HIFIPIX: BRINGING HIGH-FIDELITY REAL-WORLD SUPER-RESOLUTION BACK TO PIXEL SPACE

Anonymous authors
Paper under double-blind review



Figure 1: We propose HiFiPix, a Real-ISR framework built on a pixel-space large generative model, achieving high-fidelity restoration of real-world LQ images and superior perceptual quality compared with existing latent-space Real-ISR methods.

ABSTRACT

Diffusion-based real-world image super-resolution suffers from limited reconstruction fidelity due to information loss in VAE-based latent representations. Recently, pixel-space diffusion models offer a promising high-faithfulness alternative by avoiding latent compression, but adapting them to Real-ISR remains challenging due to inherent patch boundary artifacts and degraded generative capability. To address the issues and maximize the potential of pretrained pixel-space diffusion models, we propose **HiFiPix**, a pixel-space Real-ISR framework capable of ultra high-fidelity reconstruction and fine-grained visual generation effects. To suppress patch boundary artifacts, we penalize unrealistic high-frequency periodic components with a Cross-Scale Wavelet Discriminator (CWSD) for image-level supervision, and Mixup Contrastive Self-Distillation (MixCSD) for intermediate representation regularization. Furthermore, we uncover a heterogeneous generation-fidelity trade-off across low-rank adapters (LoRA) and propose Selective LoRA Pruning (LoRA-SP) to adaptively prune LoRA layers, enhancing generative capability with minimal deterioration of reconstruction fidelity. Extensive experiments demonstrate that HiFiPix consistently outperforms existing state-of-the-art latent-space methods in terms of both restoration accuracy and perceptual quality by a large margin. Code and models will be released.

1 INTRODUCTION

Real-world image super-resolution (Real-ISR) targets the recovery of low-quality (LQ) images suffering from complex unknown degradations, with the goal of reconstructing high-quality (HQ) counterparts that exhibit both high fidelity and satisfying perceptual quality. Initially, GAN-based algorithms Wang et al. (2021); Chen et al. (2023) employ a discriminator to match the distribution of reconstructed images with that of real-scenario images. However, lacking powerful generative priors, GAN-based models exhibit limited capability for recovering fine-grained details.

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 2: Fine-tuning pixel models faces two challenges compared with latent models: **(a)** without VAE decoders to preserve spatial continuity, pixel-space models are prone to patch boundary artifacts; **(b)** training progress tends to favor high-fidelity restoration at the expense of generative capability.

Over the past few years, diffusion-based generative models Rombach et al. (2022); Labs (2025) have advanced Real-ISR by producing highly photo-realistic details. However, existing Real-ISR methods Dong et al. (2025); Fang et al. (2026) typically rely on the Variational Autoencoder (VAE) Esser et al. (2021) to compress LQ inputs into the latent space for efficient denoising. This irreversible compression discards the subtle textures from LQ inputs, often leading to structural inconsistencies, ghosting artifacts, and ultimately compromised restoration fidelity. To mitigate this issue, prior works try to improve input quality Lin et al. (2024), reduce the VAE compression ratio Yi et al. (2025), or inject LQ features into denoising networks Yang et al. (2024). However, these strategies fail to fundamentally address the information loss introduced by latent space compression, yielding reduced generative capability (*e.g.*, blurry leaves in the top row of Fig. 1) or limited fidelity gains (*e.g.*, distorted characters in the bottom row of Fig. 1). As a result, restoring highly faithful textures while preserving excellent visual quality still remains a significant challenge in Real-ISR.

More recently, pixel-level text-to-image (T2I) diffusion models Yu et al. (2026); Cai et al. (2026) have demonstrated effectiveness in mitigating reconstruction accuracy issues by removing the VAE component. However, roughly fine-tuning pixel-space models to Real-ISR via LoRA adapters still faces two key challenges. First, pixel-space models typically adopt patch-based tokenization, while different tokens only interact within pure MLP-based attention blocks. Unlike VAE-based ones Saharia et al. (2022); Esser et al. (2024), whose convolutional decoders naturally preserve spatial continuity, pixel-space decoding does not explicitly enforce information interaction across patch boundaries. This leads to discontinuities between patches and periodic grid artifacts, particularly in texture-rich regions. As shown in Fig. 2 (a), under the same training strategy, pixel-space models suffer from more severe periodic artifacts than latent models Esser et al. (2024); Labs (2025), degrading the structural consistency across neighboring patches. Second, as depicted in Fig. 2 (b), although removing the VAE eliminates latent compression, it discards the compact semantic representations, making high-level semantic modeling more challenging in pure pixel space. Therefore, this weakens the generative prior and biases models toward deterministic reconstruction rather than plausible detail synthesis. Consequently, models tend to overfit reasonable but over-smoothed structures, while losing their ability to generate missing high-frequency details, especially under severe degradations.

To address these challenges, we propose *HiFiPix*, a pixel-space framework for high-fidelity Real-ISR. We first focus on patch-boundary artifacts arising from pixel-space decoding and observe that they mainly manifest as discontinuities in high-frequency structures. This motivates us to exploit coherent high-frequency information for cross-patch continuity. Specifically, we introduce a *Cross-Scale Wavelet Discriminator (CSWD)* that propagates high-frequency features across scales for boundary-aware image-level supervision, and further propose a *Mixup Contrastive Self-Distillation (MixCSD)* that leverages an HQ-mixed branch to regularize intermediate representations at feature levels. Together, they provide complementary external and internal supervision to suppress patch-boundary artifacts. Furthermore, we address degradation of generative capability during training. We observe that LoRA updates across layers are not equally necessary: some deep layers merely echo transformations already made by shallower layers, contributing to overfitting pixel-level characteristics rather than preserving high-level semantics. We thus propose *Selective LoRA Pruning (LoRA-SP)*, which detects and prunes such redundant layers via Centered Kernel Alignment-based representational Kornblith et al. (2019), protects fidelity-critical layers from pruning, and progressively freezes rest, thereby constraining fine-tuning capacity precisely where overfitting risks occur.

Furthermore, existing Real-ISR benchmarks are hindered by low-quality ground truth, suffering from missing fine textures Cai et al. (2019); Wei et al. (2020) or background de-focus Wu et al. (2026). As noted in A-FINE Chen et al. (2025a), generative restoration models may yield images

visually superior to these flawed GT references, rendering nearly all reference-based fidelity metrics unreliable. To address this issue, we construct *Dual-Q-Bench*, an ultra-high-quality paired real-world ISR benchmark captured via screen photography, whose HQ reference images feature rich textures and minimal background de-focus. Dual-Q-Bench enables reliable evaluation of both fidelity and perceptual quality. Our contributions are summarized as follows:

- We propose *HiFiPix*, a Real-ISR method built on a pretrained pixel-level model, shifting Real-ISR from latent-space to pixel-space modeling, achieving high-fidelity restoration without sacrificing generative capability or inference efficiency.
- We propose *CSWD* that propagates high-frequency features across scales for boundary-aware image-level supervision, and *MixCSD*, which distills clean representations to enhance patch-level discriminability, jointly mitigating patch-boundary artifacts.
- We propose *LoRA-SP*, which identifies functionally redundant LoRA layers via representational drift and progressively prunes them while protecting fidelity-critical layers, preserving the pretrained generative prior during supervised fine-tuning.
- We introduce *Dual-Q-Bench*, a high-quality paired Real-ISR benchmark that addresses the low-quality GT and background de-focus in existing benchmarks, enabling more reliable evaluation of both generative capability and reconstruction fidelity.

2 Related Works

2.1 Real-world Image Super-Resolution

With the advent of large-scale T2I models, diffusion-based Real-ISR methods have surpassed GAN-based approaches. Existing methods utilize LQ input to control diffusion models for faithful restoration by leveraging pluggable adapters Chen et al. (2025b), dual-stream architectures Yang et al. (2024), ControlNet Yu et al. (2024), initialized restoration modules Lin et al. (2024), semantic conditioning Wu et al. (2026), and large-scale high-quality training datasets Yu et al. (2024). Furthermore, to fully use input information, methods like OSEDiff Wu et al. (2024b) and TSD-SR Dong et al. (2025) replace random noise with LQ images, while S3Diff Zhang et al. (2024) and ODTSR Fang et al. (2026) incorporate GAN loss to enhance perceptual quality. TVT Yi et al. (2025) increases the VAE latent resolution to improve fidelity, whereas LucidFlux Fei et al. (2026) and StrSR Wang et al. (2026a) improve fidelity by fine-tuning FLUX Labs (2025) or Z-image Cai et al. (2025) with increased VAE channel capacity. Nevertheless, these methods still operate in latent spaces, where VAE compression inevitably limits reconstruction fidelity. In contrast, we perform Real-ISR directly in pre-trained pixel spaces Cai et al. (2026), enabling high-fidelity and perceptually pleasing Real-ISR.

2.2 Pixel-space Diffusion Models

Early pixel-space diffusion models Dhariwal & Nichol (2021); Saharia et al. (2022) suffered from prohibitive computational costs at high resolutions. Recent works Chen et al. (2025c); Wang et al. (2026b); Li & He (2026); Ma et al. (2026b); Lu et al. (2026b); Ma et al. (2026a); Yu et al. (2026) have revisited pixel-space diffusion by improving its training efficiency, sampling strategy, and architectural design. Among them, PixelFlow Chen et al. (2025c) enables efficient end-to-end pixel-space generation, while JiT Li & He (2026) and PixelDiT Yu et al. (2026) further improve the scalability and efficiency of pixel-space modeling. HiDream-O1-Image Cai et al. (2026) extends this paradigm to a large-scale pixel-space T2I diffusion transformer with a unified generative architecture. Inspired by these advances, we explore fine-tuning pre-trained pixel-space T2I models for Real-ISR, aiming to achieve high-fidelity restoration and visually pleasant generation without VAE compression.

3 Method

3.1 Empirical Observations

Following classical one-step fine-tuning paradigms Yi et al. (2025), we identify two critical insights when applying HiDream-O1-Image for Real-ISR. **First**, patch-boundary artifacts may not arise

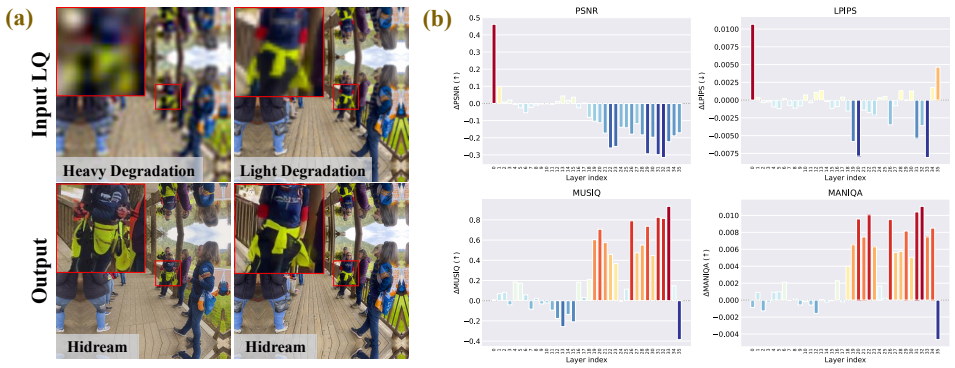


Figure 3: **(a)** Heavier boundary artifacts when super-resolving extremely blurry input. **(b)** Removal of individual LoRA layers reveals trade-offs between reconstruction and generative capabilities.

solely from the patch-wise architecture of pixel-space models. As shown in Fig. 3 (a), grid-like artifacts become more pronounced under severe degradation than under mild degradation, suggesting that imperfect high-frequency reconstruction is an important cause of boundary artifacts, as severe degradation makes high-frequency reconstruction more challenging. This observation motivates a low-cost and architecture-agnostic strategy to explicitly regularize high-frequency components. Specifically, we design the Cross-Scale Wavelet Discriminator, which penalizes spurious boundary artifacts via multi-scale information constraints. Besides, we propose Mixup Contrastive Self-Distillation to ensure faithful generation of genuine high-frequency details.

Second, simply applying LoRA layers yields SR outputs lacking fine-grained details, indicating that such a general strategy drastically reduces the generative capability of pixel-space models. As shown in Fig. 3 (b), we perform ablation studies to investigate individual LoRA layers, then reveals that generative capacity and fidelity are non-uniformly entangled across depth: removing certain layers boosts generative expressiveness at negligible fidelity cost, whereas removing others yields a similar generative gain but at the expense of severe fidelity degradation. This heterogeneity rules out uniform layer pruning or fine-tuning as an effective strategy; therefore, we propose Selective LoRA Pruning that identifies and removes layers that mainly reduce the generation ability.

3.2 Cross-Scale Wavelet Discriminator

Inspired by Wu et al. (2025); Wang et al. (2026a) that leverage DINOv3-ConvNeXt Siméoni et al. (2025) as a discriminator to enhance the visual quality, we propose **Cross-Scale Wavelet Discriminator (CSWD)** to extract multi-scale features and penalize the unrealistic high-frequency periodic components from boundaries, thereby alleviating patch boundary artifacts. As shown in Fig. 4, given the super-resolved image I_{SR}, the frozen DINOv3-ConvNeXt first extracts multi-level features $\boldsymbol{F}_1 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$, $\boldsymbol{F}_2 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2C}$, and $\boldsymbol{F}_3 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 4C}$. Different from Wu et al. (2025) that relies on pooling or convolution layers for multi-scale down-sampling, we employ the wavelet transform (WT) to decompose features into a low-frequency sub-band $\boldsymbol{LL}$ and three high-frequency sub-bands $\{\boldsymbol{LH}, \boldsymbol{HL}, \boldsymbol{HH}\}$, all with half the spatial resolution of input features. We discard $\boldsymbol{LL}$ and concatenate the three high-frequency sub-bands along the channel dimension, thereby focusing the discriminator on edges, textures, and periodic artifacts. The concatenated features are then compressed by trainable convolutional layers, while cascaded WT-convolution layers progressively extract deeper high-frequency representations and produce a pixel-wise discriminative score map.

To further exploit cross-scale information, we introduce a cross-scale high-frequency injection mechanism. Specifically, high-frequency sub-bands $\left\{\boldsymbol{LH}_i^k, \boldsymbol{HL}_i^k, \boldsymbol{HH}_i^k\right\}$ extracted by the k-th WT block WT_k from level $\boldsymbol{F}_i$ are aggregated via a 1×1 convolution, then injected into features of next level $\boldsymbol{F}_{i+1}$. This progressively propagates fine-grained high-frequency details from shallow to deep layers, enhancing the discriminator's ability to detect multi-scale periodic boundary artifacts.

$$WT_k\left(Conv\left(\boldsymbol{F}_i\right)\right) = \left\{\boldsymbol{LL}_i^k, \boldsymbol{LH}_i^k, \boldsymbol{HL}_i^k, \boldsymbol{HH}_i^k\right\}, \boldsymbol{H}_i^k = Conv\left(\boldsymbol{LH}_i^k + \boldsymbol{HL}_i^k + \boldsymbol{HH}_i^k\right), \quad (1)$$

$$WT_{k-1}\left(Conv\left(Concat\left(\boldsymbol{H}_i^k, \boldsymbol{F}_{i+1}\right)\right)\right) = \left\{\boldsymbol{LL}_{i+1}^k, \boldsymbol{LH}_{i+1}^k, \boldsymbol{HL}_{i+1}^k, \boldsymbol{HH}_{i+1}^k\right\}. \quad (2)$$

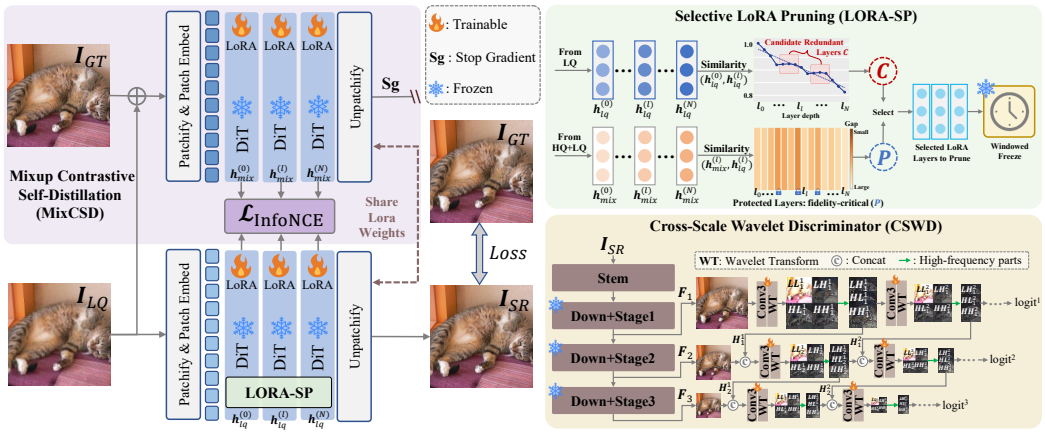


Figure 4: Architecture of HiFiPix. We introduce MixCSD and CSWD to alleviate patch boundary artifacts, and propose LoRA-SP to enhance the model's generative capability.

The multi-level discriminator progressively produces pixel-wise logit maps $\boldsymbol{M} \in \{\mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 1}, \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 1}, \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 1}\}$ for PatchGAN-style Isola et al. (2017) local realism discrimination.

3.3 Mixup Contrastive Self-Distillation

To further suppress boundary artifacts while promoting the faithful generation of clean details, as shown in Fig. 4, we introduce **Mixup Contrastive Self-Distillation (MixCSD)** at intermediate layers of DiT blocks. Specifically, for each training batch, we construct two forward branches sharing the same set of LoRA weights $\boldsymbol{f}_\theta$ and the same DiT weights. The anchor branch takes the pure LQ image $\boldsymbol{I}_{LQ}$ as input, with gradients flowing normally, and extracts patch-token hidden states at DiT layer ℓ:

$$\boldsymbol{h}^{\ell}_{lq} = f^{(\ell)}_{\theta}(\boldsymbol{I}_{LQ}) \in \mathbb{R}^{B\times N\times D}, \tag{3}$$

where B, N, and D denote the batch size, the number of patch tokens, and the hidden dimension, respectively. Here we omit the DiT weights in all equations for simplicity of expression. The positive branch takes a mixture of GT and LQ as input with mixing ratio $\alpha \in [0, 1]$, then feeds forward through sharing the anchor branch weights under a stop-gradient operator sg:

$$\boldsymbol{h}^{\ell}_{mix} = \mathrm{sg}\Big(f^{(\ell)}_{\theta}(\alpha\, \boldsymbol{I}_{GT} + (1-\alpha)\, \boldsymbol{I}_{LQ})\Big), \tag{4}$$

since the mixed input injects genuine high-frequency content and consistent cross-patch boundary structure, $\boldsymbol{h}^{\ell}_{mix}$ carries more coherent neighborhood features at patch boundaries than $\boldsymbol{h}^{\ell}_{lq}$, serving as an implicit target for the anchor branch along the "boundary-consistency" dimension. After ℓ_2-normalization of the anchor (*e.g.*, $\bar{\boldsymbol{h}}_{lq,i} = \frac{\boldsymbol{h}_{lq,i}}{\|\boldsymbol{h}_{lq,i}\|_2}$), we treat each patch-token as a contrastive unit (flattened index $i = 1, ..., BN$) and define an InfoNCE loss Oord et al. (2018) with temperature τ:

$$\mathcal{L}^{(\ell)}_{\text{InfoNCE}} = -\frac{1}{BN}\sum_{i=1}^{BN} \log \frac{\exp\Big(\bar{\boldsymbol{h}}^{\top}_{lq,i}\bar{\boldsymbol{h}}_{mix,i}/\tau\Big)}{\sum_{j=1}^{BN}\exp\big(\bar{\boldsymbol{h}}^{\top}_{lq,i}\bar{\boldsymbol{h}}_{mix,j}/\tau\big)}, \tag{5}$$

where $j = i$ term in the denominator corresponds to the positive pair and the remaining $BN-1$ terms serve as negatives, and $\top$ denotes the matrix transpose. Negatives are sampled from other patches in the same image or other images in the same training batch. This encourages each anchor to approach its high-frequency-consistent target while moving away from other patches, particularly spatially adjacent ones, thereby suppressing patch boundary artifacts. A stop-gradient operation enforces one-way supervision from the positive branch to the anchor branch, preventing representational collapse. Multi-layer aggregation and step-gated weighting follow the same formulation as follows:

$$\mathcal{L}_{\text{InfoNCE}} = \lambda_{\text{InfoNCE}}(t)\cdot\frac{1}{|\mathcal{S}|}\sum_{\ell\in\mathcal{S}}\mathcal{L}^{(\ell)}_{\text{InfoNCE}}, \lambda_{\text{InfoNCE}}(t) = \begin{cases}\lambda^{\max}_{\text{InfoNCE}}, & t \geq t_{\text{start}}\\ 0, & t < t_{\text{start}}\end{cases}, \tag{6}$$

where $\mathcal{S}$ is the number of supervised DiT layers, t is the training step. $\lambda^{\max}_{\text{InfoNCE}}$ and t_{start} denote the target loss weight and the step at which it is activated, respectively.

3.4 SELECTIVE LORA PRUNING

As shown in Fig. 4, we introduce **Selective LoRA Pruning (LoRA-SP)**, a training-time mechanism that leverages the network's internal representations to identify and freeze functionally redundant LoRA branches, thereby improving generative capacity without sacrificing restoration fidelity.

At each probing step, we forward the degraded input $\boldsymbol{I}_{LQ}$ and mixture $\boldsymbol{x}_{mix} = \alpha \boldsymbol{I}_{GT} + (1-\alpha)\boldsymbol{I}_{LQ}$ through the network to gather per-layer hidden states $h^{(l)}$. Layer-wise similarity is quantified by linear Centered Kernel Alignment (CKA) Kornblith et al. (2019) and defined as:

$$\mathrm{CKA}(\boldsymbol{H}_1, \boldsymbol{H}_2) = \frac{\|\boldsymbol{H}_1^\top \boldsymbol{H}_2\|_F^2}{\|\boldsymbol{H}_1^\top \boldsymbol{H}_1\|_F \, \|\boldsymbol{H}_2^\top \boldsymbol{H}_2\|_F}, \tag{7}$$

where $\boldsymbol{H}_1$, $\boldsymbol{H}_2$ and F denote two intermediate features and Frobenius norm, respectively. We derive two EMA-smoothed per-layer metrics: anchor similarity $\bar{\boldsymbol{a}}^{(l)}$ via $\mathrm{CKA}(\boldsymbol{h}_{lq}^{(l)}, \boldsymbol{h}_{lq}^{(r)})$ with shallow reference layer r (expected to decay under generative processing), and fidelity gap $\bar{\boldsymbol{g}}^{(l)} = 1 - \mathrm{CKA}(\boldsymbol{h}_{lq}^{(l)}, \boldsymbol{h}_{mix}^{(l)})$, measuring layer-$l$'s representation shift toward high-quality targets. A layer becomes a redundancy candidate when its anchor similarity fails to drop from the prior layer,

$$\mathcal{C} = \left\{ l \;\middle|\; \bar{\boldsymbol{a}}^{(l)} - \bar{\boldsymbol{a}}^{(l-1)} \geq 0 \right\}, \tag{8}$$

where layers in $\mathcal{C}$ introduce no distinct additional representation transformation. Candidates are ranked and filtered against a small protected set $\mathcal{P}$ of the globally most fidelity-critical layers, which are vetoed from pruning. The surviving top-ranked candidates constitute the final pruning set,

$$\mathcal{S} = \{\, l \in \mathcal{C} \setminus \mathcal{P} \,\}, \qquad |\mathcal{S}| \leq K, \tag{9}$$

bounded by a fixed pruning budget K. Selected layers are not pruned abruptly: early in training they undergo a temporary windowed freeze, later transitioning to a permanent freeze once the redundancy signal stabilizes, with a soft-decay fallback when the pruning budget is saturated. This graduated schedule keeps LoRA-SP conservative at early training stages while achieving persistent capacity reduction over time, enabling improved generation without severe fidelity degradation.

3.5 TRAINING OBJECTIVE

Following Fang et al. (2026), one-step LoRA fine-tuning paradigm is adopted. As shown in Fig. 4, we replace the original Gaussian noise with LQ image as the input. The overall backbone generates fine-grained details in pure pixel-space by predicting the residual between LQ and GT image.

We fix pretrained DiT blocks, optimize LoRA with a reconstruction loss and an adversarial loss, yielding the total objective $\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{adv}}$. $\mathcal{L}_{\mathrm{rec}}$ aggregates four terms: pixel-space regressive loss $\mathcal{L}_1$, a DINOv3-ConvNeXt-based Siméoni et al. (2025) perceptual loss $\mathcal{L}_{per}$, frequency-space loss Wang et al. (2026a) $\mathcal{L}_{freq}$, and the InfoNCE contrastive loss $\mathcal{L}_{\mathrm{InfoNCE}}$ (Sec. 3.3).

$$\mathcal{L}_{\mathrm{rec}} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{per} + \lambda_3 \mathcal{L}_{freq} + \lambda_4 \mathcal{L}_{InfoNCE}. \tag{10}$$

For $\mathcal{L}_{\mathrm{adv}}$, we employ two complementary discriminators built on the same DINOv3-ConvNeXt features: a main discriminator D_{main} that performs multi-scale assessment of structural realism, guiding the model to produce images with higher visual authenticity, and our Cross-Scale Wavelet Discriminator D_{wave} (Sec. 3.2), which applies a wavelet transform to retain accurate high-frequency sub-bands, while penalizing texture/edge artifacts, especially boundary artifacts across patches:

$$\mathcal{L}_{\mathrm{adv}} = \lambda_{\mathrm{main}} \mathbb{E}_{\hat{x}} \big[-\log \sigma(D_{\mathrm{main}}(\hat{x})) \big] + \lambda_{\mathrm{wave}}(t)\, \mathbb{E}_{\hat{x}} \big[-\log \sigma(D_{\mathrm{wave}}(\hat{x})) \big], \tag{11}$$

where $\lambda_{\mathrm{wave}}(t)$ is linearly warmed up to avoid the high-frequency term saturating at an early stage.

3.6 CONSTRUCTION OF DUAL-Q-BENCH

Due to page limitations, please refer to Appendix A.1 for construction details of Dual-Q-Bench.

Table 1: Quantitative results on real-world benchmarks, with the best/second-best results highlighted in **bold**/underlined. "-s" denotes the number of diffusion steps.

Datasets	Metrics	SUPIR -s50	FaithDiff -s20	Dit4SR -s40	VOSR -s25	LucidFlux -s15	TSDSR -s1	PiSA-SR -s1	TVT -s1	ODTSR -s1	StrSR -s1	**Ours** -s1
DrealSR	PSNR↑	26.87	27.29	25.79	27.72	20.82	27.71	28.31	28.25	28.15	23.90	**29.62**
	SSIM↑	0.6702	0.7104	0.6817	0.7376	0.5806	0.7543	0.7804	0.7898	0.7740	0.6948	**0.7990**
	LPIPS↓	0.4339	0.3497	0.3656	0.3664	0.4532	0.2997	0.2960	0.2900	**0.2591**	0.3345	0.2898
	DISTS↓	0.2793	0.2386	0.2437	0.2446	0.2771	0.2153	0.2169	0.2206	**0.1926**	0.2377	0.2302
	MANIQA↑	0.5281	0.6378	0.6232	0.5204	0.251	0.5883	0.6161	0.5784	0.6226	0.5598	**0.6429**
	MUSIQ↑	53.91	65.79	64.51	57.70	67.97	66.83	66.11	65.60	62.83	60.98	**68.18**
	Q-Align↑	3.2002	3.4284	3.3482	2.9609	3.6537	3.6185	3.5836	3.6424	3.4652	3.3464	**3.7473**
	LIQE↑	3.8725	3.9620	3.9157	2.7986	4.0107	3.9491	4.1496	4.0861	3.5717	3.4685	**4.1957**
	Q-Insight↑	3.1085	3.5806	3.5530	3.2112	**3.6777**	3.4693	3.5460	3.5133	3.4223	3.3051	3.6419
	AFINE-ALL↓	46.27	26.28	15.41	50.63	14.45	15.86	18.57	22.97	20.17	24.58	**7.61**
GGT-GeneralMix	PSNR↑	26.03	26.57	24.40	25.52	21.83	25.16	26.13	26.41	25.95	24.17	**26.86**
	SSIM↑	0.6606	0.6984	0.6518	0.6512	0.5823	0.6813	0.7161	0.7362	0.7126	0.6572	**0.7435**
	LPIPS↓	0.3637	0.2800	0.3527	0.4418	0.4275	0.3114	0.2916	0.2732	0.3040	0.3382	**0.2684**
	DISTS↓	0.1945	0.1614	0.1864	0.2750	0.2159	0.1701	0.1756	0.1631	0.1693	0.1846	**0.1605**
	MANIQA↑	0.5677	0.6041	0.6068	0.4708	0.6087	0.5880	0.5934	0.5698	0.6195	0.5597	**0.6221**
	MUSIQ↑	59.36	63.13	68.49	49.79	70.83	69.27	65.60	67.79	66.37	62.18	**71.06**
	Q-Align↑	3.7762	3.8415	3.9934	2.8977	4.2334	3.8949	3.8818	3.9229	3.9480	3.9615	**4.3009**
	LIQE↑	3.1497	3.3761	3.9904	1.9933	3.8951	4.0377	3.7018	3.9262	3.6083	3.4264	**4.2130**
	Q-Insight↑	3.4435	3.6085	3.8388	2.9985	**3.9863**	3.6901	3.6655	3.7361	3.7062	3.5839	3.9289
	AFINE-ALL↓	39.66	32.82	18.46	60.49	24.57	17.30	15.16	24.84	12.40	26.32	**2.6303**
Dual-Q-Bench	PSNR↑	22.71	22.59	21.67	21.90	19.56	21.89	22.32	22.64	22.35	21.74	**23.21**
	SSIM↑	0.6225	0.6196	0.6152	0.5701	0.5594	0.6290	0.6380	0.6492	0.6507	0.6129	**0.6662**
	LPIPS↓	0.3046	0.2736	0.3157	0.4095	0.3191	0.2648	0.2722	0.2746	0.2862	0.3069	**0.2639**
	DISTS↓	0.1784	0.1695	0.1793	0.2824	0.1829	0.1663	0.1675	0.1662	0.1646	0.1788	**0.1590**
	MANIQA↑	0.6257	0.6590	0.6328	0.4830	0.6623	0.6307	0.6520	0.6250	0.6611	0.6061	**0.6663**
	MUSIQ↑	67.80	70.80	70.87	59.12	74.05	73.83	72.17	73.07	71.19	70.25	**74.22**
	Q-Align↑	4.2362	4.2789	4.2239	2.8755	4.5148	4.2297	4.3351	4.3144	4.3569	4.2257	**4.6683**
	LIQE↑	4.1563	4.3841	4.3776	2.3265	4.4996	4.6336	4.5401	4.6342	4.4962	4.3966	**4.8425**
	Q-Insight↑	3.8389	3.9413	3.9807	3.2694	**4.1875**	3.9661	4.0004	4.0005	3.9700	4.0608	4.1103
	AFINE-ALL↓	47.01	43.44	41.68	69.75	42.64	39.00	35.10	40.22	30.95	43.59	**12.62**

Table 2: Comparison of computational costs on GGT-GeneralMix test set.

Methods	SUPIR -s50	FaithDiff -s20	Dit4SR -s40	VOSR-1.4B -s25	LucidFlux -s15	TSDSR -s1	PiSA-SR -s1	TVT -s1	ODTSR -s1	StrSR -s1	**Ours** -s1
Inference Time (s) ↓	10.8	5.0	57.6	7.9	23.8	0.48	0.45	2.4	5.2	0.97	**0.27**
GPU Memory (G) ↓	52.46	11.10	22.06	6.91	35.0	6.99	**6.89**	14.62	40.83	27.69	17.63

User Study of Real-ISR Benchmarks. We conduct a user study to evaluate the quality of Dual-Q-Bench against existing Real-ISR benchmarks, including RealSR Cai et al. (2019), DRealSR Wei et al. (2020), GGT Kong et al. (2026), and ScreenSR Wu et al. (2026). We randomly sample 50 GT-LQ image pairs from each dataset for evaluation. In each trial, 17 participants are presented with five image pairs from different datasets and asked to select the best one based on GT quality (clarity, background de-focus, and realism), LQ authenticity, and GT–LQ alignment (color, brightness, and position). As shown in Tab. 6, Dual-Q-Bench achieves the highest user preference rate of 76%, substantially surpassing the second-best, GGT (10.7%), confirming the best quality.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Protocols. We use LSDIR Li et al. (2023) and the first $10,000$ images from FFHQ Kornblith et al. (2019) as the training sets. LQ images are synthesized using the degradation pipeline in Real-ESRGAN Wang et al. (2021). We evaluate competing methods on DRealSR Wei et al. (2020), GGT-GeneralMix Kong et al. (2026), RealSR Cai et al. (2019) and our constructed *Dual-Q-Bench*. To comprehensively evaluate all methods, we employ both full-reference (FR) and no-reference (NR) metrics. The PSNR, SSIM Wang et al. (2004), LPIPS Zhang et al. (2018) and DISTS are used for fidelity evaluation, while MANIQA Yang et al. (2022), MUSIQ Ke et al. (2021), Q-Insight Li et al. (2026), Q-Align Wu et al. (2024a), and LIQE Zhang et al. (2023) are computed for visual generation assessment. In addition, we adopt AFINE-ALL Chen et al. (2025a) to assess the consistency between full-reference and no-reference evaluation even the reference image is imperfect.

Implementation Details. LoRA Hu et al. (2022) layers with rank 64 are fine-tuned based on the open-source HiDream-O1-Image Cai et al. (2026) model. λ_1, λ_2, λ_3, λ_4, λ_{main}, λ_{wave} is set to 1, 1, $1e^{-1}$, 1, 1, $1e^{-1}$, respectively. The timestep t is fixed to 0.5 in training and inference stages. We train HiFiPix on 4 NVIDIA H20 GPUs for $4,000$ iterations with an accumulated batch size of 32. Following previous methods Fei et al. (2026); Wang et al. (2026a), the training patch size is set to 1024^2. The learning rate is set to $5e^{-5}$ with the AdamW Loshchilov & Hutter (2019) optimizer. We discard image captions and remove the text encoder to accelerate training and inference.

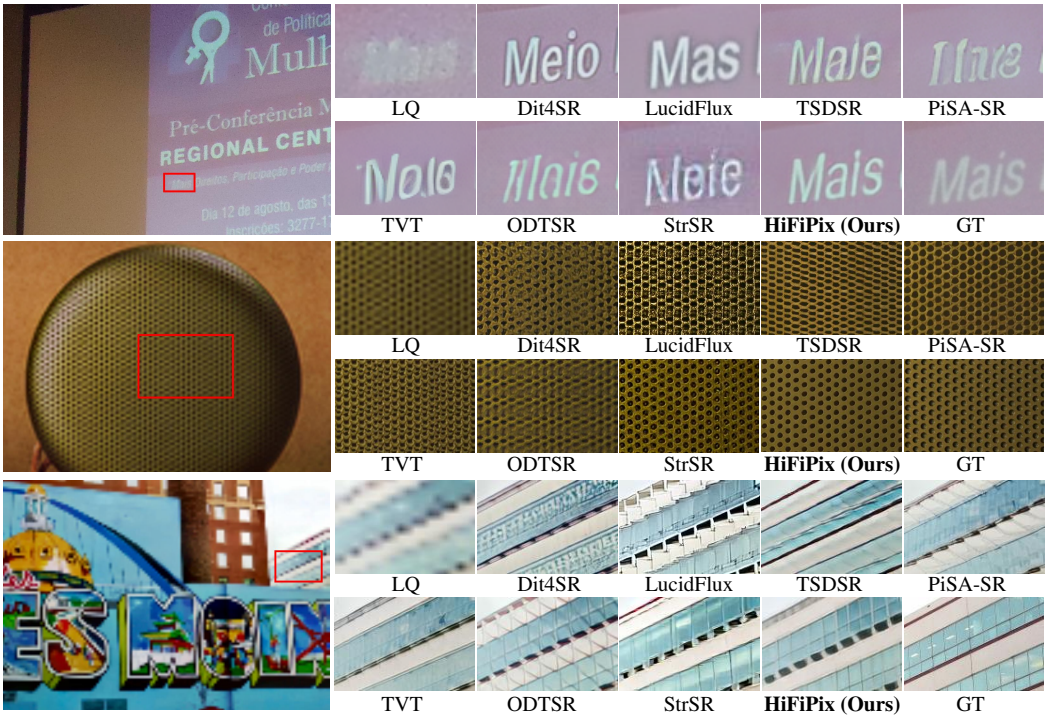


Figure 5: Qualitative comparisons with existing Real-ISR methods on real-world test sets.



Figure 6: Qualitative visualization of CSWD and MixCSD for patch-boundary artifact mitigation.

CSWD	MixCSD	PSNR↑	LPIPS↓	MUSIQ↑	MANIQA↑
✗	✗	25.53	0.3132	**70.67**	**0.6247**
✓	✗	25.83	0.2781	68.62	0.6048
✗	✓	25.77	0.2839	68.88	0.6084
✓	✓	**26.12**	**0.2759**	69.86	0.6183

Table 3: Ablation results of CSWD and MixCSD on quantitative metrics.

4.2 Experiment Results

We compare HiFiPix with existing state-of-the-art Real-ISR methods, including multi-steps methods (*i.e.*, SUPIR Yu et al. (2024), FaithDiff Chen et al. (2025b), Dit4SR Duan et al. (2025), VOSR Wu et al. (2026) in version of 1.4B Params, LucidFlux Fei et al. (2026)), and one-step methods (*i.e.*, TSDSR Dong et al. (2025), PiSA-SR Sun et al. (2025), TVT Yi et al. (2025), ODTSR Fang et al. (2026), and StrSR Wang et al. (2026a) based on Z-Image-Turbo Cai et al. (2025)).

Quantitative Results. As shown in Tab. 1, HiFiPix consistently outperforms existing methods in restoration fidelity, with notable gains on high-quality GT benchmarks (GGT and Dual-Q-Bench). It also excels on most NR metrics, confirming enhanced fidelity without sacrificing generation ability.

Qualitative Results. As shown in Fig. 5, HiFiPix produces both photo-realistic results and faithfully restored details, e.g., fine text in the first set and faithful textures in the second and third sets. By contrast, existing VAE-based methods generate entangled artifacts and over-smoothed details.

Computational Cost. As shown in Tab. 2, we compare the computational cost of different methods on GGT-GeneralMix at 1024^2 resolution on a single NVIDIA H20 GPU. Benefiting from the removal of the text encoder and the efficient vanilla patch-wise tokenization in HiDream-O1-Image that greatly reduces complexity, HiFiPix is much more efficient than multi-step diffusion methods and runs 40% faster than PiSA-SR, the second-best single-step method. Moreover, its GPU memory footprint is considerably smaller than ODTSR and StrSR, which also leverage large-scale pre-trained models. Due to page limitations, comparisons of parameters can be found in Tab. 10 of Appendix A.3.

Table 4: Ablation on the teacher network input composition: mixed LQ-HQ vs. HQ-only.

Method	PSNR↑	LPIPS↓	MUSIQ↑	MANIQA↑
HQ-only	**26.15**	**0.2726**	68.72	0.6079
LQ+HQ (**Ours**)	26.12	0.2759	**69.86**	**0.6183**

Table 5: Ablation on loss formulation: distillation loss ($\mathcal{L}_{\text{KD}}$) vs. contrastive learning ($\mathcal{L}_{\text{InfoNCE}}$).

Method	PSNR↑	LPIPS↓	MUSIQ↑	MANIQA↑
$\mathcal{L}_{\text{KD}}$	26.01	0.2794	69.51	0.6155
$\mathcal{L}_{\text{InfoNCE}}$(**Ours**)	**26.12**	**0.2759**	**69.86**	**0.6183**

Table 6: Results of user study, showing vote percentages for different Real-ISR benchmarks.

Benchmarks	Preferences↑
RealSR	1.3%
DrealSR	1.5%
GGT-GeneralMix	10.7%
ScreenSR	10.5%
Dual-Q-Bench (**Ours**)	76.0%

Table 7: Effectiveness of redundancy detection and fidelity-aware veto in LoRA-SP. Further effectiveness of LoRA-SP in VAE-based models.

Method	PSNR↑	LPIPS↓	MUSIQ↑	MANIQA↑
Hidream Baseline (Pixel-based)	**26.12**	**0.2759**	69.86	0.6183
+Random LoRA Pruning	25.27	0.2910	69.52	0.6124
+Redundancy detection (RD)	25.81	0.2846	**71.17**	**0.6303**
+RD & Fidelity-aware veto (**Ours**)	26.06	0.2775	70.92	0.6274
Z-Image (VAE-based)	**23.81**	**0.6327**	72.21	**0.6293**
Z-Image+LoRA-SP (VAE-based)	23.05	0.6205	**72.72**	0.6267

4.3 Ablation Study

CSWD & MixCSD. As shown in Fig. 6, both CSWD and MixCSD effectively suppress periodic boundary artifacts, and their combination yields the best visual results. From Tab. 3, both terms also substantially improve reconstruction fidelity (PSNR/LPIPS), yet using either alone degrades generation quality (MUSIQ/MANIQA). In contrast, combining them achieves the best fidelity without excessive loss of generation capability. CSWD and MixCSD penalize artifacts from image and intermediate feature domains, respectively, making them complementary.

GT-LQ Mixing Strategy. We conduct comparison experiments by replacing the GT-LQ mixture branch with GT-only inputs. Tab. 4 validates the benefits of mixed LQ-HQ teacher inputs. An HQ-only teacher creates a huge domain gap relative to LQ student inputs, degrading performance.

Contrastive Loss. We substitute contrastive loss with knowledge distillation loss Dong et al. (2025). As shown in Tab. 5, contrastive learning is superior to distillation by leveraging negative samples to avoid representation collapse and maintain feature diversity, thereby improving performance.

LoRA-SP. LoRA-SP identifies redundant LoRA layers for pruning and incorporates a fidelity-aware veto mechanism, which retains critical layers by postponing their pruning to later candidates. As shown in Tab. 7, redundancy detection alone boosts generation quality yet causes notable fidelity drops. Equipped with the fidelity-aware veto, LoRA-SP yields an improved generation-fidelity trade-off and consistently surpasses uniform and random LoRA pruning. These results validate our design and suggest an interpretable path to enhance pixel-space model generation via selective LoRA pruning. Moreover, Tab. 7 indicates that this strategy yields marginal gains for latent-space models (*e.g.*, Z-Image), probably because VAE-driven semantic compression produces little layer redundancy. Hence, LoRA-SP is tailored for boosting the generation performance of pixel-space models.

Due to page limitations, more ablation studies can be found in Appendix. A.4.

4.4 Limitations

HiFiPix suffers from two main limitations. First, while patch-boundary artifacts are mostly mitigated, several extremely blurred inputs still contain faint residual seams within texture-rich areas. Adding a generative decoder (*e.g.*, PiD Lu et al. (2026a)) could suppress these artifacts further, yet we abandon it for the additional consumption. Second, as our pipeline does not rely on text prompts, we exclude text-conditioned metrics such as CLIPIQA. These metrics focus mainly on text-image alignment, thereby our prompt-free setting yields inherently low scores that offer limited evaluation results.

5 Conclusion

We present HiFiPix, a pixel-space generative framework which enables faithful restoration by removing VAE compression. To mitigate patch-boundary artifacts and generative degradation, we propose CSWD, MixCSD for enhanced high-frequency reconstruction and cross-patch consistency, and LoRA-SP to generate fine-grained details with minimal fidelity sacrifice. We further construct Dual-Q-Bench, a high-quality benchmark for reliable evaluation. Experiments show HiFiPix outperforms VAE-based methods, demonstrating the promise of pixel-space modeling for Real-ISR.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability, grammar, and language of the manuscript, and to assist with the creation and modification of software code. Generative AI tools were not used to develop the methodology, formulate mathematical claims, design experiments, or analyze or interpret experimental results. All AI-assisted text and code were reviewed, revised, and verified by the authors. We take full responsibility for the final content of this work, including all text, code, claims, and conclusions.

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.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

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 *ICCV*, pp. 3086–3095, 2019.

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Tao Mei, Yingwei Pan, Yi Peng, Zhaofan Qiu, Ting Yao, Kai Yu, Yiheng Zhang, et al. Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer. *arXiv preprint arXiv:2605.11061*, 2026.

Du Chen, Jie Liang, Xindong Zhang, Ming Liu, Hui Zeng, and Lei Zhang. Human guided ground-truth generation for realistic image super-resolution. In *CVPR*, pp. 14082–14091. IEEE, 2023.

Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang. Toward generalized image quality assessment: Relaxing the perfect reference quality assumption. In *CVPR*, pp. 12742–12752, 2025a.

Junyang Chen, Jinshan Pan, and Jiangxin Dong. Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In *CVPR*, pp. 28188–28197, 2025b.

Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. Pixelflow: Pixel-space generative models with flow. *arXiv preprint arXiv:2504.07963*, 2025c.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, volume 34, pp. 8780–8794, 2021.

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.

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 *CVPR*, pp. 23174–23184, 2025.

Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S Ren, Chunle Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. In *ICCV*, pp. 18948–18958, 2025.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pp. 12873–12883, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

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 *CVPR*, pp. 23440–23450, 2026.

Song Fei, Tian Ye, Lujia Wang, and Lei Zhu. Lucidflux: Caption-free photo-realistic image restoration via a large-scale diffusion transformer. In *ICLR*, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, volume 1, pp. 3, 2022.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pp. 5967–5976. IEEE, 2017.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pp. 5148–5157, 2021.

Xiangtao Kong, Jixin Zhao, Lingchen Sun, Rongyuan Wu, and Lei Zhang. Ggt-100k: Generative ground truth for generalizable real-world image restoration. *arXiv preprint arXiv:2605.31039*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, pp. 3519–3529. PMlR, 2019.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *CVPR*, pp. 36115–36125, 2026.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Jian Zhang, et al. Q-insight: Understanding image quality via visual reinforcement learning. *Advances in Neural Information Processing Systems*, 38:36802–36827, 2026.

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 *CVPR*, 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 *ECCV*, pp. 430–448. Springer, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. 2019.

Yifan Lu, Qi Wu, Jay Zhangjie Wu, Zian Wang, Huan Ling, Sanja Fidler, and Xuanchi Ren. Pid: Fast and high-resolution latent decoding with pixel diffusion. *arXiv preprint arXiv:2605.23902*, 2026a.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026b.

Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. Deco: Frequency-decoupled pixel diffusion for end-to-end image generation. In *CVPR*, pp. 43600–43610, 2026a.

Zehong Ma, Ruihan Xu, and Shiliang Zhang. Pixelgen: Pixel diffusion beats latent diffusion with perceptual loss. *arXiv preprint arXiv:2602.02493*, 2026b.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 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, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

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 *CVPR*, pp. 2333–2343, 2025.

Jingkai Wang, Yixin Tang, Jue Gong, Jiatong Li, Shu Li, Libo Liu, Jianliang Lan, Yutong Liu, and Yulun Zhang. Spectral and trajectory regularization for diffusion transformer super-resolution. In *ECCV*, 2026a.

Shuai Wang, Ziteng Gao, Chenhui Zhu, Weilin Huang, and Limin Wang. Pixnerd: Pixel neural field diffusion. In *ICLR*, 2026b.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pp. 1905–1914, 2021.

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 *ECCV*, pp. 101–117. Springer, 2020.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *ICML*, 2024a.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *NeurIPS*, volume 37, pp. 92529–92553, 2024b.

Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, Xiangtao Kong, Jixin Zhao, Shihao Wang, and Lei Zhang. Vosr: A vision-only generative model for image super-resolution. In *CVPR*, pp. 16311–16321, 2026.

Zhiqiang Wu, Zhaomang Sun, Tong Zhou, Bingtao Fu, Ji Cong, Yitong Dong, Huaqi Zhang, Xuan Tang, Mingsong Chen, and Xian Wei. Omgsr: You only need one mid-timestep guidance for real-world image super-resolution. *arXiv preprint arXiv:2508.08227*, 2025.

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 *CVPR*, 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 *ECCV*, pp. 74–91. Springer, 2024.

Qiaosi Yi, Shuai Li, Rongyuan Wu, Lingchen Sun, Yuhui Wu, and Lei Zhang. Fine-structure preserved real-world image super-resolution via transfer vae training. In *ICCV*, pp. 12415–12426, 2025.

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 *CVPR*, pp. 25669–25680, 2024.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. Pixeldit: Pixel diffusion transformers for image generation. In *CVPR*, pp. 14273–14282, 2026.

Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *arXiv preprint arXiv:2409.17058*, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *CVPR*, pp. 14071–14081, 2023.

A Appendix

In this Appendix file, we provide the following materials:

- Construction of Dual-Q-Bench. Please refer to Sec. 3.6 in the main paper.
- More quantitative and qualitative comparisons. Please refer to Sec. 4.2 in the main paper.
- Model parameters of different Real-ISR methods. Please refer to Sec. 4.2 in the main paper.
- More ablation studies of our method. Please refer to Sec. 4.3 in the main paper.

A.1 Construction of Dual-Q-Bench

Existing paired Real-ISR benchmarks often suffer from low-quality or unrealistic GT, which can compromise the reliability of fidelity and generation evaluation. As shown in Fig. 7, existing benchmarks (*e.g.*, RealSR Cai et al. (2019) and DrealSR Wei et al. (2020)) contain GT with blur details or background de-focus. Such imperfections introduce an undesirable evaluation bias: outputs with richer and more realistic details may deviate from deficient GT and receive lower fidelity scores, while outputs that closely mimic the imperfect GT may be favored instead, finally leading to unreasonable measurements, which is also indicated in Chen et al. (2025a). Besides, background de-focus in DIV2K Agustsson & Timofte (2017) and ScreenSR Wu et al. (2026) complicates fidelity assessment, as recovering sharp structures in the background can be incorrectly penalized as deviation from the GT. Furthermore, large-model-generated GT (*e.g.*, GGT-GeneralMix Kong et al. (2026)) do not necessarily guarantee faithful details due to potential hallucinations. To address these limitations, we construct *Dual-Q-Bench*, a high-quality real-world paired benchmark with realistic and sharp GT images at 1024^2 resolution, enabling more reliable evaluation of both fidelity and generation quality.

GT Construction: We crawl high-resolution photographs from Unsplash and filter them using a combination of NR metrics (*e.g.*, MANIQA Yang et al. (2022), MUSIQ Ke et al. (2021), Q-Insight Li et al. (2026)), retaining only images whose aggregated score exceeds GT images in LSDIR Li et al. (2023) by an average of 30% to ensure superior intrinsic quality. Laplacian-gradient response is then used to crop texture-rich regions, followed by manual inspection to discard images with background de-focus or painterly appearance, yielding 100 GT images with sharp, rich details at 1024^2 resolution.

LQ Acquisition: We obtain real-world LQ images following the same screen photographing strategy in Wu et al. (2026). Specifically, the original GT images are shown on an ultra-high-definition screen in the Imaging Laboratory. We implement a timed routine to capture LQ from GT images displayed on-screen by using a mobile phone for real-world degradation simulation. The specific mobile phone model used for image capture will be disclosed upon acceptance of our paper.

GT-LQ Alignment Strategy: Each LQ is registered to its GT via SIFT keypoint matching with RANSAC homography estimation, falling back to multi-scale template matching when correspondences are insufficient, and further refined by ECC for sub-pixel accuracy. The warped LQ then undergoes locally-adaptive brightness and color correction in LAB space to compensate for capture-induced illumination shifts, after which both images are center-cropped to 1024^2 to form the final aligned pair. Tab. 8 shows the differences between Dual-Q-Bench and existing benchmarks.

A.2 Quantitative and Qualitative Comparisons

Quantitative Comparisons. As shown in Tab. 9, on existing test sets, HiFiPix achieves competitive pixel-level restoration fidelity and obtains the best MANIQA Yang et al. (2022), Q-Align Wu et al. (2024a), and AFINE-ALL Chen et al. (2025a) scores. Its relatively lower LPIPS Zhang et al. (2018) and DISTS Ding et al. (2020) scores stem partly from limited perceptual quality within GT images,

Table 8: Comparisons between existing paired Real-ISR benchmarks and our Dual-Q-Bench.

Benchmark	# of Pairs	Resolution	MANIQA↑ of GT	MUSIQ↑ of GT	Real LQ	Real GT	Construction Manner
DIV2K	100	512^2	0.6274	63.7	✗	✓	Synthesized LQ
RealSR	100	512^2	0.6017	57.5	✓	✓	DSLR acquisition
DrealSR	100	512^2	0.5589	50.1	✓	✓	DSLR acquisition
ScreenSR	130	512^2	0.7187	72.2	✓	✓	Screen photography
GGT-GeneralMix	250	1024^2	0.6488	68.3	✓	✗	Generated GT
Dual-Q-Bench (Ours)	100	1024^2	**0.7316**	**74.9**	✓	✓	Screen photography

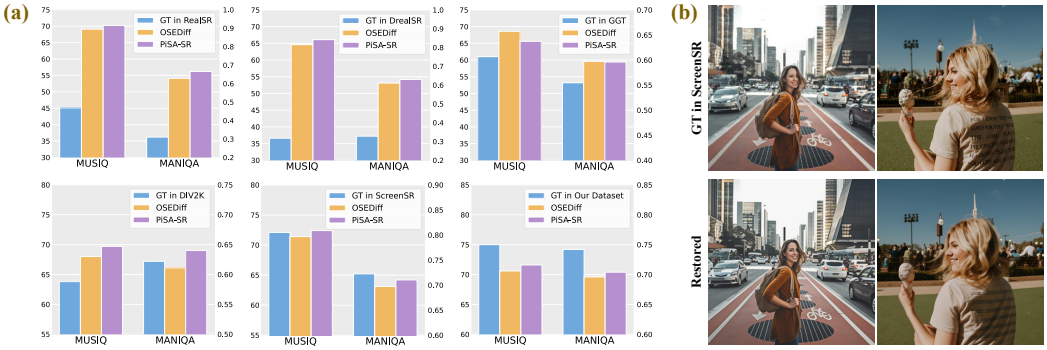


Figure 7: **(a) Insufficient GT Quality.** Existing Real-ISR benchmarks often provide GT with limited sharpness. Consequently, restored results may contain substantially richer or sharper details than GT, achieving higher MUSIQ and MANIQA scores than GT. This mismatch makes such GTs unreliable for jointly evaluating restoration fidelity and visual generation effects. **(b) Background De-focus in GT.** Existing benchmarks, such as ScreenSR, contain GT images with de-focused backgrounds. Real-ISR models often recover details in these regions that are absent from the GT, increasing the discrepancy with the reference and consequently challenging fidelity evaluation.

which lack fine details that HiFiPix faithfully reconstructs, rendering these FR metrics unreliable. Overall, HiFiPix improves fidelity without degrading generative performance.

Qualitative Comparisons. As shown in Fig. 8, Fig. 9, Fig. 10, and Fig. 11, HiFiPix achieves the best visual quality while preserving high-fidelity image details on various real-world test sets (*e.g.*, RealSR, DrealSR, GGT-GeneralMix, RealLQ250, and our proposed Dual-Q-Bench), including legible text and natural textures. In contrast, VAE-based approaches tend to introduce twisted artifacts and over-smoothing details. Those visual comparisons validate the superior performance of HiFiPix.

A.3 Comparison of Model Parameters

We compare generative backbone, language model, and trainable component parameters across existing Real-ISR methods in Tab. 10. Our generative backbone is substantially more compact than methods built on large generative architectures: ODTSR and LucidFlux use 22B Qwen and 12B FLUX-dev backbones, respectively, incurring heavy parameter overhead. Furthermore, removing the text encoder cuts our total parameters relative to SDXL-, SD3-, SD3.5-, and Z-Image-based approaches such as SUPIR, FaithDiff, DiT4SR, and StrSR. Our method requires merely 207.1M trainable parameters, surpassing all multi-step methods and delivering better parameter efficiency than single-step methods relying on large-scale generative backbones (e.g., ODTSR and StrSR). While SD2.1-based single-step methods have fewer trainable parameters, they yield poorer restoration performance. Together with our competitive restoration quality, these results verify our method's superior efficiency in both model size and fine-tuning cost.

A.4 Ablation Studies

Number of Selected Layers in LoRA-SP. With 35 LoRA layers total for fine-tuning, Tab. 11 examines how many layers should be pruned to boost generative capability while preserving fidelity. We find that both insufficient and excessive pruning yield suboptimal results. Specifically, pruning

Table 9: Quantitative comparisons with existing Real-ISR methods on real-world benchmarks, with the best/second-best results in **bold**/underlined. "-s" denotes the number of diffusion steps.

Datasets	Metrics	SUPIR -s50	FaithDiff -s20	Dit4SR -s40	VOSR -s25	LucidFlux -s15	TSDSR -s1	PiSA-SR -s1	TVT -s1	ODTSR -s1	StrSR -s1	**Ours** -s1
RealSR	PSNR↑	25.09	25.28	23.39	25.29	18.58	24.80	25.50	25.81	25.07	21.66	**26.70**
	SSIM↑	0.6728	0.7073	0.6661	0.7150	0.5544	0.7159	0.7418	**0.7597**	0.7361	0.6531	0.7553
	LPIPS↓	0.3731	0.2872	0.3193	0.2961	0.3856	0.2997	0.2672	0.2593	**0.2395**	0.3126	0.2622
	DISTS↓	0.2515	0.2110	0.2234	0.2226	0.2658	0.2153	0.2044	0.2061	**0.1892**	0.2358	0.2191
	MANIQA↑	0.5632	0.6737	0.6569	0.6510	0.6645	0.6342	0.6551	0.6225	0.6620	0.5967	**0.6739**
	MUSIQ↑	57.83	68.91	67.72	70.47	**71.40**	71.23	70.15	69.86	68.29	66.15	70.78
	Q-Align↑	3.3511	3.5679	3.3780	3.2089	3.8472	3.7364	3.6345	3.7679	3.6065	3.6615	**3.9069**
	LIQE↑	2.9247	3.9748	3.9351	3.3794	**4.2172**	4.0649	4.0982	4.1567	3.9295	3.6873	4.2034
	AFINE-ALL↓	46.55	20.86	15.38	30.68	10.84	17.45	16.87	23.79	15.07	28.37	**6.32**
RealLQ250	MANIQA↑	0.6547	0.6877	0.6832	0.6811	0.6832	0.6726	0.6912	0.6317	**0.7103**	0.6348	0.6997
	MUSIQ↑	69.55	71.61	72.00	71.52	73.80	**74.16**	73.13	69.37	73.16	71.35	73.82
	Q-Align↑	4.3184	4.0015	3.9845	3.8657	4.2847	4.1936	4.1849	4.0287	4.1307	4.0631	**4.4402**
	LIQE↑	4.3553	4.4080	4.4912	4.2704	4.6271	4.6227	4.6213	4.2508	4.5284	4.2024	**4.7264**

Table 10: Comparison of model parameters, where the generative model includes the text encoder, the VAE encoder/decoder, and the diffusion backbones; a language model is utilized for generating image captions. Since HiFiPix does not require image captions or text encoders, the final model parameters are fewer than the original $9B$ HiDream-O1-Image (with text encoder).

Methods	SUPIR -s50	FaithDiff -s20	Dit4SR -s40	VOSR -s25	LucidFlux -s15	TSDSR -s1	PiSA-SR -s1	TVT -s1	ODTSR -s1	StrSR -s1	**Ours** -s1
Generative Model params (B)↓	4.80	3.51	8.38	1.74	14.03	7.79	1.29	1.26	21.86	6.24	7.77
Language Model Params (B)↓	13.35	13.35	13.35	0	0	0	0	0.47	8.29	4.44	0
Trainable Parts Params (M)↓	1243.4	2612.2	350.9	1439	1630.8	95.2	8.1	30.07	1231.2	641.6	207.1

Table 11: Ablation on the number of selected layers in our LoRA-SP.

Layers	PSNR↑	LPIPS↓	MUSIQ↑	MANIQA↑
1 layer	26,11	0.2764	70.15	0.6207
2 layers	26.09	0.2768	70.51	0.6236
4 layers (Ours)	26.06	0.2775	70.92	0.6274
6 layers	25.92	0.2798	71.07	0.6282
8 layers	25.63	0.2856	71.22	0.6298

Table 12: Ablation on the number of distillation layers in MixCSD.

Layers	PSNR↑	LPIPS↓	MUSIQ↑	MANIQA↑
3 layers	25.90	0.2804	69.37	0.6138
4 layers	26.02	0.2766	69.69	0.6168
5 layers (Ours)	26.12	**0.2759**	**69.86**	**0.6183**
6 layers	26.16	0.2761	69.61	0.6152
7 layers	**26.21**	0.2767	69.15	0.6120

too few layers has only a limited effect on improving generative performance. While pruning more layers can improve generative performance, it severely compromises the model's fidelity. Pruning 4 layers of all LoRA layers achieves the optimal trade-off between fidelity and generative capability.

Number of Distillation Layers in MixCSD. Tab. 12 analyzes the effect of intermediate layer quantity for MixCSD-based contrastive distillation. We find that insufficient distillation layers provide limited supervision, hindering the generation of high-frequency representations and thus impeding performance improvements. Conversely, distilling excessive layers imposes redundant constraints on intermediate features and reduces generative flexibility. Accordingly, we uniformly adopt 5 layers (roughly $1/7$ of total layers) for contrastive distillation in final settings.

Layer-Selective Freezing vs. Layer-Adaptive Attenuation. We initially expected that layer-adaptive attenuation of LoRA across all layers would improve generative quality, as weakening the adapters should release the generative prior of the pretrained backbone. However, no gain was observed in practice: such attenuation only rescales the magnitude of each layer's update while preserving its direction, so the adaptation drift is weakened but never removed, and every layer is left in an intermediate state that matches neither the pretrained backbone nor the converged adapter. We therefore adopt a more decisive layer-selective freezing of LoRA instead.

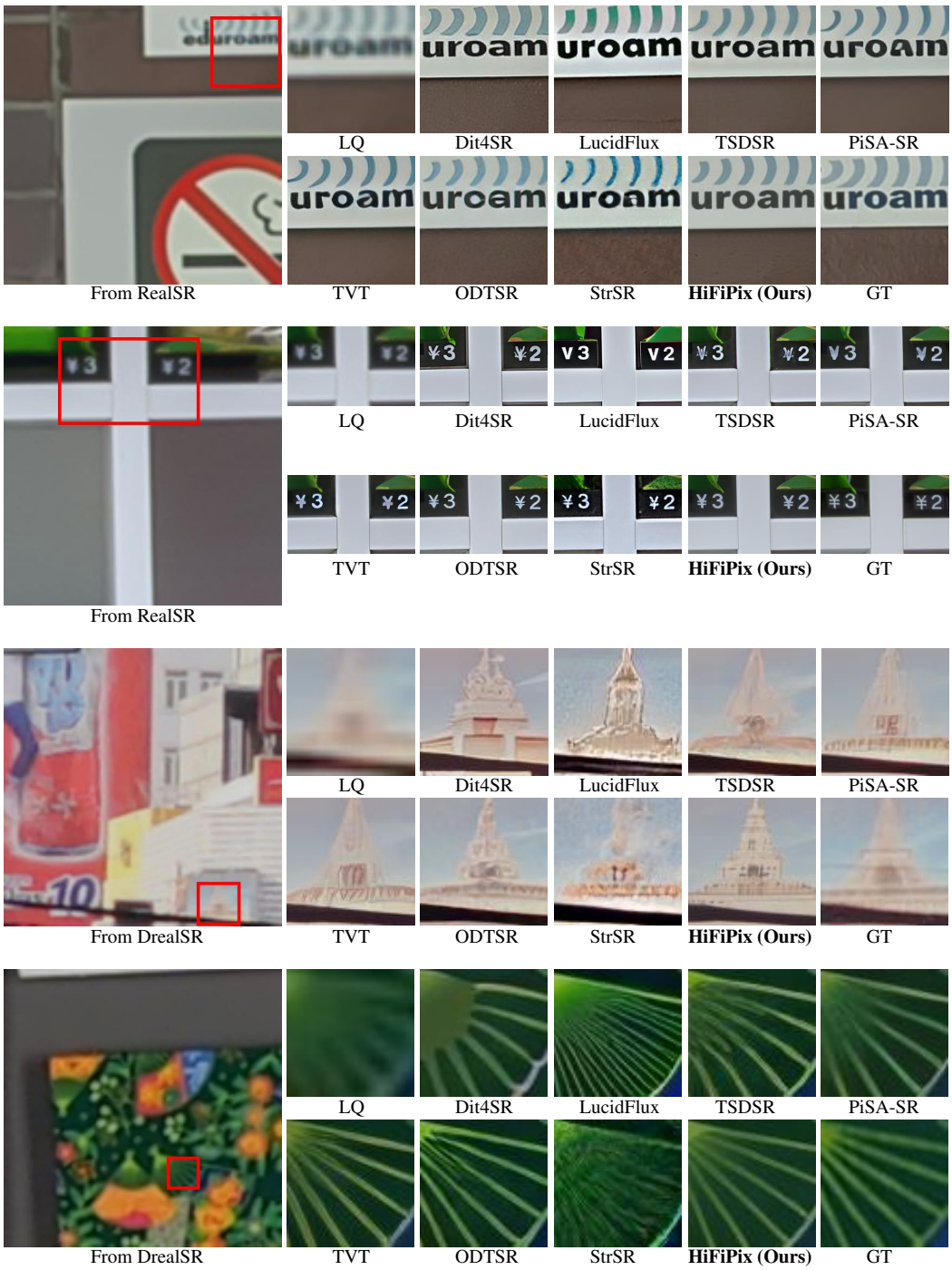


Figure 8: Qualitative comparisons with existing Real-ISR methods on paired LQ-GT real-world test sets: RealSR and DrealSR.

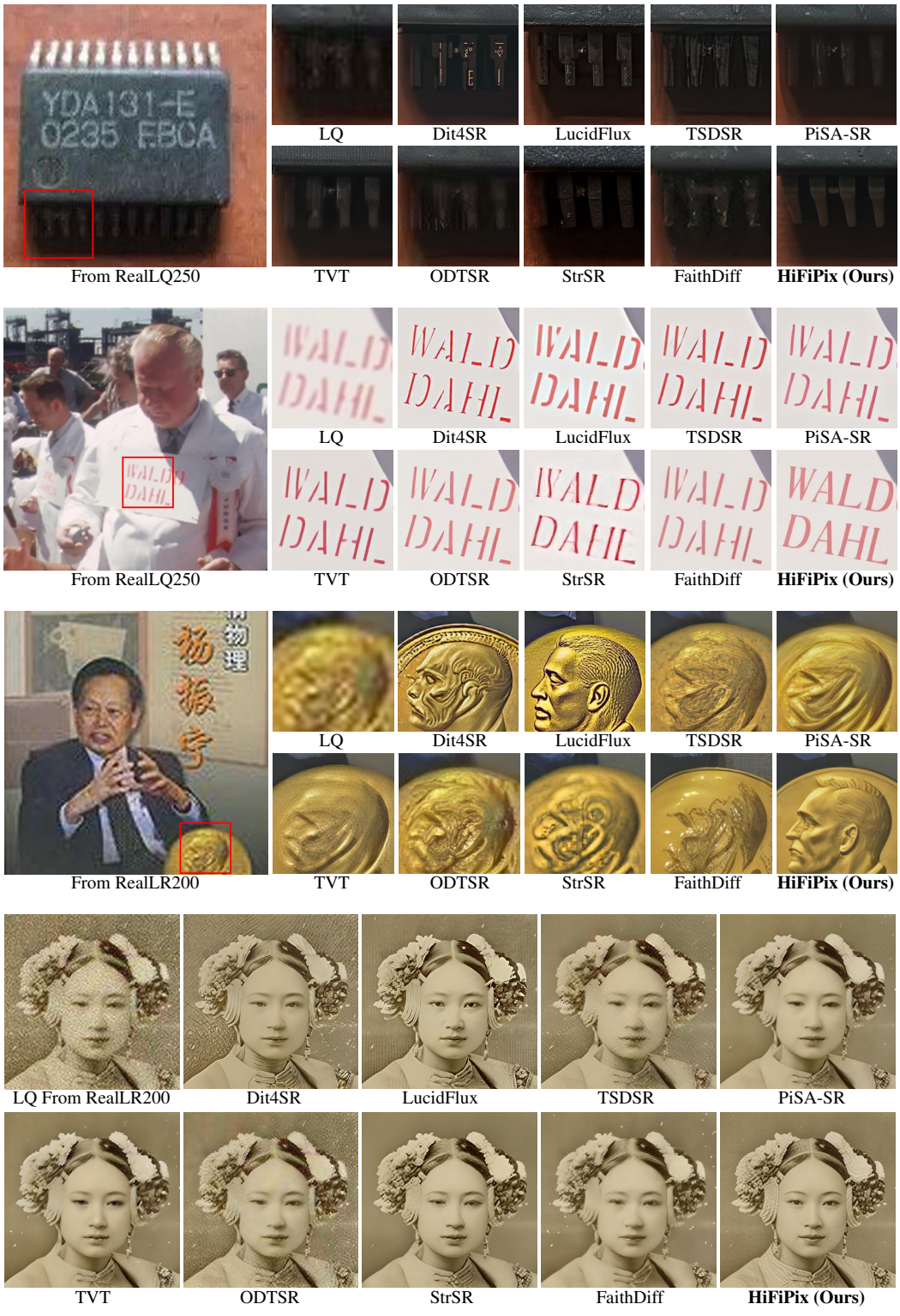


Figure 9: Qualitative comparisons with existing methods on unpaired RealLQ250 and RealLR200 test sets.



Figure 10: Qualitative comparisons with existing Real-ISR methods on paired LQ-GT real-world test sets: GGT-GeneralMix.

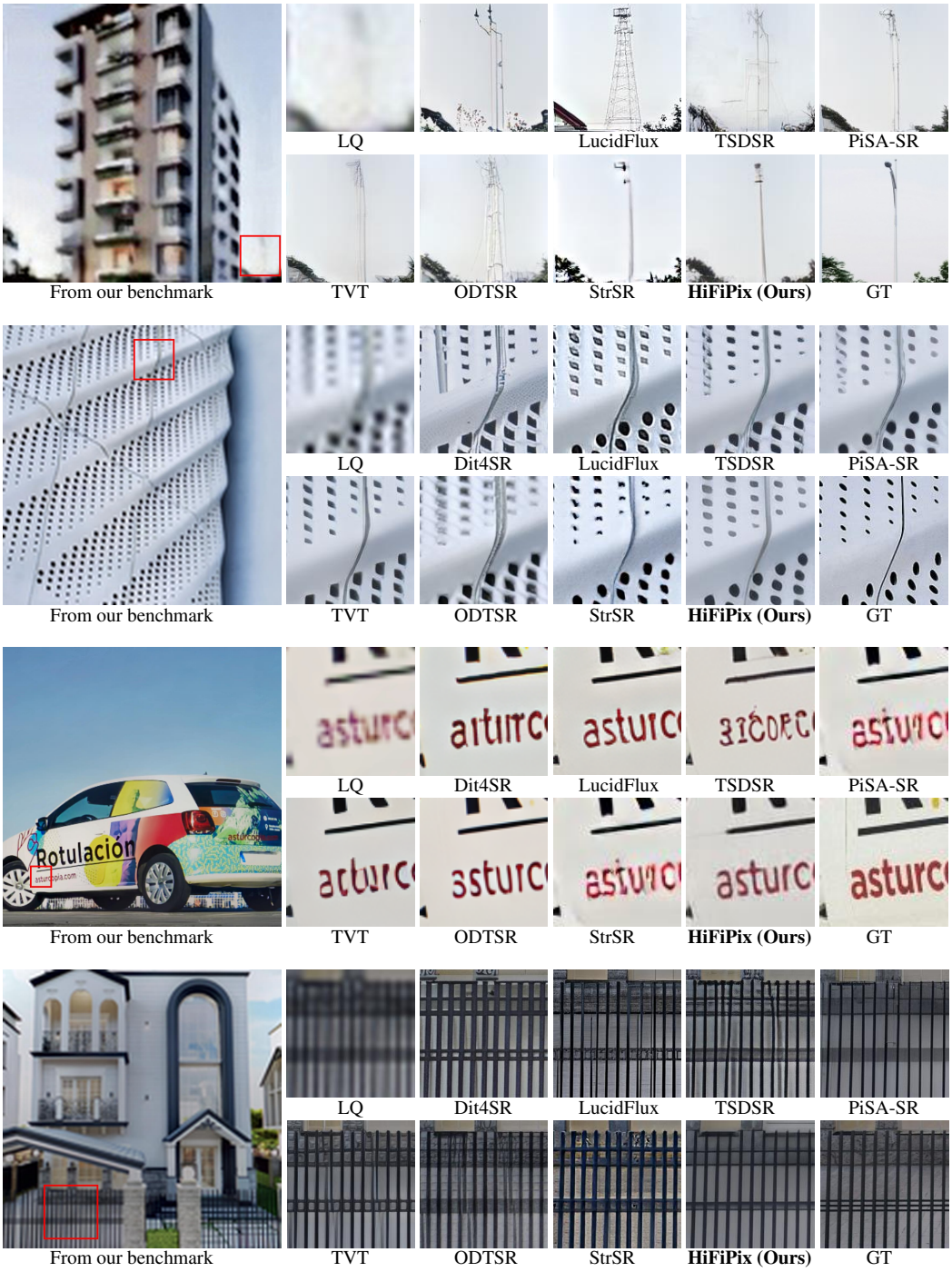


Figure 11: Qualitative comparisons with existing Real-ISR methods on our paired LQ-GT real-world benchmark: Dual-Q-Bench.