Following LoRA Hu et al. (2022), branch $q \in \{\mathrm{pix}, \mathrm{sem}\}$ updates frozen weight W_l as $W_l'^{,q} = W_l + \Delta W_l^q$, where $\Delta W_l^q = \frac{\alpha_l}{r_l^q} B_l^q A_l^q$. The trainable down- and up-projections A_l^q and B_l^q have sizes $r_l^q \times d_{\mathrm{in}}$ and $d_{\mathrm{out}} \times r_l^q$; α_l is the scaling coefficient, and d_{in} and d_{out} are the input and output dimensions.

Both branches adapt the same set $\mathcal{A}$ of 258 convolutional and linear U-Net modules. Its subset $\mathcal{P} \subset \mathcal{A}$ contains 226 modules whose outputs form square spatial grids. For $l \in \mathcal{P}$, let F_l have shape $N \times C \times h \times w$, where N, C, h, and w denote batch size, channels, height, and width. We measure its radial Fourier high-frequency energy fraction

$$\begin{aligned} E_{lnc}(\Omega') &= \textstyle\sum_{\omega \in \Omega'} \left|\mathcal{F}_{2\mathrm{D}}(F_{lnc})(\omega)\right|^2, \\ \rho_l &= \frac{1}{NC} \textstyle\sum_{n=1}^{N} \sum_{c=1}^{C} \frac{E_{lnc}(\Omega_{\mathrm{HF}})}{E_{lnc}(\Omega)+\epsilon}. \end{aligned} \tag{2}$$

Here F_{lnc} is the spatial map of sample n and channel c; $\mathcal{F}_{2\mathrm{D}}$ is the two-dimensional Fourier transform, ω is a frequency coordinate, Ω is the full grid, and $\Omega' \subseteq \Omega$. Ω_{HF} contains frequencies beyond half the Nyquist radius, and ϵ stabilizes the denominator. Larger ρ_l indicates a greater high-frequency fraction. Orthonormal scaling and a centered zero frequency are used before masking; per-sample, per-channel fractions remove sensitivity to scalar activation magnitude.

The frozen VAE encodes center-cropped training targets into clean sampled latents without added diffusion noise. At timestep 1, null-text conditioning isolates image-driven U-Net responses collected by forward hooks before LoRA adaptation. Convolutional maps and square token grids enter $\mathcal{P}$; non-square sequences are excluded from the spatial Fourier estimator. Averaging these responses yields a reusable spectral allocation profile that configures both adaptation branches.

The profile $\{\rho_l : l \in \mathcal{P}\}$ is averaged over 200 training images and min–max normalized to $\bar{\rho}_l \in [0, 1]$. Centering it at the midpoint, $\delta_l = \bar{\rho}_l - \frac{1}{2}$, gives complementary multipliers $s_l^{\mathrm{sem}} = 1 + \delta_l$ and $s_l^{\mathrm{pix}} = 1 - \delta_l$. The rank is

$$r_l^q = \mathrm{clip}\left(\mathrm{round}(\bar{r} s_l^q),\ \frac{\bar{r}}{2},\ \frac{3\bar{r}}{2}\right), \tag{3}$$

where $q \in \{\mathrm{pix}, \mathrm{sem}\}$ and $\bar{r}$ is the inherited uniform backbone rank. The midpoint retains uniform allocation; positive and negative deviations allocate more rank to the semantic and pixel branches, respectively. Since $\delta_l \in [-\frac{1}{2}, \frac{1}{2}]$, the multipliers lie in $[0.5, 1.5]$ and sum to 2. With $\bar{r} = 4$, ranks lie in $[2, 6]$ and every profiled module preserves $r_l^{\mathrm{pix}} + r_l^{\mathrm{sem}} = 8$. The resulting rank map is fixed after profiling and reused throughout training and inference without routing overhead. The remaining 32 non-spatial text-side cross-attention key/value projections retain shared block-wise ranks in both branches; the complete module schema is reported in the supplement.

3.4 Frequency-Selective Decoder Feature Modulation

Using the Haar basis defined above, we decompose a selected late-decoder feature F into F^{LL} for coarse structure and $F^{\mathrm{LH}}, F^{\mathrm{HL}}, F^{\mathrm{HH}}$ for orientation-sensitive detail. We concatenate the three high-frequency subbands and compute a joint channel gate. Late-decoder placement targets detail refinement at high spatial resolution:

$$\begin{aligned} F^{\mathrm{HF}} &= [F^{\mathrm{LH}}; F^{\mathrm{HL}}; F^{\mathrm{HH}}], \quad G = g(F^{\mathrm{HF}}), \\ \tilde{F}^b &= F^b \odot (1 + G^b), \quad b \in \{\mathrm{LH}, \mathrm{HL}, \mathrm{HH}\}, \end{aligned} \tag{4}$$

Bracketed semicolons denote channel concatenation, g is a squeeze-and-excitation (SE) gate, G is split into $\{G^{\mathrm{LH}}, G^{\mathrm{HL}}, G^{\mathrm{HH}}\}$, and $\odot$ denotes element-wise multiplication. The gate uses global average pooling and two 1×1 convolutions with ReLU and sigmoid activations Hu et al. (2018). Here F^{HF} is the concatenated detail tensor and $\tilde{F}^b$ is the modulated subband; G^b is the gate slice for

Dataset	Method	Time (s/img)↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
DIV2K-Val	ResShift-S15	6.1152	24.69	**0.6175**	0.3374	0.2215	36.01	6.82	0.6089	60.92	0.5450
	StableSR-S200	14.0200	23.31	0.5728	0.3129	0.2138	**24.67**	4.76	0.6682	65.63	0.6188
	DiffBIR-S50	7.1054	23.67	0.5653	0.3541	0.2129	30.93	4.71	0.6652	65.66	0.6204
	PASD-S20	1.4597	23.14	0.5489	0.3607	0.2219	29.32	4.40	0.6711	68.83	**0.6484**
	SeeSR-S50	4.8887	23.71	0.6045	0.3207	0.1967	25.83	4.82	0.6857	68.49	0.6239
	SinSR-S1	0.6441	24.43	0.6012	0.3262	0.2066	35.45	6.02	0.6499	62.80	0.5395
	OSEDiff-S1	0.1597	23.72	0.6108	0.2941	0.1976	26.32	4.71	0.6683	67.97	0.6148
	PiSA-SR-S1	**0.1351**	23.87	0.6058	0.2823	0.1934	25.07	4.55	0.6927	**69.68**	0.6400
	MoR-DASR-S1†	–	24.01	0.606	0.289	–	–	–	0.681	68.09	0.475
	AdcSR-S1†	–	23.74	0.6017	0.2853	**0.1899**	25.52	**4.36**	0.6764	68.00	0.6090
	AddSR-S1	–	23.26	0.5902	0.3623	0.2123	29.68	4.76	0.5657	63.39	0.5734
	Ours	0.1355	**24.70**	0.6147	**0.2783**	0.1929	26.62	4.54	**0.6938**	68.34	0.6402
RealSR	ResShift-S15	6.1152	**26.31**	0.7411	0.3489	0.2498	142.81	7.27	0.5450	58.10	0.5305
	StableSR-S200	14.0200	24.69	0.7052	0.3091	0.2167	127.20	5.76	0.6195	65.42	0.6211
	DiffBIR-S50	7.1054	24.88	0.6673	0.3567	0.2290	124.56	5.63	0.6412	64.66	0.6231
	PASD-S20	1.4597	25.22	0.6809	0.3392	0.2259	123.08	**5.18**	0.6502	68.74	0.6461
	SeeSR-S50	4.8887	25.33	0.7273	0.2985	0.2213	125.66	5.38	0.6594	69.37	0.6439
	SinSR-S1	0.6441	26.30	0.7354	0.3212	0.2346	137.05	6.31	0.6204	60.41	0.5389
	OSEDiff-S1	0.1597	25.15	0.7341	0.2921	0.2128	123.50	5.65	0.6693	69.09	0.6339
	PiSA-SR-S1	**0.1351**	25.50	**0.7417**	0.2672	0.2044	124.09	5.50	0.6702	**70.15**	0.6560
	MoR-DASR-S1†	–	25.32	0.728	0.291	–	–	–	**0.691**	69.78	0.512
	AdcSR-S1†	–	25.47	0.7301	0.2885	0.2129	**118.41**	5.35	0.6731	69.90	0.6360
	AddSR-S1	–	24.79	0.7077	0.3091	0.2191	132.05	5.54	0.6098	66.18	0.5722
	Ours	0.1355	**26.31**	0.7356	**0.2575**	**0.1947**	118.82	5.59	0.6731	67.44	**0.6561**
DRealSR	ResShift-S15	6.1152	28.45	0.7632	0.4073	0.2700	175.92	8.28	0.5259	49.86	0.4573
	StableSR-S200	14.0200	28.04	0.7460	0.3354	0.2287	147.03	6.51	0.6171	58.50	0.5602
	DiffBIR-S50	7.1054	26.84	0.6660	0.4446	0.2706	167.38	6.02	0.6292	60.68	0.5902
	PASD-S20	1.4597	27.48	0.7051	0.3854	0.2535	157.36	**5.57**	0.6714	64.55	0.6130
	SeeSR-S50	4.8887	28.26	0.7698	0.3197	0.2306	149.86	6.52	0.6672	64.84	0.6026
	SinSR-S1	0.6441	28.41	0.7495	0.3741	0.2488	177.05	7.02	0.6367	55.34	0.4898
	OSEDiff-S1	0.1597	27.92	**0.7835**	0.2968	0.2165	135.29	6.49	0.6963	64.65	0.5899
	PiSA-SR-S1	**0.1351**	28.31	0.7804	0.2960	0.2169	**130.61**	6.20	0.6970	66.11	0.6156
	MoR-DASR-S1†	–	28.37	0.776	0.307	–	–	–	0.717	65.94	0.509
	AdcSR-S1†	–	28.10	0.7726	0.3046	0.2200	134.05	6.45	0.7049	**66.26**	0.5927
	AddSR-S1	–	27.77	0.7722	0.3196	0.2242	150.18	6.93	0.549	60.85	0.6188
	Ours	0.1355	**28.93**	0.7736	**0.2820**	**0.2085**	131.64	6.12	**0.7172**	64.79	**0.6202**

Table 1: Quantitative comparison on three benchmarks. Unified baselines follow PiSA-SR Sun et al. (2025); daggered MoR-DASR/AdcSR values are author-reported He et al. (2026); Chen et al. (2025a), and FreqDiSR reports three-seed means. Bold/underline denote best/second; S1 denotes one-step inference and time is in s/image.

orientation b. Joint gating exploits correlations among horizontal, vertical, and diagonal responses instead of treating each orientation independently. The unchanged LL path preserves low-frequency structure, while the residual high-frequency path controls how strongly each decoder block refines local detail. The multiplicative form $1 + G^b$ retains the incoming response and learns an orientation-aware correction.

The inverse DWT $\mathcal{W}^{-1}$ reconstructs $\tilde{F} = \mathcal{W}^{-1}(\{F^{\mathrm{LL}}, \tilde{F}^{\mathrm{LH}}, \tilde{F}^{\mathrm{HL}}, \tilde{F}^{\mathrm{HH}}\})$, followed by the residual path

$$F' = F + \gamma(\tilde{F} - F), \tag{5}$$

where F' is the modulated output and γ is a learnable scalar initialized to zero. Zero initialization preserves the pretrained mapping at the start of adaptation and lets each block learn its frequency-selective modulation strength.

3.5 Low-Frequency Structural Anchoring

Using the same Haar basis as decoder modulation, we add an image-space structural anchor on the level-3 low-frequency component alongside the inherited full-image objectives. Pixel-space ℓ_2 encourages reconstruction fidelity, while SSIM encourages structural consistency Wang et al. (2004); the auxiliary constraint therefore concentrates on coarse structure while perceptual objectives model detail Ledig et al. (2017); Blau & Michaeli (2018). For restoration $\hat{y}$ and target y, a recursive three-level Haar decomposition produces $\mathrm{LL}_3(\hat{y})$ and $\mathrm{LL}_3(y)$. The additional image-space term anchors these low-frequency components:

$$\begin{aligned}\mathcal{L}_{\mathrm{low}} = {} & w_{\ell_2} \|\mathrm{LL}_3(\hat{y}) - \mathrm{LL}_3(y)\|_2^2 \\ & + w_{\mathrm{SSIM}} \left(1 - \mathrm{SSIM}\left(\mathrm{LL}_3(\hat{y}), \mathrm{LL}_3(y)\right)\right),\end{aligned} \tag{6}$$

where $\mathrm{LL}_3(\cdot)$ is the three-level Haar low-frequency subband, and w_{ℓ_2} and w_{SSIM} weight its ℓ_2 and SSIM components. These terms emphasize coarse geometry and low-frequency color, while

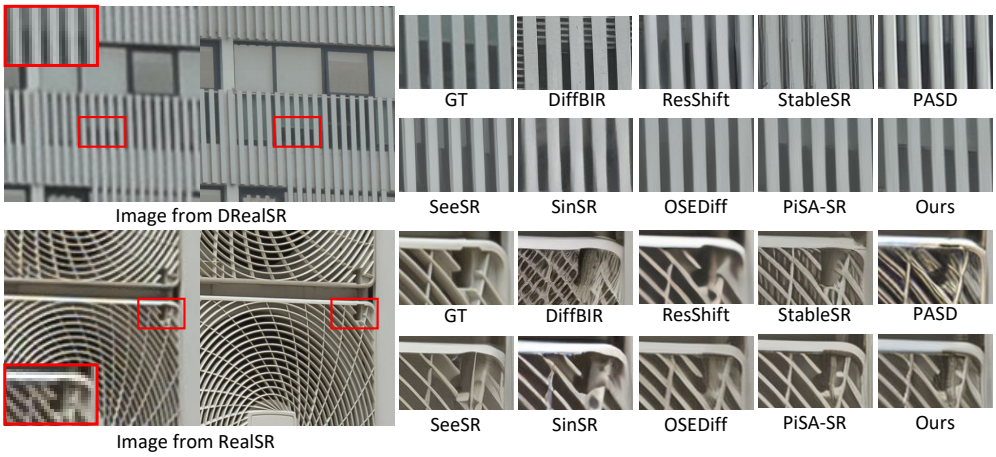

Figure 3: Qualitative comparison on real-world paired benchmarks. Enlarged regions show FreqDiSR's recovery of continuous lines, coherent geometry, and natural local textures.

the inherited image-fidelity, perceptual, and CSD objectives optimize the complete output and its details. This division places deterministic guidance on the component most directly supported by the LR observation and complements branch allocation without changing the one-step inference graph.

Overall objective. FreqDiSR follows the pixel-to-semantic training protocol of the dual-LoRA backbone. The pixel stage learns the fidelity-oriented branch; the semantic stage activates the perceptual branch together with decoder modulation and low-frequency anchoring. Its objective is

$$\begin{aligned}\mathcal{L}_{\text{total}} = \lambda_{\ell_2}\mathcal{L}_{\ell_2} + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}} \\ +\lambda_{\text{CSD}}\mathcal{L}_{\text{CSD}} + \lambda_{\text{low}}\mathcal{L}_{\text{low}}.\end{aligned} \tag{7}$$

λ_{ℓ_2}, λ_{LPIPS}, λ_{CSD}, and λ_{low} weight the four losses; w_{ℓ_2} and w_{SSIM} act within $\mathcal{L}_{\text{low}}$. The first three objectives are inherited from the dual-LoRA backbone and retain their image-fidelity, perceptual, and semantic roles. $\mathcal{L}_{\text{low}}$ adds the low-frequency image-space anchor only during training, leaving the two adapted branches and decoder modulation in one denoising pass.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. FreqDiSR is built on Stable Diffusion 2.1-base Rombach et al. (2022) for $\times 4$ SR. Training combines LSDIR Li et al. (2023) and the first 10,000 FFHQ images Karras et al. (2019) with Real-ESRGAN degradations Wang et al. (2021). Evaluation covers a synthetic 3,000-patch DIV2K-Val split Agustsson & Timofte (2017) and the paired real-camera benchmarks RealSR Cai et al. (2019) and DRealSR Wei et al. (2020), containing 100 and 93 image pairs. DIV2K-Val is generated by the training degradation pipeline, whereas RealSR and DRealSR provide paired observations from real camera systems. Together, these benchmarks cover controlled and real acquisition degradations. PSNR and SSIM Wang et al. (2004) measure fidelity on the luminance channel; LPIPS Zhang et al. (2018) and DISTS Ding et al. (2022) compare restored and reference RGB images in learned perceptual spaces; FID Heusel et al. (2017) measures set-level distributional quality; and NIQE Mittal et al. (2013), CLIPIQA Wang et al. (2023), MUSIQ Ke et al. (2021), and MANIQA Yang et al. (2022) assess no-reference quality. Higher is better for PSNR, SSIM, CLIPIQA, MUSIQ, and MANIQA; lower is better for LPIPS, DISTS, FID, and NIQE. The complete metric protocol is provided in the supplement.

Implementation. Following the one-step dual-LoRA SR protocol Sun et al. (2025), training uses a learning rate of 5×10^{-5}, a batch size of 4, and 10,500 optimization steps; the first 4,000 steps form the pixel-oriented stage.

Variant	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	CLIPIQA↑
Component contribution					
FreqDiSR	28.98	0.7742	0.2845	0.2082	0.7180
w/o decoder mod.	28.68	0.7704	0.2871	0.2079	0.6986
All-module rank-4	28.95	0.7605	0.2885	0.2096	0.6855
No add. guidance	28.43	0.7745	0.2887	0.2120	0.7066
Supervision domain					
Full-band guidance	28.80	0.7787	0.2916	0.2133	0.6883
Allocation correspondence					
Inverse mapping	28.94	0.7715	0.2904	0.2157	0.6831
Depth-based	29.07	0.7772	0.2881	0.2128	0.6974
Permuted (mean)	28.98	0.7752	0.2918	0.2129	0.6921

Table 2: Component, supervision-domain, and allocation analyses on DRealSR. The permuted row averages three profiles; full results are in the supplement.

The dual-LoRA SR backbone uses base rank 4 for both pixel and semantic branches Sun et al. (2025); the adapters use the LoRA parameterization Hu et al. (2022). Frequency-aware rank allocation uses a data-driven Fourier response profile for 226 of the 258 shared adapted modules; its midpoint-centered complementary mapping yields multipliers in $[0.5, 1.5]$ and ranks in $[2, 6]$ with a per-module two-branch rank sum of 8. Frequency-selective decoder feature modulation operates in the penultimate and final decoder blocks. The semantic stage adds the low-frequency anchor from a three-level Haar decomposition Mallat (1989), using ℓ_2/SSIM weights $w_{\ell_2} = 0.5$ and $w_{\mathrm{SSIM}} = 0.2$ Wang et al. (2004), with $\lambda_{\mathrm{low}} = 0.5$ in the overall objective. The anchor is training-only; inference remains one-step.

4.2 Main Quantitative Comparison

The evaluated baselines include ResShift Yue et al. (2023), StableSR Wang et al. (2024a), DiffBIR Lin et al. (2024), PASD Yang et al. (2024), SeeSR Wu et al. (2024b), SinSR Wang et al. (2024b), OSEDiff Wu et al. (2024a), PiSA-SR Sun et al. (2025), MoR-DASR He et al. (2026), AdcSR Chen et al. (2025a), and AddSR Tai et al. (2026), with results summarized in Table 1. The unified-protocol entries follow PiSA-SR Sun et al. (2025); AddSR-S1 uses the same benchmarks, FreqDiSR reports three-seed means, and dagger-marked MoR-DASR and AdcSR values are author-reported. FreqDiSR runs at 0.1355 s/image, matching PiSA-SR while remaining faster than the other baselines with reported timings.

FreqDiSR shows its clearest advantage on perceptual distance, achieving the lowest LPIPS on all three benchmarks and the lowest DISTS on RealSR and DRealSR. The largest gains occur on DRealSR: relative to PiSA-SR, FreqDiSR raises PSNR by $0.62\,\mathrm{dB}$ while lowering LPIPS by 0.0140 and DISTS by 0.0084. Its leading CLIPIQA and MANIQA on this benchmark agree with the full-reference perceptual metrics. Across the three benchmarks, the perceptual gains, best or tied-best PSNR, and competitive SSIM place FreqDiSR on a favorable empirical fidelity–perception trade-off under one-step inference. A paired re-evaluation on 93 DRealSR pairs uses a common protocol for FreqDiSR, PiSA-SR, OSEDiff, and SeeSR (PSNR/SSIM on Y; LPIPS/DISTS in RGB). FreqDiSR wins on a majority of images for PSNR, LPIPS, and DISTS against every baseline; the corresponding 95% paired-bootstrap confidence intervals exclude zero, and all one-sided Wilcoxon tests yield $p < 0.002$. Complete paired statistics, including SSIM, are provided in the supplement.

4.3 Qualitative Comparison

Representative comparisons against diffusion-based baselines on paired real-world benchmarks appear in Figure 3. Enlarged regions show continuous lines, coherent geometry, and sharper natural textures, consistent with Table 1.

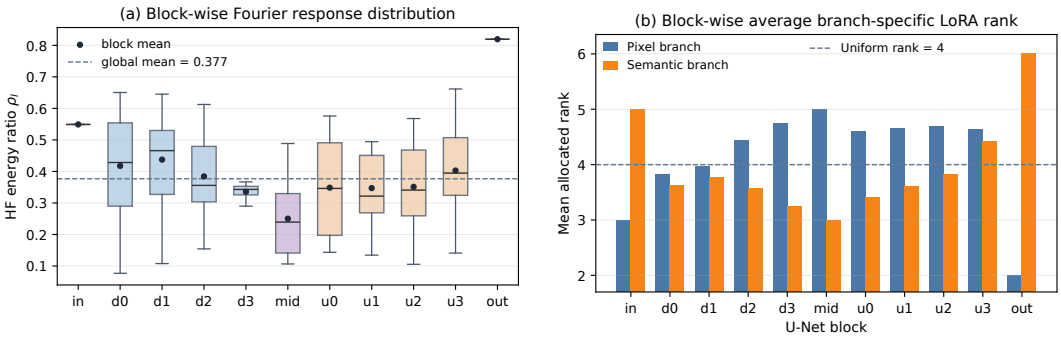

Figure 4: Block-wise Fourier responses (226 profiled modules) and installed branch ranks (258 modules); dashed lines mark the global mean and rank-4 baselines.

4.4 Ablation Study

Controlled ablations use a fixed random seed, with each group varying one design factor relative to the full model, with results reported in Table 2. Measured, inverse, depth-based, and permuted mappings share the same two-branch LoRA budget and 32 non-spatial ranks; the all-module baseline assigns rank 4 throughout. Measured allocation attains the lowest LPIPS and DISTS among the matched mappings, improving them by 0.0073 and 0.0047 over the permuted mean while retaining comparable PSNR and SSIM. In this controlled run, decoder feature modulation improves PSNR and LPIPS while strengthening no-reference quality. Low-frequency-only guidance leads eight of nine metrics; full-band guidance gives the best SSIM.

Appendix F provides the profile-stability, supervision-response, and decoder-modulation diagnostics supporting these controlled comparisons.

5 Conclusion

One-step diffusion SR faces the challenge of preserving LR-supported structure while synthesizing plausible details under a limited adaptation budget. FreqDiSR uses layer-wise Fourier responses to divide a fixed rank budget between pixel and semantic LoRA branches, and extends this frequency guidance to Haar-based decoder modulation and low-frequency supervision. It achieves the lowest LPIPS on all three benchmarks and the lowest DISTS on RealSR and DRealSR with one-step inference; budget-matched allocation controls favor the measured map on LPIPS and DISTS, while profile analyses show stable extraction across the evaluated subsets and degradations. The current method uses a dataset-level rank profile and fixed Haar operators. Future work may explore input- and degradation-aware allocation and learned multiscale transforms while retaining the fixed-budget, one-step setting. These results support layer-wise spectral response as a practical signal for static, fixed-budget capacity allocation in the evaluated one-step SR setting.

Reproducibility Statement

Implementation details, training settings, evaluation protocols, controlled comparisons, and diagnostic analyses are provided in the appendix. The appendix also reports the complete rank-allocation configuration and the statistical evidence supporting the main results. Anonymous core code is available at https://anonymous.4open.science/r/FreqDiSR-6ED5/.

AI Use Statement

Generative AI tools were used solely to edit and polish author-written text for grammar, clarity, and readability. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

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

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6228–6237, 2018.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3086–3095, 2019.

Bin Chen, Gehui Li, Rongyuan Wu, Xindong Zhang, Jie Chen, Jian Zhang, and Lei Zhang. Adversarial diffusion compression for real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28208–28220, 2025a.

Junyang Chen, Jinshan Pan, and Jiangxin Dong. Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28188–28197, 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, 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23174–23184, 2025.

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7323–7334, 2023.

Xiao He, Zhijun Tu, Kun Cheng, Mingrui Zhu, Jie Hu, Nannan Wang, and Xinbo Gao. Mixture of ranks with degradation-aware routing for one-step real-world image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4681–4689, 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. *Advances in neural information processing systems*, 30, 2017.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *Proceedings of the IEEE international conference on computer vision*, pp. 1689–1697, 2017.

Yi Huang, Jiancheng Huang, Jianzhuang Liu, Mingfu Yan, Yu Dong, Jiaxi Lv, Chaoqi Chen, and Shifeng Chen. Wavedm: Wavelet-based diffusion models for image restoration. *IEEE Transactions on Multimedia*, 26:7058–7073, 2024.

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

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Lujun Li, Dezhi Li, Cheng Lin, Wei Li, Wei Xue, Sirui Han, and Yike Guo. AIRA: Activation-informed low-rank adaptation for large models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1729–1739, 2025.

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

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. DiffBIR: Toward blind image restoration with generative diffusion prior. In *European conference on computer vision*, pp. 430–448. Springer, 2024.

Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *ACM Transactions on Graphics (TOG)*, 44(6):1–21, 2025.

Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 773–782, 2018.

Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2013.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2023.

Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE international conference on computer vision*, pp. 4491–4500, 2017.

Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4733–4743, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2333–2343, 2025.

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.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):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 *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1905–1914, 2021.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. SinSR: Diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25796–25805, 2024b.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25456–25467, 2024b.

Chendong Xiang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. A closer look at parameter-efficient tuning in diffusion models. *arXiv preprint arXiv:2303.18181*, 2023.

Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4230–4239, 2023.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1191–1200, 2022.

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.

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.

Chuyan Zhang, Kefan Wang, and Yun Gu. Beyond low-rank tuning: Model prior-guided rank allocation for effective transfer in low-data and large-gap regimes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3337–3345, 2025.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023.

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

U-Net stage	Profiled	Pixel	Semantic
Encoder blocks	81	93	93
Bottleneck	14	16	16
Decoder blocks	129	147	147
Input/output	2	2	2
Total	226	258	258

Table 3: Profiled and adapted U-Net module schema. Pixel and semantic LoRA use the same 258-module insertion set.

A Implementation Details

A.1 Backbone and Training Pipeline

FreqDiSR is built on a one-step latent diffusion SR framework with dual-LoRA adaptation branches. The generative prior is Stable Diffusion 2.1-base. The model uses a frozen VAE, a frozen text encoder, and a U-Net adapted by two LoRA branches: a pixel-oriented branch and a semantic-oriented branch.

The training schedule follows a two-stage protocol. The first 4,000 optimization steps train the pixel-oriented stage. After these steps, training activates the semantic branch, the perceptual and CSD objectives, and FreqDiSR's low-frequency structural anchor. The complete run uses 10,500 optimization steps, a batch size of 4, a learning rate of 5×10^{-5}, and a gradient-accumulation factor of 1. xFormers memory-efficient attention is enabled. The CSD guidance scale is set to 7.5. The diffusion timestep for the one-step model is fixed to 1. For CSD, the minimum and maximum denoising-step ratios are set to 0.02 and 0.5, respectively. AdaIN is used for color-statistics alignment.

A.2 Frequency-Aware Rank Allocation

Both LoRA branches adapt the same 258 convolutional and linear U-Net modules. Before adaptation, forward hooks measure Fourier high-frequency energy ratios for the 226 modules whose outputs form square spatial grids. The frozen VAE encodes center-cropped training targets into clean sampled latents without added diffusion noise. The profile is averaged over 200 images with timestep 1, null-text conditioning, and a radial high-frequency mask beyond half the Nyquist radius. The null condition isolates image-driven layer responses from prompt-specific variation. Its raw ratios range from 0.077 to 0.819, with an average of 0.377, revealing strong spectral heterogeneity across the U-Net.

The 226 scores are normalized across modules to $\bar{\rho}_l \in [0, 1]$ and centered as $\delta_l = \bar{\rho}_l - \frac{1}{2}$. The symmetric mapping $s_l^{\mathrm{sem}} = 1 + \delta_l$ and $s_l^{\mathrm{pix}} = 1 - \delta_l$ treats the normalized midpoint as the uniform allocation and uses the response deviation directly to redistribute capacity between branches. Its multipliers lie in $[0.5, 1.5]$ and sum to 2. With the inherited base rank $\bar{r} = 4$, clipping and rounding install integer ranks in $[2, 6]$ with a pairwise rank sum of 8: the semantic branch receives larger ranks for high-response modules, and the pixel branch receives larger ranks for low-response modules. The other 32 modules are text-side cross-attention K/V projections with non-spatial sequence outputs. Both branches use the same block-wise ranks for these projections: 2/3/4 in the first three encoder blocks, 4 in the bottleneck, and 5/6/8 in the last three decoder blocks. The resulting rank map is reused throughout training and evaluation.

The 226 profiled modules comprise 64 self-attention projections, 32 spatial cross-attention projections, 32 feed-forward projections, 32 transformer input/output projections, 58 ResNet convolutions, 6 sampling convolutions, and 2 input/output convolutions. The 32 additional adapted modules are the non-spatial cross-attention K/V projections described above.

Item	Value
Base model	Stable Diffusion 2.1-base
Training images	LSDIR plus the first 10,000 FFHQ images
Training degradation	Real-ESRGAN online degradation
Learning rate	5×10^{-5}
Batch size	4
Training steps	10,500
Pixel-stage steps	4,000
Base LoRA rank	4 for pixel branch, 4 for semantic branch
Loss weights	$\lambda_{\mathrm{L2}} = 1.0$, $\lambda_{\mathrm{LPIPS}} = 2.0$, $\lambda_{\mathrm{CSD}} = 2.0$
CSD scale	7.5
Inference resolution	512, with $\times 4$ upscaling

Table 4: Main training configuration used by FreqDiSR.

Algorithm 1: Training and Inference Pipeline of FreqDiSR

1: Initialize the pretrained one-step latent diffusion SR backbone.
2: Measure Fourier responses for the 226 spatial-output adapted modules.
3: Allocate complementary pixel/semantic ranks on the profiled modules and retain shared block-wise ranks on the other 32 projections.
4: Train the pixel-oriented branch for the first 4,000 optimization steps.
5: After the first 4,000 steps, activate the semantic LoRA branch, frequency-selective decoder feature modulation, and low-frequency structural anchoring.
6: Continue training through step 10,500.
7: At inference, perform one-step SR with the learned LoRA branches and decoder feature modulation.

A.3 Frequency-Selective Decoder Feature Modulation

Frequency-selective decoder feature modulation is applied to the penultimate and final decoder blocks. Its parameters are frozen during the pixel stage and trained jointly with the semantic LoRA branch during the semantic stage.

The modulation applies a single-level Haar decomposition to a decoder feature map. The three high-frequency subbands are concatenated and passed through an SE-style channel gate with global average pooling, a 1×1 convolution, ReLU, a second 1×1 convolution, and a sigmoid. The gated high-frequency subbands are recomposed by inverse Haar transform. The modulated output is

$$F' = F + \gamma(\tilde{F} - F),$$

where F is the input decoder feature map, $\tilde{F}$ is its Haar-recomposed feature after detail-subband gating, F' is the modulated output, and γ is a learnable scalar initialized to zero. At the beginning of training, the operation preserves the initial mapping of the pretrained one-step SR model.

A.4 Low-Frequency Structural Anchoring

Alongside the inherited full-image objectives, the structural anchor operates in image space. Given prediction $\hat{y}$ and target y, repeated Haar LL extraction defines the auxiliary loss on the low-frequency component:

$$\begin{aligned}\mathcal{L}_{\mathrm{low}} = w_{\ell_2} \|\mathrm{LL}_L(\hat{y}) - \mathrm{LL}_L(y)\|_2^2 \\ + w_{\mathrm{ssim}} \left(1 - \mathrm{SSIM}(\mathrm{LL}_L(\hat{y}), \mathrm{LL}_L(y))\right).\end{aligned}$$

Here $\mathrm{LL}_L(\cdot)$ is the level-L recursively extracted Haar low-frequency component, and L is the number of decompositions. w_{ℓ_2} and w_{ssim} balance the two internal terms. The paper configuration uses $L = 3$, $w_{\ell_2} = 0.5$, $w_{\mathrm{ssim}} = 0.2$, and $\lambda_{\mathrm{low}} = 0.5$, which weights $\mathcal{L}_{\mathrm{low}}$ in the semantic-stage overall objective. The anchor is used in the semantic stage after step 4,000 and omitted at inference.

B Relation to Rank-Allocation Methods

Method	Signal	Allocation unit	Schedule	Budget organization	Test-time router
PiSA-SR Sun et al. (2025)	Branch objectives	Pixel/semantic branches	Static uniform	Uniform rank in both branches	No
AdaLoRA Zhang et al. (2023)	Parameter importance	Singular components across matrices	Training-adaptive	Scheduled global budget	No
AIRA Li et al. (2025)	Activation outlier statistics	Layer/projection total rank	Offline assignment	Cross-layer rank budget	No
SR-LoRA Zhang et al. (2025)	Pretrained-weight stable rank	Layer total rank	Static prior	Cross-layer rank budget	No
MoR-DASR He et al. (2026)	Input degradation cues	Rank-level experts	Input-adaptive	Routed expert budget	Yes
FreqDiSR	Fourier HF activation response	Pixel/semantic ranks per module	Dataset-level static	Fixed local two-branch rank sum	No

Table 5: Mechanism-level comparison with representative dual-branch and adaptive-rank LoRA methods. Local budget denotes the two-branch rank sum at each profiled module.

FreqDiSR differs in both allocation unit and budget organization. Existing adaptive-rank methods primarily vary the total capacity assigned to individual matrices, layers, or routed experts. FreqDiSR preserves a fixed two-branch budget at every profiled module and uses a frozen-backbone spectral profile to redistribute that budget complementarily between fidelity- and perception-oriented branches. The rank map is computed once and reused without inference-time routing.

C Dataset and Evaluation Protocol

C.1 Training Data

Training uses LSDIR and the first 10,000 FFHQ images, following the main paper. Each image is randomly cropped, resized to the target resolution, randomly flipped, and then degraded online using the Real-ESRGAN degradation pipeline. Both LR conditioning images and HR targets are normalized to $[-1, 1]$.

C.2 Test Sets

The evaluation protocol uses three paired benchmarks. DRealSRVal contains 93 image pairs, RealSRVal contains 100 image pairs, and our DIV2K-Val evaluation split contains 3,000 paired patches generated under the same degradation protocol as training. Paired SR and GT images are matched by image identifier, and the same split is used for all methods.

C.3 Metrics

PSNR and SSIM are computed on the luminance (Y) channel in YCbCr space using the ITU-R BT.601 conversion and a data range of 255. LPIPS and DISTS are computed on RGB image pairs using PyIQA. NIQE, CLIPIQA, MUSIQ, and MANIQA are computed on SR images with PyIQA; MANIQA uses its PIPAL-trained `maniqa-pipal` model. FID is a set-level distributional metric computed between the SR and GT sets with Clean-FID in clean mode and a batch size of 16. The crop border is 0 for all reported results. Higher is better for PSNR, SSIM, CLIPIQA, MUSIQ, and MANIQA; lower is better for LPIPS, DISTS, FID, and NIQE.

D Additional Quantitative Evidence

D.1 Main Result Consistency

The FreqDiSR rows reported in the main paper are reproduced in Table 6. The model is trained for 10,500 optimization steps, with the first 4,000 steps used for the pixel-oriented stage. Main-comparison values are means over three training seeds; the controlled ablations and diagnostics use seed 123.

Dataset	Pairs	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
DIV2K-Val	3000	24.70	0.6147	0.2783	0.1929	26.62	4.54	0.6938	68.34	0.6402
RealSR	100	26.31	0.7356	0.2575	0.1947	118.82	5.59	0.6731	67.44	0.6561
DRealSR	93	28.93	0.7736	0.2820	0.2085	131.64	6.12	0.7172	64.79	0.6202

Table 6: Three-seed mean FreqDiSR rows from the main paper. Arrows indicate the preferred direction for each metric.

Mapping	226 P/S	32 P/S	All P/S	$\sum r$	LoRA (M)
Measured	994/814	158/158	1152/972	2124	8.208704
Inverse	814/994	158/158	972/1152	2124	8.208704
Depth-based	848/960	158/158	1006/1118	2124	8.208704
Each permutation	994/814	158/158	1152/972	2124	8.208704
All-module rank-4	904/904	128/128	1032/1032	2064	8.111424

Table 7: Exact rank and LoRA-parameter budgets. P/S denotes pixel/semantic branch rank sums; 226 and 32 identify the spatially profiled and non-spatial module groups.

D.2 Controlled Rank-Allocation Analysis

We compare the measured layer-to-branch correspondence with controls using nonuniform rank distributions. Measured, inverse, depth-based, and permuted mappings change the allocation over the same 226 profiled modules while retaining all other components, training settings, and the 32 non-spatial projection ranks. The all-module rank-4 baseline assigns rank 4 to all 258 adapted modules.

Let $\bar{\rho}_l \in [0,1]$ be the normalized measured response of layer l. The inverse control uses $1-\bar{\rho}_l$. Shuffle$_k$ uses $\tilde{\rho}_l^{(k)} = \bar{\rho}_{\pi_k(l)}$ for three permutations $k \in \{0,1,2\}$, preserving the score multiset while changing its module correspondence. Each map produces complementary branch multipliers. The depth-based control replaces the response with normalized block position. All nonuniform controls share the midpoint-centered mapping, rank interval $[2,6]$, rounding and clipping rule, and two-branch LoRA parameter count; the shuffle row reports the mean and sample SD of the three maps.

Measured-response allocation achieves the best LPIPS, DISTS, CLIPIQA, and MANIQA among the tested mappings while maintaining PSNR and SSIM close to the shuffle mean.

The exact allocation budget is reported in Table 7. Complementary ranks sum to 8 at every profiled module, giving measured, inverse, depth-based, and permuted mappings the same 2,124 two-branch rank sum and 8,208,704 trainable LoRA parameters. The all-module rank-4 baseline has a rank sum of 2,064 and 8,111,424 trainable LoRA parameters. Decoder modulation is shared unchanged across these variants.

D.3 Full-Training Supervision Control

We further compare the supervised frequency support under the complete training schedule. The low-frequency-only configuration applies the additional image-space ℓ_2/SSIM terms to the level-3 low-frequency component, whereas the full-band variant applies them to the complete decoded image. Both retain the inherited full-image objectives; all other components and training settings remain unchanged.

Low-frequency-only guidance achieves higher PSNR, lower LPIPS, DISTS, FID, and NIQE, and higher CLIPIQA, MUSIQ, and MANIQA. Full-band guidance achieves higher SSIM. Low-frequency-only guidance achieves the best value on eight of the nine reported metrics.

D.4 Unified Paired Re-evaluation on DRealSR

To complement the main comparison, we uniformly re-evaluate FreqDiSR, PiSA-SR, OSEDiff, and SeeSR on the same 93 paired DRealSR validation images. PSNR and SSIM are computed on the

Allocation	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
Measured response	28.9767	0.7742	**0.2845**	**0.2082**	134.3062	6.1241	**0.7180**	**65.1864**	**0.6228**
All-module rank-4	28.9492	0.7605	0.2885	0.2096	**132.9202**	**6.0129**	0.6855	63.9022	0.6174
Inverse mapping	28.9414	0.7715	0.2904	0.2157	134.7780	6.1038	0.6831	64.9822	0.6148
Depth-based	**29.0702**	**0.7772**	0.2881	0.2128	133.0778	6.1301	0.6974	64.8351	0.6132
Shuffle mean (sample SD)	28.9846 ±0.0382	0.7752 ±0.0015	0.2918 ±0.0031	0.2129 ±0.0014	133.6556 ±0.9325	6.1800 ±0.1971	0.6921 ±0.0103	64.2838 ±0.9856	0.6030 ±0.0113

Table 8: Controlled rank-allocation results on DRealSR. The shuffle row reports the descriptive mean±sample SD over three distinct permuted profiles.

Supervision	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
Full-band	28.8032	**0.7787**	0.2916	0.2133	134.5376	6.2589	0.6883	64.2235	0.6112
Low-frequency	**28.9767**	0.7742	**0.2845**	**0.2082**	**134.3062**	**6.1241**	**0.7180**	**65.1864**	**0.6228**

Table 9: Full-training supervision control on DRealSR. Both settings use the same 10,500-step configuration and ℓ_2/SSIM coefficients.

Baseline	Metric	Advantage (95% CI)	Win (%)	p
PiSA-SR	PSNR	0.6588 [0.4713, 0.8481]	82.8	2.1×10^{-10}
PiSA-SR	SSIM	-0.0062 [-0.0127, 0.0000]	31.2	9.9×10^{-1}
PiSA-SR	LPIPS	0.0114 [0.0043, 0.0185]	61.3	1.4×10^{-3}
PiSA-SR	DISTS	0.0087 [0.0041, 0.0133]	67.7	3.0×10^{-5}
OSEDiff	PSNR	1.0533 [0.8484, 1.3046]	93.5	9.4×10^{-16}
OSEDiff	SSIM	-0.0093 [-0.0156, -0.0033]	30.1	9.9×10^{-1}
OSEDiff	LPIPS	0.0122 [0.0045, 0.0197]	62.4	5.4×10^{-4}
OSEDiff	DISTS	0.0081 [0.0031, 0.0131]	66.7	3.4×10^{-4}
SeeSR	PSNR	0.7328 [0.3792, 1.0849]	64.5	5.0×10^{-5}
SeeSR	SSIM	0.0044 [-0.0065, 0.0155]	48.4	2.1×10^{-1}
SeeSR	LPIPS	0.0365 [0.0251, 0.0487]	69.9	7.9×10^{-8}
SeeSR	DISTS	0.0232 [0.0170, 0.0316]	73.1	4.5×10^{-8}

Table 10: Unified paired re-evaluation on 93 DRealSR image pairs. Positive values favor FreqDiSR; brackets indicate the 95% paired-bootstrap CI, and p values are from one-sided Wilcoxon signed-rank tests.

Y channel, while LPIPS and DISTS are computed in RGB. For PSNR and SSIM, the per-image advantage is FreqDiSR minus the baseline; for LPIPS and DISTS, it is the baseline minus FreqDiSR, so a positive value always favors FreqDiSR. We use 10,000 paired bootstrap resamples with seed 20260707 for the 95% confidence intervals and report one-sided Wilcoxon signed-rank test p values. Win (%) is the proportion of images with strictly positive advantage.

The unified paired analysis shows consistent PSNR, LPIPS, and DISTS advantages over all three baselines, with the corresponding confidence intervals excluding zero. SSIM exhibits baseline-dependent preferences across the three comparisons.

E Runtime and Reproducibility

The experiments were conducted on one NVIDIA A40 GPU with 46 GB of memory. Inference uses a single denoising pass. The low-frequency anchor is training-only, while the trained decoder feature modulation operates in two selected blocks during the one-step pass.

E.1 Runtime, Parameters, and FLOPs

Pure inference latency, model size, and compute cost are reported in Table 11. The timing uses 100 LR images from RealSRVal because the DRealSR validation set contains 93 images. The reported

Method	Steps	s/image	Params	GMACs	GFLOPs
PiSA-SR	1	0.1351	1289.95	2142.62	4285.23
OSEDiff	1	0.1597	1504.77	2312.48	4624.96
SinSR	1	0.6441	173.92	2094.66	4189.32
ResShift	15	6.1152	173.92	4901.11	9802.22
PASD	20	1.4597	1675.76	23565.85	47131.71
SeeSR	50	4.8887	2250.23	66504.70	133009.41
DiffBIR	50	7.1054	1645.58	47489.56	94979.12
StableSR	200	14.0200	1415.64	82626.85	165253.70
FreqDiSR	1	0.1355	1290.34	2142.62	4285.23

Table 11: Pure inference latency and computational cost. Latency is measured on 100 RealSRVal images; Params are reported in millions, and lower values are better.

Pair	Pearson	Spearman	Norm. MAE	Rank agree.	Rank MAD
A–B	0.9999	0.9998	0.0022	99.12%	0.0088
A–C	0.9999	0.9998	0.0021	99.12%	0.0088
B–C	0.9998	0.9997	0.0024	99.12%	0.0088

Table 12: Profile stability across three disjoint 200-image subsets. Rank agreement and MAD match across branches.

latency excludes one-time initialization and image-I/O overhead. For methods affected by first-run CUDA warm-up, we report the mean after discarding the first inference.

FLOPs are estimated from the executed single-image inference path. The profiler counts unique parameters exercised by that path and multiply-accumulate operations of convolution, transposed-convolution, and linear layers. We report GMACs as the primary compute measure and also include GFLOPs using the common convention $1\ \text{MAC} = 2\ \text{FLOPs}$. All methods use the same hook profiler and input resolution.

F Diagnostic Analyses

Spectral heterogeneity and allocation correspondence. Forward-hook Fourier HF-energy fractions span 0.077–0.819 across U-Net blocks in Figure 4, showing substantial module-wise spectral heterogeneity. Across three disjoint 200-image subsets, Table 12 reports Pearson/Spearman correlations above 0.9997, normalized MAE below 0.0025, and 99.12% exact rank agreement.

Across training-degraded inputs, blur, noise, and compression, profiles retain Spearman correlations of 0.9545–0.9989, with only one-level rank changes. Layer identity explains 89.46% of response variation versus 0.46% for degradation. Profiles from 25 images reach 0.9990 median correlation; dense HF cutoffs and a spectral centroid preserve the ordering with minima of 0.9676 and 0.9859. Complete protocols and per-setting results below support a stable fine-grained capacity map.

Supervised frequency support shapes output spectra. The controlled fine-tuning probe in Figure 5 initializes all configurations from the same trained checkpoint. At the final displayed iteration, low-frequency-only supervision records an output HF energy of 0.0479 versus 0.0355 for full-band supervision, a 35% higher endpoint value. In full training, low-frequency-only supervision outperforms full-band supervision on eight of nine evaluated metrics.

Decoder modulation produces layer-dependent changes. Fourier, spectral-centroid, and multi-level Haar measurements below show complementary spectral redistribution across the selected blocks. The component ablation improves PSNR, CLIPIQA, MUSIQ, and MANIQA and reduces LPIPS.

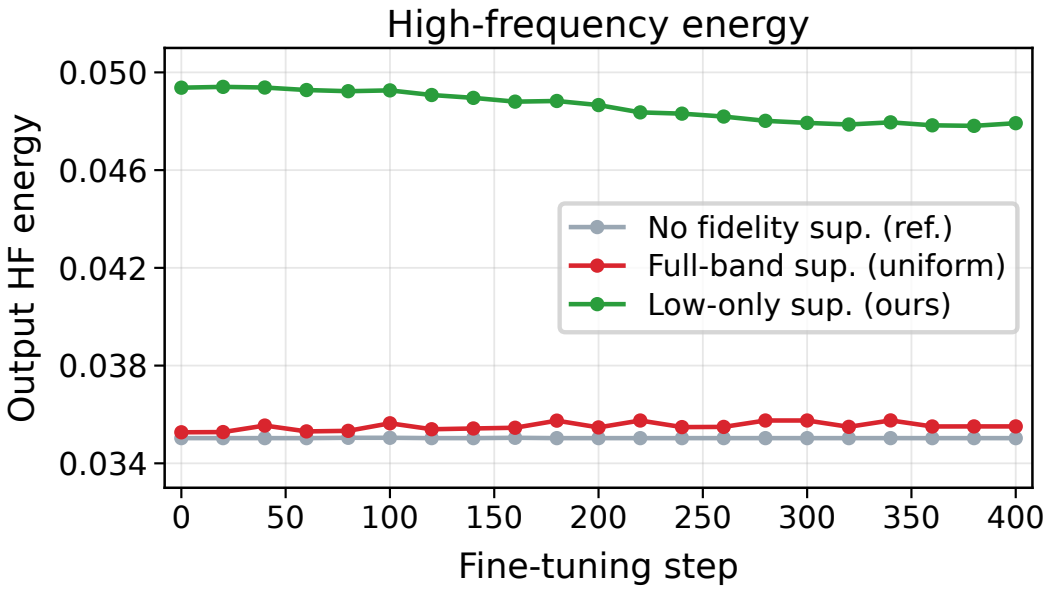

Figure 5: Output HF-energy trajectories; the low-frequency-only endpoint is 35% higher than the full-band endpoint.

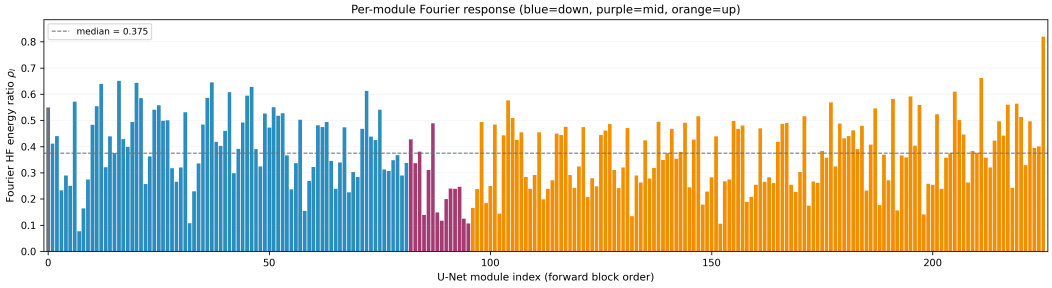

Figure 6: Layer-wise Fourier high-frequency energy profile over 226 frozen spatial-output modules, measured from 200 training targets.

F.1 Layer Frequency Profile

The measured Fourier energy profile is visualized in Figure 6. The response varies both across blocks and within the same block. The final U-Net output convolution has the highest raw response of 0.819, while the minimum response is 0.077. This heterogeneity motivates replacing uniform LoRA rank assignment with frequency-aware allocation.

F.2 Relation to U-Net Structure

We associate each of the 226 responses with its U-Net depth, spatial resolution, output-channel width, parameter count, module family, and module class. Table 13 reports their associations. Continuous size variables are log-transformed. Spearman correlation measures monotonic associations, and η^2 measures the variance associated with categorical descriptors. Layer depth shows negligible response correlation, while spatial resolution and channel width show weak associations.

An ordinary least-squares model using U-Net block, module family, log output-channel width, and log parameter count explains 49.49% of the response variation (45.10% adjusted R^2). A depth-only model explains 1.12%, and a model using spatial resolution, channel width, and parameter count explains 10.38%. The measured profile therefore provides a fine-grained layer description beyond stage position and tensor dimensions.

Descriptor	Statistic	Value	p
Layer depth	Spearman ρ	-0.0741	0.267
Spatial resolution	Spearman ρ	0.2153	0.0011
Output channels	Spearman ρ	-0.2376	0.0003
Parameter count	Spearman ρ	-0.1150	0.0845
U-Net block	η^2	0.1657	–
Module family	η^2	0.2997	–
Module class	η^2	0.0546	–

Table 13: Associations between the measured Fourier response and structural descriptors of the profiled modules.

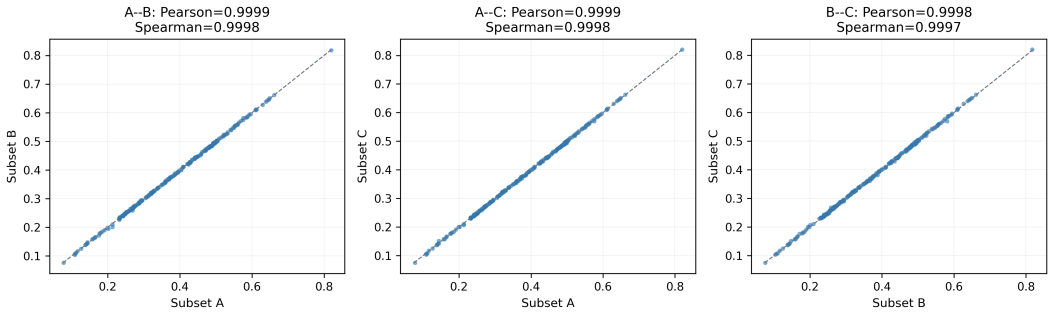

Figure 7: Pairwise comparison of layer-wise Fourier profiles independently extracted from three disjoint 200-image subsets. Each point is one of the 226 spatial-output modules; the dashed line denotes equality.

F.3 Profile Reproducibility across Disjoint Subsets

We evaluate profile reproducibility on three mutually disjoint subsets, denoted A, B, and C, with 200 training images in each subset. Each profile is independently extracted from the same frozen pre-adaptation backbone with timestep 1, null-text conditioning, and LoRA adapters disabled. The comparison covers the same 226 spatial-output modules. For each subset, the estimator averages Fourier high-frequency energy ratios over 50 batches of four images.

We compare profile pairs using Pearson/Spearman correlations and the MAE between independently min–max-normalized responses. Each profile is converted into pixel- and semantic-branch ranks using the same base rank of 4, rank interval $[2, 6]$, and rounding rule as FreqDiSR. Table 3 of the main paper reports the correlations, normalized MAE, exact rank agreement, and rank MAD.

Close alignment among the independently extracted profiles for all three subset pairs is shown in Figure 7. Pearson/Spearman correlations span 0.9998–$0.9999/0.9997$–0.9998, and normalized MAE spans 0.0021–0.0024. For ranks 2–6, the pixel-branch counts are 1/26/98/86/15, 1/26/98/86/15, and 1/27/96/87/15 for subsets A, B, and C; the semantic-branch counts follow the complementary allocation. Every pair agrees on 224 of 226 modules, and each disagreement is one rank level. All three profiles also select the same minimum-response cross-attention projection and the final U-Net output convolution as the maximum-response module. The stable ordering and discretized maps show that the 200-image extraction reproducibly produces the layer-wise rank map.

F.4 Allocation Stability across Spectral Settings

Using subset A and the same frozen profiling protocol, we recompute the response at radial HF thresholds from $\tau = 0.30$ to 0.70 in increments of 0.05 within the same U-Net forward passes. We also compute a threshold-free radial spectral centroid. The latent inputs, timestep, text condition, 226-module set, normalization, and rank conversion remain unchanged. Table 14 compares each profile with the half-Nyquist setting $\tau = 0.50$.



Figure 8: Additional enlarged examples comparing zoomed LR inputs with FreqDiSR outputs. The results illustrate sharper text, signs, and foliage details under real-world degradations.

τ	Pearson	Spearman	Exact	Agree.	max $\lvert\Delta r\rvert$
0.30	0.9797	0.9798	166/226	73.45%	1
0.35	0.9697	0.9676	163/226	72.12%	1
0.40	0.9921	0.9901	211/226	93.36%	1
0.45	0.9854	0.9829	202/226	89.38%	1
0.50	1.0000	1.0000	226/226	100.00%	0
0.55	0.9962	0.9956	217/226	96.02%	1
0.60	0.9961	0.9954	213/226	94.25%	1
0.65	0.9862	0.9848	202/226	89.38%	1
0.70	0.9784	0.9789	193/226	85.40%	1
Centroid	0.9855	0.9859	184/226	81.42%	1

Multiplier	Ranks	Pair sum	$\sum r$	LoRA (M)
$[0.75, 1.25]$	3–5	8	2124	8.208704
$[0.50, 1.50]$	2–6	8	2124	8.208704
$[0.25, 1.75]$	1–7	8	2124	8.208704

Table 14: Response-derived allocations across spectral definitions. The upper block reports rank-map agreement with $\tau = 0.50$; the lower block shows allocation contrast under exact complementary and global budgets.

Response ordering remains strongly aligned across the dense sweep in $[0.30, 0.70]$, with Pearson/Spearman correlations of at least $0.9697/0.9676$. The threshold-free centroid reaches $0.9855/0.9859$ Pearson/Spearman correlation, and every rank change across all spectral definitions is limited to one level. Symmetric multiplier ranges adjust allocation contrast while preserving $r_l^{\text{pix}} + r_l^{\text{sem}} = 8$, the total two-branch rank sum of 2,124, and 8.208704M trainable LoRA parameters. The base rank $\bar{r} = 4$ follows the underlying dual-LoRA configuration.

F.5 Cross-Input Profile Stability

Using the same 200 training images, each center-cropped target is paired with the conditioning image produced by the training degradation pipeline and resized to the target resolution. Both profiles use clean sampled VAE latents, the same frozen pre-adaptation backbone, timestep 1, null-text conditioning, and no added diffusion noise. Table 15 compares their responses and discretized ranks.

Profiles	Pearson	Spearman	NMAE	Exact	Max. $\lvert\Delta r\rvert$
Target–LR	0.9591	0.9544	0.0418	194/226	1
Target repeat	0.99999	0.99997	0.00083	226/226	0

Table 15: Cross-input stability of the dataset-level spectral profile. Exact-rank agreement applies identically to both complementary branches.

Input	Pearson	Spearman	Exact	Agree.	$\max \lvert\Delta r\rvert$
Training degradation	0.9591	0.9545	194/226	85.84%	1
Mild blur	0.9974	0.9967	218/226	96.46%	1
Severe blur	0.9713	0.9680	191/226	84.51%	1
Mild noise	0.9992	0.9989	223/226	98.67%	1
Severe noise	0.9936	0.9916	211/226	93.36%	1
JPEG-60	0.9862	0.9861	211/226	93.36%	1
JPEG-20	0.9737	0.9715	204/226	90.27%	1

Table 16: Profile stability across training and controlled degradations. Rank agreement applies identically to the complementary branches.

Target and training-degraded profiles show strong agreement, with Pearson/Spearman correlations of $0.9591/0.9544$. Their rank maps match exactly on 194 of 226 modules, and every remaining assignment differs by one level. An independent target rerun reproduces all 226 ranks. Together with the 99.12% exact agreement across disjoint subsets, these results show a persistent module ordering across the evaluated inputs and subsets.

F.6 Robustness across Degradations and Profile Sizes

The same 200 paired contents and posterior sampling noise are used to compare the target profile with the training degradation and controlled transformations. Gaussian blur uses standard deviations $1.2/3.0$, additive Gaussian noise uses standard deviations $5/20$ on the $[0, 255]$ scale, and JPEG compression uses quality factors $60/20$. Table 16 reports profile and rank-map agreement with the target setting.

Group-wise Spearman correlations span 0.9545–0.9989, and every changed assignment is limited to one rank level. Individual-image profiles retain median Spearman correlations of 0.9727–0.9862 with their group profile. A balanced sums-of-squares analysis uses

$$\rho_{l,d,i} = \mu + a_l + b_d + c_i + \varepsilon_{l,d,i}, \tag{8}$$

where l, d, and i index the profiled module, degradation, and image; a_l, b_d, and c_i are their additive effects. The remaining term collects interactions and residual variation.

Source	Response variation (%)
Layer identity	89.46
Image content	3.82
Degradation	0.46
Residual interactions	6.26

Table 17: Balanced sums-of-squares partition of profile variation.

Profile-size convergence is measured by drawing 200 random subsets without replacement at each size. Table 18 reports the median correlation and the median, 5th-percentile, and 95th-percentile rank agreement with the 200-image profile.

Images	Median ρ_s	Median agree.	5th pct.	95th pct.
25	0.9990	98.23%	95.13%	99.12%
50	0.9995	98.67%	96.90%	99.56%
100	0.9998	99.12%	98.23%	100.00%
200	1.0000	100.00%	100.00%	100.00%

Table 18: Convergence of response-derived ranks with the number of profiling images. Every observed disagreement is one rank level.

F.7 Supervision Probe

The three supervision settings initialize from the same trained checkpoint. Full-band and low-frequency-only supervision use their designated additional ℓ_2 terms, while the no-additional-supervision curve evaluates the unchanged model. Four batches are cycled within each setting, and output HF energy is recorded every 20 iterations after the corresponding update. HF energy is the sum of the mean absolute LH, HL, and HH coefficients after a single-level Haar decomposition, with each mean taken over batch, channel, and spatial elements. At the final displayed iteration, low-frequency-only supervision records 0.0479 versus 0.0355 for full-band supervision, a 35% higher endpoint value. The same supervision settings are evaluated under the complete training schedule across all nine metrics in Table 9.

F.8 Decoder Feature Modulation Analysis

The modulation control only disables decoder modulation; all other components and training settings remain unchanged. The diagnostic compares internal high-frequency feature energy with decoder feature modulation enabled and with its learned residual gate set to zero. The same batches are used in both cases. HF feature energy uses the same single-level Haar statistic as the supervision probe. The percentages are relative to the corresponding $\gamma = 0$ mean energy. As shown in Figure 9, the penultimate decoder block increases high-frequency energy by 1.66%, while the final decoder block decreases it by 0.83%. The opposite signs show a layer-specific modulation pattern.

The learned gates are evaluated on the 93 DRealSR pairs using Fourier HF ratio, radial spectral centroid, and three recursively extracted Haar detail levels. Each value in Table 19 is measured relative to the same model with both residual gates set to zero. Fourier, centroid, and first-level Haar statistics agree on enhanced fine-scale response in the penultimate block and attenuation in the final block. The deeper Haar levels show the complementary redistribution of coarser detail responses.

Decoder block	FFT HF	Centroid	Haar L1	Haar L2	Haar L3
Penultimate	+1.38	+0.73	+2.75	−0.54	−0.54
Final	−1.25	−1.16	−1.49	+0.38	+0.34

Table 19: Relative spectral-response changes (%) produced by the learned decoder gates. All statistics use identical DRealSR inputs and paired latent samples.

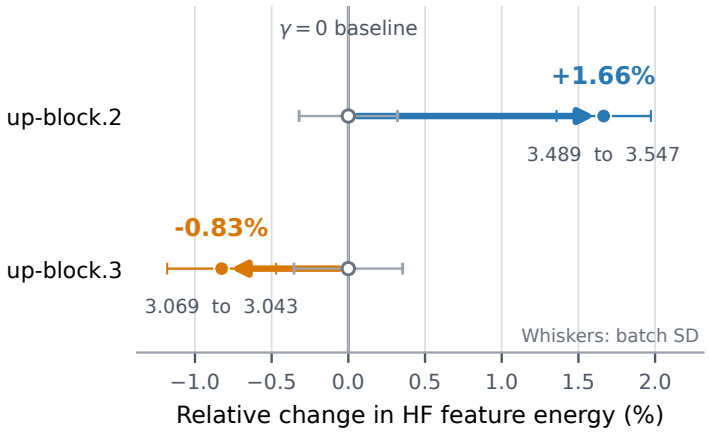

Figure 9: Layer-specific change in internal high-frequency feature energy under decoder feature modulation. Each block is normalized to its $\gamma = 0$ mean; labels report raw means and whiskers show batch SD.

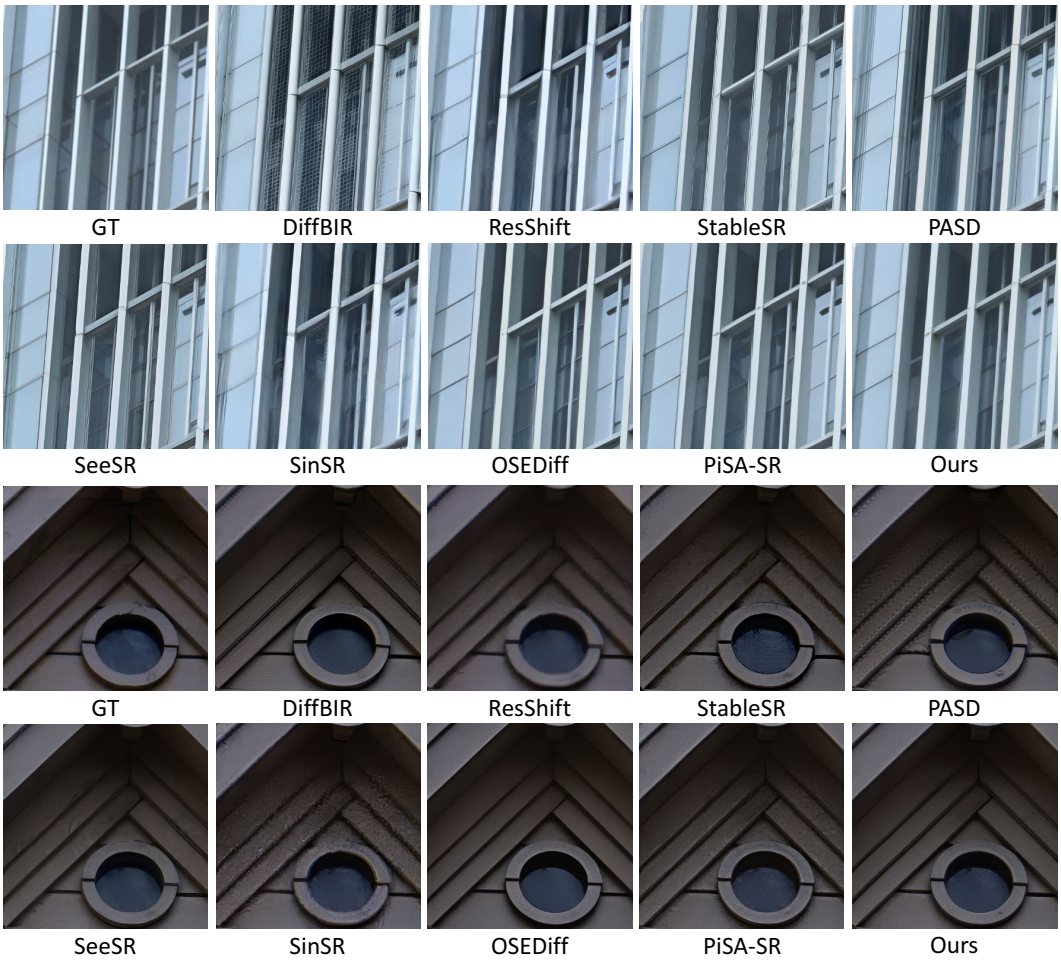

Figure 10: Additional qualitative comparison with representative SR baselines. FreqDiSR produces clearer local structures and more natural textures while preserving the overall geometry of the input scene.

G Additional Visual Comparisons

Additional qualitative comparisons with representative diffusion-based SR baselines are provided in Figure 10. The examples cover structured facades and low-light architectural details, where both geometric consistency and texture recovery are important. FreqDiSR preserves straighter local structures with clearer, naturally scaled textures.

Enlarged LR inputs and the corresponding FreqDiSR outputs on challenging real-world regions appear in Figure 8. These examples include text, signs, and foliage, where the LR inputs contain severe blur or missing high-frequency information. FreqDiSR recovers sharper visual details while maintaining plausible local structure.

H Limitations and Responsible Use

FreqDiSR combines a reusable dataset-level Fourier profile, single-level Haar decoder modulation, and a level-3 Haar training anchor without adding denoising steps. Its outputs are perceptual reconstructions and should be accompanied by the source input in evidence-sensitive applications.