DYNAORTHO-SR: ONLINE ORTHOGONAL CO-EVOLUTION FOR REAL-WORLD IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world image super-resolution based on pre-trained diffusion models requires balancing faithful restoration with natural detail synthesis. Spatially non-uniform degradations demand region-adaptive restoration, while continued optimization of the pixel restoration branch can cause a fixed subspace constraint to become misaligned with its evolving representation. We propose DynaOrtho-SR, a one-step diffusion framework that integrates local degradation-aware expert modeling with online orthogonal co-evolution (OOCE). A local degradation-aware convolutional mixture-of-experts (C-MoE) jointly exploits intermediate features and local degradation descriptors for spatially adaptive routing. During alternating optimization of pixel experts and semantic adapters, we periodically update the pixel restoration subspace and constrain semantic updates through gradient projection and adapter-weight retraction. Since parameter-space orthogonality does not necessarily ensure functional complementarity, we further introduce functional orthogonality regularization (F-Ortho) on shared probes to suppress correlations between branch responses, providing complementary constraints in parameter and function spaces. At inference time, adjusting branch contributions yields different trade-offs between fidelity and perceptual quality without retraining or additional iterative diffusion sampling steps. Experiments on synthetic and real-world degradation benchmarks demonstrate competitive reconstruction performance, with a favorable balance between reconstruction fidelity and reference-based perceptual quality on real-world benchmarks.

1 INTRODUCTION

Image super-resolution (ISR) aims to recover high-resolution (HR) images from low-resolution (LR) observations (Anwar et al., 2020). Unknown blur, noise, and compression make real-world ISR ill-posed, requiring input-consistent reconstruction and plausible detail synthesis (Zhang et al., 2021; Wang et al., 2021b; Blau & Michaeli, 2018). Pixel-wise regression methods (Dong et al., 2015; Lim et al., 2017; Liang et al., 2021) preserve fidelity but tend to oversmooth uncertain details. Perceptual losses compare pre-trained feature representations (Johnson et al., 2016), while generative adversarial networks (Goodfellow et al., 2014) exploit natural-image distributions to synthesize realistic textures (Ledig et al., 2017; Wang et al., 2018). However, structural distortions, artifacts, and training instability remain challenging under complex degradations, motivating richer generative priors for faithful and perceptually realistic restoration.

Recently, large-scale pre-trained diffusion models have demonstrated strong image generation capabilities (Ho et al., 2020; Rombach et al., 2022), opening new directions for real-world ISR. Existing methods transfer their generative priors to ISR through conditional feature injection, semantic guidance, and model adaptation. For example, StableSR (Wang et al., 2024a) and PASD (Yang et al., 2024) guide reconstruction using conditions derived from LR inputs; SeeSR (Wu et al., 2024) and XPSR (Qu et al., 2024) introduce semantic priors to improve the plausibility of generated details; and FaithDiff (Chen et al., 2025) strengthens the guidance provided by degraded inputs through feature alignment and joint optimization. Although these methods improve visual quality, multi-step iterative sampling incurs substantial computational overhead. To improve inference efficiency,

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.

methods such as SinSR (Wang et al., 2024b) distill restoration into a single step, advancing efficient diffusion-based ISR.

Building on these developments, recent work has explored explicit coordination between faithful restoration and perceptual enhancement. PiSA-SR (Sun et al., 2025) models pixel-level regression and semantic-level enhancement using two low-rank adaptation branches, enabling different restoration effects through decoupled training and inference-time adjustment of branch strengths. However, separating branch parameters does not necessarily eliminate overlap between their functional responses. FreqOrtho-SR (Hoang et al., 2026) further introduces frequency-guided mixture-of-experts, using degradation information to guide expert selection and restricting overlap between perceptual updates and the pixel restoration subspace. These studies provide an important foundation for balancing fidelity and perceptual quality, while raising the question of how to integrate local degradation adaptation with dynamic optimization constraints.

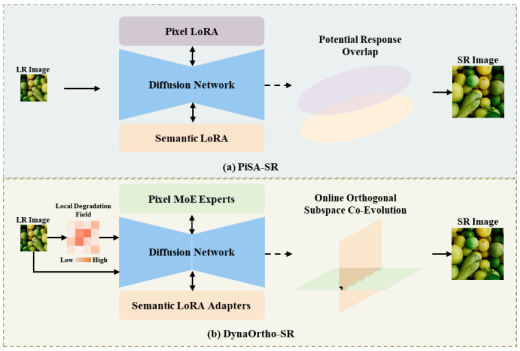

Figure 1: Conceptual comparison with PiSA-SR. DynaOrtho-SR combines local degradation-aware pixel experts with online subspace constraints and functional regularization to promote spatial adaptation and pixel-semantic complementarity.

To address these challenges, we propose DynaOrtho-SR, a one-step diffusion framework integrating local degradation-aware expert modeling with online orthogonal co-evolution (OOCE). As illustrated in Fig. 1, a convolutional mixture-of-experts (C-MoE) jointly exploits intermediate features and local degradation descriptors for spatially adaptive restoration. During alternating pixel and semantic optimization, OOCE refreshes the pixel restoration subspace and constrains semantic updates through gradient projection and adapter-weight retraction. Shared-probe functional orthogonality regularization (F-Ortho) further suppresses response redundancy, complementing the parameter-space constraint. Adjusting branch contributions enables inference-time fidelity-perception control without retraining or additional iterative diffusion sampling steps. Our main contributions are:

- We propose DynaOrtho-SR with a local degradation-aware C-MoE that jointly exploits content and degradation cues for spatially adaptive restoration.
- We introduce OOCE and shared-probe F-Ortho to coordinate evolving pixel and semantic branches through complementary parameter-space and functional constraints.
- Experiments on synthetic and real-world benchmarks demonstrate competitive restoration performance and inference-time fidelity-perception controllability.

2 Related Work

2.1 Diffusion-Based Super-Resolution

Diffusion models provide rich structural and textural priors for real-world ISR through condition injection and model adaptation. StableSR (Wang et al., 2024a) uses a time-aware encoder to guide a frozen diffusion backbone and controllable feature fusion to balance content fidelity and visual quality. PASD (Yang et al., 2024) introduces pixel-aware cross-attention to preserve local input structures, while FaithDiff (Chen et al., 2025) combines feature alignment with joint fine-tuning of the encoder and diffusion model. However, multi-step sampling remains computationally expensive, and generated details are not necessarily consistent with the true content. To reduce sampling costs, subsequent studies explore one-step generation (Song et al., 2023). SinSR (Wang et al., 2024b) distills a teacher's deterministic sampling mapping into a student with consistency-preserving supervision. CTMSR (You et al., 2025) learns a one-step mapping through consistency training and distribution trajectory matching; GTASR (Deng et al., 2026) combines trajectory alignment and structural correction; and VOSR (Wu et al., 2026) uses visual semantic conditions and distillation within a visual-only training framework. Although these methods improve efficiency, compressing

sampling does not directly address regional differences in degradation removal and detail synthesis or interference between fidelity and perceptual objectives. We therefore integrate local degradation-aware modeling with explicit co-optimization in a one-step diffusion framework.

2.2 Degradation-Aware Restoration

Degradation-aware restoration uses input degradation information to adjust feature extraction, sampling, or restoration parameters (Zhang et al., 2018a; Gu et al., 2019; Wang et al., 2021a). In diffusion-based ISR, PASD (Yang et al., 2024) extracts robust content features through degradation removal and guides local restoration with pixel-aware cross-attention. FaithDiff (Chen et al., 2025) combines feature alignment and joint optimization to reduce discrepancies between degraded inputs and generative priors. Both approaches improve conditional feature reliability rather than directly predicting degradation parameters. Region-adaptive noise control further exploits local uncertainty, injecting less noise into low-uncertainty regions to preserve input information (Zhang et al., 2025). However, improving conditional features or adjusting perturbation strength differs from assigning region-specific restoration experts. We jointly exploit intermediate features and local degradation descriptors for spatially adaptive routing, incorporating degradation awareness directly into the allocation of pixel restoration capacity.

2.3 Disentanglement and Orthogonal Optimization

Faithful reconstruction and perceptual enhancement impose different optimization requirements, which existing studies reconcile through inference-time control, branch decoupling, and gradient constraints. StableSR (Wang et al., 2024a) adjusts restoration effects through controllable feature fusion. PiSA-SR (Sun et al., 2025) further employs separate pixel and semantic LoRA (Hu et al., 2021) branches to learn pixel regression and perceptual enhancement, enabling controllable restoration through branch-strength adjustment. However, parameter separation does not directly guarantee complementarity between update directions or functional responses. Orthogonal optimization provides a complementary approach: orthogonal gradient descent in continual learning restricts update directions to reduce task interference (Farajtabar et al., 2020; Saha et al., 2021), while FreqOrtho-SR (Hoang et al., 2026) projects semantic gradients onto the orthogonal complement of the pixel restoration subspace to reduce subspace overlap. In contrast, we focus on dynamic coordination between pixel restoration and perceptual enhancement, combining online subspace constraints with functional response regularization to alleviate interference and redundancy in both parameter and function spaces.

3 Method

3.1 Preliminaries

Problem formulation. Real-world ISR recovers an HR image I_{HR} from a degraded observation $I_{\mathrm{LR}} = \mathcal{T}(I_{\mathrm{HR}}; \xi)$, where $\mathcal{T}$ includes downsampling, blur, noise, and compression under unknown conditions ξ. Given paired training images, we learn a mapping $\widehat{I}_{\mathrm{SR}} = \mathcal{G}_\Theta(I_{\mathrm{LR}})$ that preserves content while restoring natural details.

Latent representation. A pre-trained variational autoencoder (VAE) (Kingma & Welling, 2013; Rombach et al., 2022) encodes the input upsampled to the target size, $\widetilde{I}_{\mathrm{LR}}$, and its HR reference into equal-sized latent representations:

$$z_{\mathrm{LR}} = \mathcal{E}(\widetilde{I}_{\mathrm{LR}}), \qquad z_{\mathrm{HR}} = \mathcal{E}(I_{\mathrm{HR}}), \tag{1}$$

where $\mathcal{E}$ includes VAE encoding and latent scaling.

One-step residual restoration. Following latent residual modeling (Sun et al., 2025), we define $r^\star = z_{\mathrm{LR}} - z_{\mathrm{HR}}$ and adapt a pre-trained diffusion U-Net to predict the correction:

$$\widehat{r} = \mathcal{R}_\Theta(z_{\mathrm{LR}}, t_\star, c), \quad \widehat{z}_{\mathrm{SR}} = z_{\mathrm{LR}} - \widehat{r}, \quad \widehat{I}_{\mathrm{SR}} = \mathcal{D}(\widehat{z}_{\mathrm{SR}}), \tag{2}$$

where $t_\star$ is the fixed timestep, c is the backbone conditioning embedding, and Θ denotes model parameters. The mapping $\mathcal{D}$ includes inverse latent scaling and VAE decoding. This formulation restores images through latent residual correction without iterative diffusion sampling.

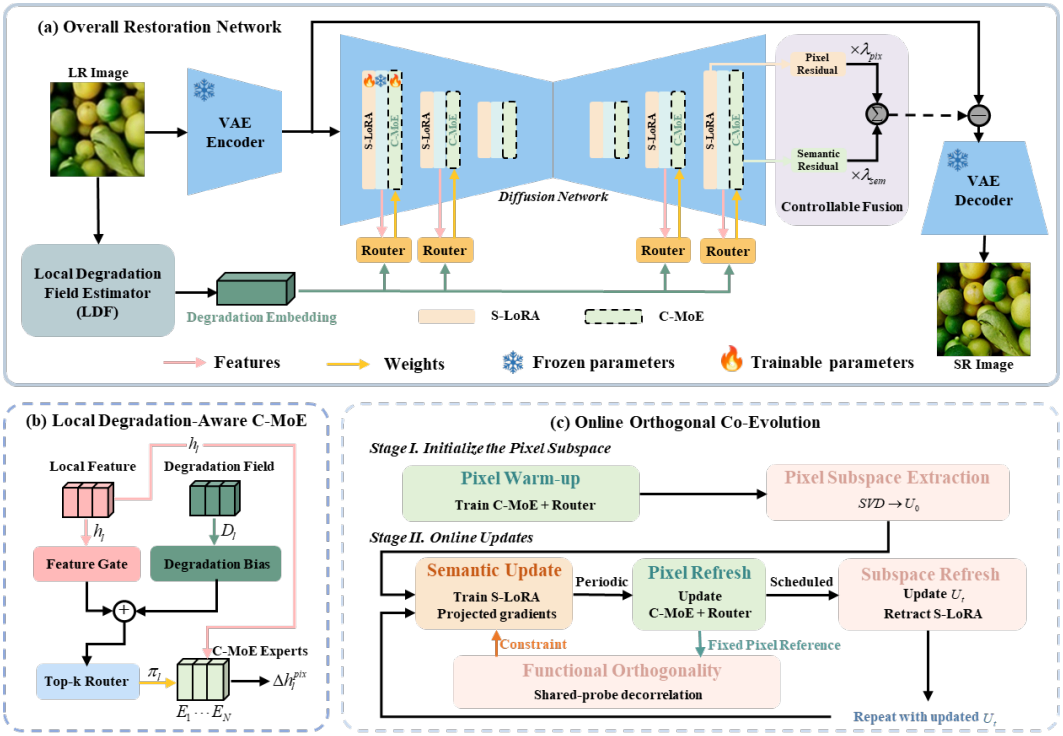

Figure 2: Overview of DynaOrtho-SR. (a) Pixel experts and semantic LoRA adapters predict controllable residuals for latent correction followed by VAE decoding. (b) Local degradation-aware C-MoE combines content gating and degradation bias for spatially adaptive Top-k routing. (c) After pixel warm-up, alternating optimization uses periodic subspace refresh, gradient projection, weight retraction, and shared-probe F-Ortho to coordinate pixel restoration and perceptual enhancement.

3.2 Framework Overview

As shown in Fig. 2, DynaOrtho-SR combines a local degradation-aware restoration network with online orthogonal co-evolution. With the diffusion backbone frozen, pixel experts and semantic LoRA adapters perform region-specific restoration and perceptual enhancement. The upsampled LR input is encoded by the frozen VAE, while its local degradation field conditions pixel-expert routing. Both adapter groups jointly predict a latent residual, which is subtracted from the input latent before VAE decoding. Their contributions can be adjusted at inference for different fidelity-perception trade-offs. Fig. 2(b) details C-MoE, where content gating and degradation bias determine Top-k routing for spatially adaptive expert aggregation. Fig. 2(c) shows the training procedure: pixel warm-up and initial subspace extraction are followed by alternating pixel and semantic updates. Periodic subspace refresh, gradient projection, and adapter-weight retraction constrain semantic optimization, while shared-probe F-Ortho suppresses response redundancy, providing complementary constraints in parameter and function spaces.

3.3 Local Degradation-Aware C-MoE

Local restoration depends on both content and degradation, whereas high-frequency content responses alone may not distinguish textures from artifacts. We construct a degradation field from spectral and gradient statistics on a regular grid of the upsampled input $\tilde{I}_{LR}$, with logarithmic compression and local-global fusion (Appendix A.1.2). As shown in Fig. 2(b), a content gate on layer-ℓ features h_ℓ and a degradation bias jointly produce expert scores:

$$\mathbf{a}_\ell(u) = \mathcal{G}_\ell^h(h_\ell)(u) + \mathcal{I}_\ell\left[\mathcal{G}_\ell^d(\mathbf{F}_{\text{deg}})\right](u), \tag{3}$$

Here, $\mathcal{U}$ denotes a spatial location, and $\mathcal{I}_\ell$ interpolates degradation scores to the feature resolution. Top-k selection followed by softmax normalization over selected experts (Shazeer et al., 2017), with zero weights for unselected experts, yields the spatially weighted pixel adaptation response:

$$\mathcal{M}_\ell(h_\ell, \mathbf{F}_{deg})(u) = \sum_{e=1}^{E} \pi_{\ell,e}(u) \mathcal{B}_{\ell,e}^{pix} \left(\mathcal{A}_{\ell,e}^{pix}(h_\ell) \right)(u), \tag{4}$$

Here, $A_{t,e}^{pix}$ and $B_{t,e}^{pix}$ form a low-rank convolutional expert, with scaling absorbed into its response. Content features feed both routers and experts, while the degradation field modulates expert preferences. This conditional aggregation (Yang et al., 2019; Chen et al., 2020) extends uniform adaptation to position-dependent restoration operators. Evolving expert weights and usage also change the effective restoration directions during training.

3.4 Online Orthogonal Co-Evolution

We alternate pixel-expert and semantic-adapter optimization to sustain restoration adaptation while constraining perceptual updates. For an adaptation layer, omitting its index, let $\Delta \mathbf{W}_e^{pix} = \mathbf{B}_e^{pix} \mathbf{A}_e^{pix}$ denote the effective expert weight, with convolutional weights unfolded along output channels. After warm-up, truncated SVD (Eckart & Young, 1936) of horizontally concatenated effective expert weights yields U_0. During online training, $col(\mathbf{B}_e^{pix} \mathbf{A}_e^{pix}) \subseteq col(\mathbf{B}_e^{pix})$ motivates a compact usage-weighted proxy:

$$\mathbf{M}_n = \left[\sqrt{\bar{q}_{1,n}} \mathbf{B}_{1,n}^{\mathrm{pix}}, ..., \sqrt{\bar{q}_{E,n}} \mathbf{B}_{E,n}^{\mathrm{pix}} \right], \quad \mathbf{U}_n = \mathrm{SVD}_\tau(\mathbf{M}_n), \tag{5}$$

Here, $\bar{q}_{e,n}$ is the exponential moving average of expert usage, SVD_τ truncates at a cumulative squared-singular-value ratio τ, and $\mathbf{P}_n = \mathbf{U}_n \mathbf{U}_n^\top$. The proxy retains dominant geometric directions of the output factors without repeatedly forming full weights; it is not the fidelity-loss gradient space of the full network. As the pixel subspace evolves, directions orthogonal under an old constraint can overlap with the current subspace, with the residual controlled by subspace drift (Appendix A.2.1). We therefore periodically refresh the reference basis and project the semantic LoRA output-factor gradient:

$$\tilde{\mathbf{G}}_n = (\mathbf{I} - \mathbf{P}_n) \nabla_{\mathbf{B}^{sem}} \mathcal{L}_{sem}. \tag{6}$$

Projection constrains subsequent increments but cannot remove components of existing weights in the new subspace. After each subspace refresh, we therefore retract the semantic output factor:

$$\mathbf{B}_{new}^{sem} = (\mathbf{I} - \mathbf{P}_n) \mathbf{B}_{old}^{sem} = \arg \min_{\mathbf{X}: \mathbf{U}_n^\top \mathbf{X} = 0} \| \mathbf{X} - \mathbf{B}_{old}^{sem} \|_F^2. \tag{7}$$

This operation restores feasibility with the minimum perturbation in Frobenius norm (Vandenberghe & Boyd, 2004). Retraction and gradient projection correct historical parameters and subsequent updates, respectively (Appendix A.2.2). The optimizer first moment is projected at refresh as well. Since Adam preconditioning need not preserve the constraint, feasibility is restored periodically rather than assuming strict orthogonality after every update.

3.5 Functional Orthogonality Regularization

Parameter orthogonality does not directly guarantee complementary responses after network propagation, and truncated subspaces cover only dominant directions (Appendix A.2.3). On selected layers Q, fixed shared probes $\{\nu_{\ell,m}\}_{m=1}^{M}$ are passed through the pixel experts and semantic adapter. We average pixel-expert responses and detach them, then flatten and center the pixel and semantic responses as $\bar{p}_{\ell,m}$ and $\bar{s}_{\ell,m}$, respectively, and define:

$$\mathcal{L}_{FO} = \frac{1}{\mid \mathcal{Q} \mid M} \sum_{\ell,m} \frac{\langle \bar{p}_{\ell,m}, \bar{s}_{\ell,m} \rangle^2}{\max(\|\bar{p}_{\ell,m}\|_2^2 \|\bar{s}_{\ell,m}\|_2^2, \epsilon)}. \tag{8}$$

For nonzero responses with an inactive stabilizer, each term measures the relative projection energy of the semantic response along the pixel reference, suppressing normalized directional overlap rather than overall response magnitude. Finite shared probes provide a low-cost local functional constraint complementary to the dynamic parameter constraint, without guaranteeing functional orthogonality of the full network.

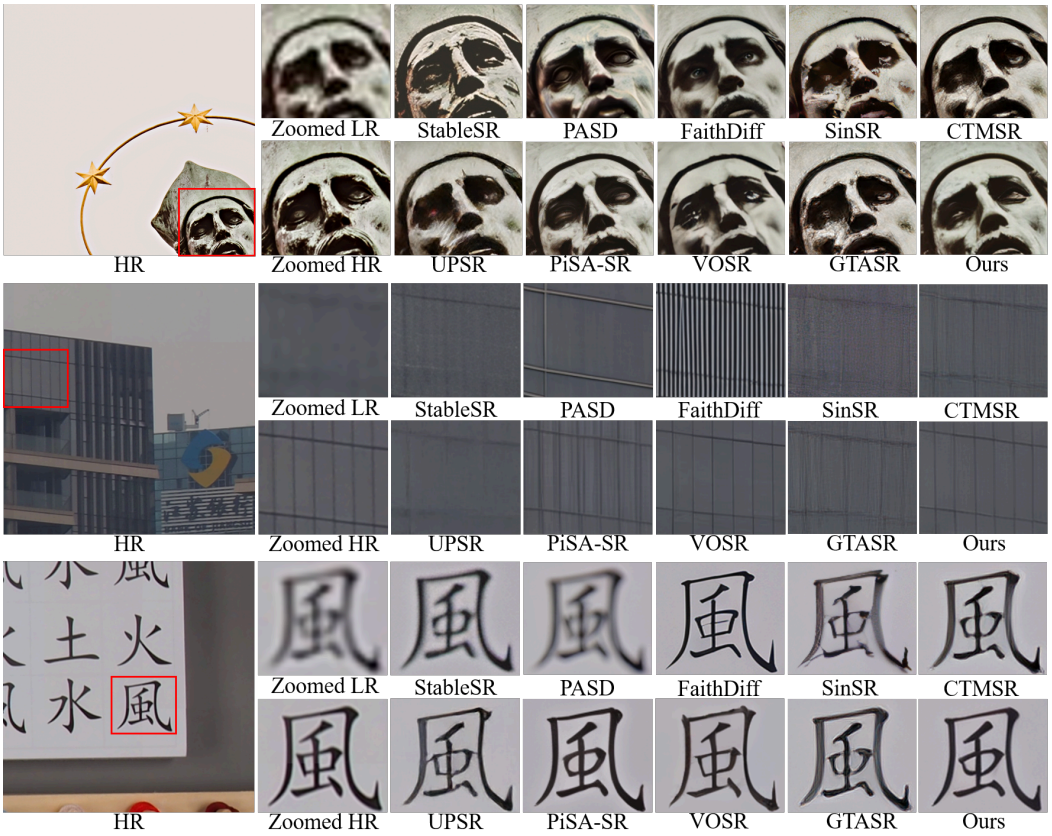

Figure 3: Visual comparisons of different Real-ISR methods on structure-sensitive regions. Please zoom in for a better view.

3.6 Training and Inference

Training objectives. As shown in Fig. 2(c), warm-up optimizes only pixel experts and routers and establishes the initial subspace. Online training periodically updates the pixel branch; other steps activate all adapters but update only semantic parameters. The objectives are:

$$\begin{aligned} \mathcal{L}_{pix} &= \alpha_{pix}\mathcal{L}_{MSE} + \alpha_{LB}\mathcal{L}_{LB}, \\ \mathcal{L}_{sem} &= \alpha_{sem}\mathcal{L}_{MSE} + \alpha_{LP}\mathcal{L}_{LPIPS} + \alpha_{CD}\mathcal{L}_{CD} + \alpha_{FO}\mathcal{L}_{FO}. \end{aligned} \tag{9}$$

MSE and LPIPS (Zhang et al., 2018b) compare restored and reference images; CSD (Yu et al., 2024) provides perceptual supervision from a frozen diffusion model. We use a routing load-balancing loss (Fedus et al., 2022), detailed in Appendix A.1.2. Pixel updates use pixel-only predictions, whereas semantic updates use full predictions. Loss weights and update intervals are provided in Appendix A.1, and the complete procedure in Appendix A.3. Controllable inference. Let $\hat{r}_{pix}$ and $\hat{r}_{all}$ denote residual predictions with pixel-only and all adapters active. The semantic contribution is their difference, combined using branch coefficients:

$$\hat{r}_{sem} = \hat{r}_{all} - \hat{r}_{pix}, \quad \hat{r} = \lambda_{pix}\hat{r}_{pix} + \lambda_{sem}\hat{r}_{sem}. \tag{10}$$

Latent correction and decoding are then performed. When both coefficients are set to 1, only a single full-model prediction is required; otherwise, the pixel-only and full predictions are computed separately, without retraining or additional iterative diffusion sampling steps.

Table 1: Quantitative comparison of DynaOrtho-SR with existing methods under the fidelity-oriented setting on RealSR, DRealSR, and DIV2K-Val. The best and second-best results are highlighted in bold and underlined, respectively.

Datasets	Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓
DRealSR	StableSR	28.8538	0.7629	0.3351	0.2307	<u>65.2159</u>
	PASD	<u>30.2606</u>	<u>0.8187</u>	0.3105	0.2207	65.2572
	FaithDiff	28.1783	0.7267	0.3703	0.2488	70.3612
	SinSR	29.3134	0.7635	0.3732	0.2494	82.8683
	CTMSR	29.7052	0.7980	0.3237	0.2387	76.1516
	UPSR	30.1253	0.8184	<u>0.2834</u>	<u>0.2048</u>	75.1472
	PiSA-SR	29.1291	0.7937	0.3033	0.2236	65.2861
	VOSR	28.5584	0.7588	0.3549	0.2409	76.7483
	GTASR	29.6172	0.7919	0.3342	0.2613	79.0682
	Ours ($\lambda_{\text{sem}} = 1.0$)	**30.2993**	**0.8227**	**0.2713**	**0.2011**	**56.9333**
RealSR	StableSR	24.8207	0.7204	0.3008	0.2042	84.4364
	PASD	<u>26.7446</u>	<u>0.7687</u>	0.2815	<u>0.1938</u>	81.2981
	FaithDiff	25.3066	0.7167	0.2884	0.2053	**73.8834**
	SinSR	26.5877	0.7546	0.3125	0.2286	101.5325
	CTMSR	26.2160	0.7686	0.2887	0.2114	100.1637
	UPSR	26.2335	0.7611	0.2815	0.1968	97.4483
	PiSA-SR	25.7448	0.7566	<u>0.2648</u>	0.1952	85.5313
	VOSR	25.3045	0.7307	0.2844	0.1994	80.7002
	GTASR	26.1174	0.7615	0.3016	0.2348	99.4343
	Ours ($\lambda_{\text{sem}} = 1.0$)	**26.8333**	**0.7737**	**0.2412**	**0.1772**	<u>78.6160</u>
DIV2K-Val	StableSR	21.6015	0.5545	0.3111	0.2058	66.6114
	PASD	22.4057	0.5900	0.4000	0.2301	77.0936
	FaithDiff	21.6626	0.5589	0.3121	0.2035	65.8863
	SinSR	22.2887	0.5681	0.3402	0.2142	81.5854
	CTMSR	**22.8578**	<u>0.5966</u>	0.3163	0.2086	78.6887
	UPSR	<u>22.6121</u>	0.5915	0.3333	0.2222	87.8312
	PiSA-SR	22.1988	0.5838	**0.2850**	**0.1972**	<u>64.4425</u>
	VOSR	21.6296	0.5629	0.3135	<u>0.1998</u>	64.9062
	GTASR	22.3647	0.5853	0.3011	0.2108	73.9245
	Ours ($\lambda_{\text{sem}} = 1.0$)	22.5436	**0.5974**	<u>0.2980</u>	0.2007	**63.9852**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets and metrics. We train on LSDIR (Li et al., 2023) and the first 10,000 FFHQ images (Karras et al., 2019) using Real-ESRGAN degradations (Wang et al., 2021b), and evaluate $4\times$ SR on DIV2K-Val (Agustsson & Timofte, 2017), DRealSR (Wei et al., 2020), and RealSR (Cai et al., 2019). Metrics include PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018b), DISTS (Ding et al., 2020), FID (Heusel et al., 2017), CLIPIQA (Wang et al., 2023), MUSIQ (Ke et al., 2021), and MANIQA (Yang et al., 2022).

Compared methods. We compare with StableSR (Wang et al., 2024a), PASD (Yang et al., 2024), FaithDiff (Chen et al., 2025), SinSR (Wang et al., 2024b), CTMSR (You et al., 2025), UPSR (Zhang et al., 2025), PiSA-SR (Sun et al., 2025), VOSR (Wu et al., 2026), and GTASR (Deng et al., 2026).

Implementation details. We freeze the Stable Diffusion 2.1-base backbone and VAE (Rombach et al., 2022), training only pixel experts, routers, and semantic LoRA adapters. Both evaluation settings share model weights with $\lambda_{\text{pix}} = 1.0$: $\lambda_{\text{sem}} = 1.0$ is the default fidelity-oriented setting, and 1.2 is the perception-oriented setting. Detailed training configurations, degradation settings, and test cropping and pairing protocols are provided in Appendix A.1.

Table 2: Quantitative comparison of DynaOrtho-SR with existing methods under the perception-oriented setting on RealSR and DRealSR. The best and second-best results are highlighted in bold and underlined, respectively.

Datasets	Methods	LPIPS↓	DISTS↓	FID↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
DRealSR	StableSR	0.3351	0.2307	<u>65.2159</u>	0.6096	57.0808	0.5517
	PASD	0.3105	0.2207	65.2572	0.5038	51.3888	0.4720
	SinSR	0.3732	0.2494	82.8683	0.6281	53.9569	0.4882
	CTMSR	0.3237	0.2387	76.1516	0.6466	58.4113	0.4833
	UPSR	**0.2834**	**0.2048**	75.1472	0.4802	54.8334	0.5098
	PiSA-SR	0.3033	0.2236	65.2861	<u>0.6853</u>	**64.3577**	<u>0.6063</u>
	Ours ($\lambda_{\text{sem}} = 1.2$)	<u>0.2970</u>	<u>0.2101</u>	**58.1852**	**0.6858**	<u>62.6833</u>	**0.6115**
RealSR	StableSR	0.3008	0.2042	84.4364	0.6279	66.0692	0.6197
	PASD	0.2815	<u>0.1938</u>	<u>81.2981</u>	0.4792	58.1208	0.5284
	SinSR	0.3125	0.2286	101.5325	0.6208	61.4910	0.5380
	CTMSR	0.2887	0.2114	100.1637	0.6403	65.7129	0.5157
	UPSR	0.2815	0.1968	97.4483	0.5233	66.1367	0.5535
	PiSA-SR	<u>0.2648</u>	0.1952	85.5313	<u>0.6644</u>	**70.9166**	<u>0.6534</u>
	Ours ($\lambda_{\text{sem}} = 1.2$)	**0.2611**	**0.1834**	**77.8680**	**0.6688**	<u>68.8955</u>	**0.6608**

4.2 Comparisons with State-of-the-Art Methods

Quantitative comparisons. Table 1 reports fidelity-oriented results on DRealSR, RealSR, and DIV2K-Val. DynaOrtho-SR achieves the best PSNR, SSIM, LPIPS, DISTS, and FID on DRealSR, and the best PSNR, SSIM, LPIPS, and DISTS with competitive FID on RealSR. On DIV2K-Val, our method achieves the best SSIM and FID and remains competitive in LPIPS. Under the perception-oriented setting (Table 2), DynaOrtho-SR ranks first in FID, CLIPIQA, and MANIQA and second in LPIPS, DISTS, and MUSIQ on DRealSR. On RealSR, it achieves the best LPIPS, DISTS, FID, CLIPIQA, and MANIQA, while PiSA-SR obtains higher MUSIQ. These results demonstrate competitive fidelity and perceptual quality under both inference settings.

Qualitative comparisons. Figure 3 compares structure-sensitive regions. Some methods exhibit oversmoothing, structural distortions, or high-frequency artifacts on sculpted faces, buildings, and text. DynaOrtho-SR better preserves facial contours, building lines, and character strokes. Texture-rich comparisons in Appendix A.7 further examine flowers, small fruits, and animal textures.

User preference. DynaOrtho-SR receives the highest preference across all five evaluated categories, with an overall preference rate of 27.6%. The study details and category-wise results are provided in Appendix A.4.

4.3 Ablation Study

To evaluate the core designs of DynaOrtho-SR, we individually remove C-MoE, OOCE, and F-Ortho and evaluate the resulting variants on DRealSR and RealSR. As shown in Table 3, the full model achieves the best overall performance. Removing C-MoE degrades multiple metrics, supporting the contribution of region-adaptive expert routing, while removing OOCE consistently reduces performance on both datasets. Removing F-Ortho causes smaller but consistent drops, supporting its complementary contribution beyond parameter-space constraints. Qualitative comparisons in Fig. 4 show blurred edges, discontinuous repetitive textures, and weakened details in some ablation variants, particularly on building lines and resolution test charts. The full model produces sharper edges and more continuous structures with fewer local distortions, supporting the combined effectiveness of local degradation-aware restoration, online orthogonal co-evolution, and functional response regularization.

4.4 Fidelity–Perception Trade-off Analysis

With $\lambda_{\text{pix}} = 1.0$, we vary λ_{sem} to assess inference-time controllability. As shown in Fig. 5, increasing semantic strength initially improves perceptual quality while reducing PSNR on all three

Table 3: Ablation study of DynaOrtho-SR on DRealSR and RealSR. The best results are highlighted in bold.

Settings	DRealSR				RealSR			
	PSNR↑	SSIM↑	DISTS↓	FID↓	PSNR↑	SSIM↑	DISTS↓	FID↓
Baseline	29.1291	0.7937	0.2236	65.2861	25.7448	0.7566	0.1952	85.5313
w/o C-MoE	30.0107	0.8200	**0.2011**	57.1349	26.6709	0.7709	0.1786	81.3471
w/o OOCE	30.0647	0.8166	0.2035	58.2473	26.5871	0.7669	0.1799	81.4573
w/o F-Ortho	30.2547	0.8208	0.2017	57.8653	26.7964	0.7727	0.1773	78.6223
DynaOrtho-SR	**30.2993**	**0.8227**	**0.2011**	**56.9333**	**26.8333**	**0.7737**	**0.1772**	**78.6160**

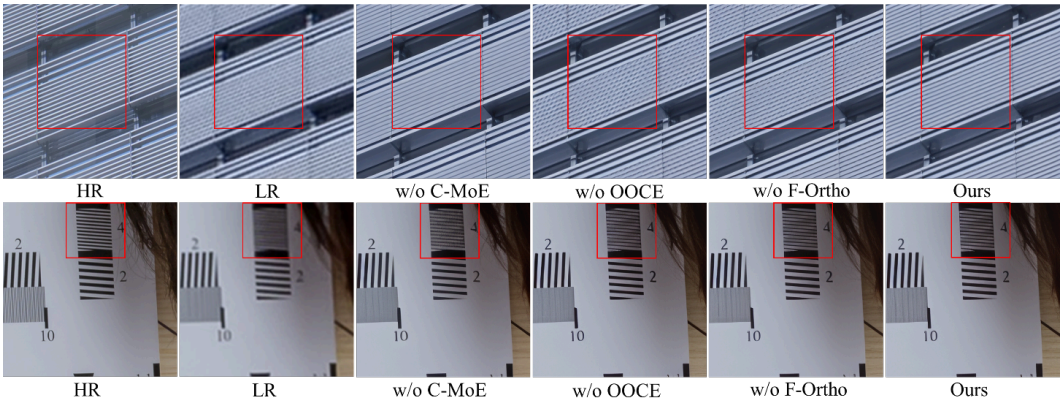

Figure 4: Qualitative ablation comparisons of DynaOrtho-SR. Please zoom in for a better view.

datasets; further increases yield saturating gains and eventually lower perceptual quality. A single model thus provides adjustable fidelity-perception operating points without retraining or additional iterative diffusion sampling steps.

5 Conclusion

We propose DynaOrtho-SR, a one-step diffusion-based real-world ISR model combining local degradation-aware C-MoE with dual-space OOCE. Spatially adaptive expert routing addresses regional degradation differences, while dynamic subspace constraints and shared-probe F-Ortho coordinate pixel restoration and perceptual enhancement. Experiments demonstrate competitive fidelity and perceptual quality on synthetic and real-world benchmarks, with adjustable inference-time operating points without retraining or additional iterative diffusion sampling steps. Future work will explore more efficient dynamic subspace modeling and cross-domain functional complementarity.

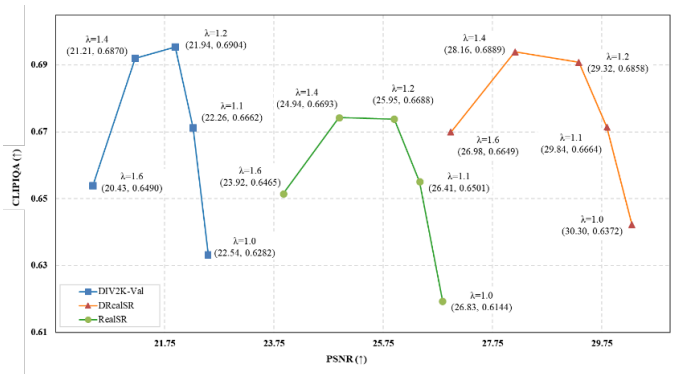

Figure 5: Fidelity-perception trade-offs obtained by varying λ_{sem} with $\lambda_{\text{pix}} = 1.0$.

AI USE STATEMENT

We used ChatGPT and DeepSeek to assist with translation and language editing, manuscript organization, literature and reference searches, methodological discussions, experimental design feedback, mathematical exposition and proof drafting, pseudocode preparation, figure layout, and code annotation. All AI-assisted outputs were reviewed and verified by the authors. The authors take responsibility for the final content of this work, including its claims, mathematical arguments, code, and reported results.

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.

Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *ACM computing surveys (CSUR)*, 53(3):1–34, 2020.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237. IEEE, 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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3086–3095. IEEE, 2019.

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

Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11027–11036. IEEE, 2020.

Chengyan Deng, Zhangquan Chen, Li Yu, Kai Zhang, Xue Zhou, and Wang Zhang. Joint geometric and trajectory consistency learning for one-step real-world super-resolution. *arXiv preprint arXiv:2602.24240*, 2026.

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.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2): 295–307, 2015.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics*, pp. 3762–3773. PMLR, 2020.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1604–1613. IEEE, 2019.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Minh Son Hoang, Dinh Phu Tran, Quyen Nguyen Duc, Dam Hoang Phuong, and Daeyoung Kim. Freqortho-sr: Frequency-guided orthogonal expert learning for real-world image super-resolution. In *European Conference on Computer Vision*, pp. 129–149. Springer, 2026.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

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

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

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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 *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 105–114. IEEE, 2017.

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

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *2021 IEEE/CVF international conference on computer vision workshops (ICCVW)*, pp. 1833–1844. IEEE, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 1132–1140. IEEE, 2017.

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

Yunpeng Qu, Kun Yuan, Kai Zhao, Qizhi Xie, Jinhua Hao, Ming Sun, and Chao Zhou. Xpsr: Cross-modal priors for diffusion-based image super-resolution. In *European Conference on Computer Vision*, pp. 285–303. Springer, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. *arXiv preprint arXiv:2103.09762*, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/song23a.html.

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

Lieven Vandenberghe and Stephen Boyd. *Convex optimization*, volume 1. Cambridge university press Cambridge, 2004.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

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

Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10576–10585. IEEE, 2021a.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *European conference on computer vision*, pp. 63–79. Springer, 2018.

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

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25796–25805. IEEE, 2024b.

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

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

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25456–25467. IEEE, 2024.

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. *arXiv preprint arXiv:2604.03225*, 2026.

Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in neural information processing systems*, 32, 2019.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1190–1199. IEEE, 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.

Weiyi You, Mingyang Zhang, Leheng Zhang, Xingyu Zhou, Kexuan Shi, and Shuhang Gu. Consistency trajectory matching for one-step generative super-resolution. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12747–12756. IEEE, 2025.

Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3d with classifier score distillation. In *International conference on learning representations*, volume 2024, pp. 19163–19177, 2024.

Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3262–3271. IEEE, 2018a.

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

Leheng Zhang, Weiyi You, Kexuan Shi, and Shuhang Gu. Uncertainty-guided perturbation for image super-resolution diffusion model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17980–17989. IEEE, 2025.

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

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

A Appendix

A.1 Implementation and Evaluation Details

A.1.1 Training and Degradation Settings

Data sampling and preprocessing. The training data comprise 84,991 LSDIR images (Li et al., 2023) and the first 10,000 FFHQ images (Karras et al., 2019). The data loader samples from LSDIR and FFHQ with probabilities of 0.1 and 0.9, respectively. During training, we randomly crop 512×512 image patches and apply random horizontal flipping. A two-stage Real-ESRGAN (Wang et al., 2021b) degradation pipeline generates LR observations with fourfold downsampling, which are then upsampled to the target size using randomly selected area, bilinear, or bicubic interpolation. Both inputs and supervision images are normalized to $[-1, 1]$. Color jitter, random grayscale conversion, and GT sharpening are not used.

Degradation settings. In the two degradation stages, the probabilities of upsampling, downsampling, and maintaining the size are $(0.2, 0.7, 0.1)$ and $(0.3, 0.4, 0.3)$, respectively, with scaling-factor ranges of $[0.3, 1.5]$ and $[0.6, 1.2]$. Each stage selects Gaussian or Poisson noise with equal probability. On the 8-bit image intensity scale, Gaussian noise standard deviations are sampled from $[1, 15]$ and $[1, 12]$, and Poisson noise scales from $[0.05, 2.0]$ and $[0.05, 1.0]$, respectively. The grayscale noise probability is 0.4 in both stages. Blur-kernel standard deviations range over $[0.2, 3.0]$ and $[0.2, 1.5]$, with a 0.5 probability of applying blur in the second stage. JPEG quality factors are sampled from $[60, 95]$ and $[60, 100]$, respectively. The sinc-kernel sampling probability is 0.1 in both stages, and the final sinc-filtering probability is 0.8. The order of the final JPEG compression and the combined resizing and sinc-filtering operations is randomized.

Optimization settings. We train on four NVIDIA RTX 4090 GPUs using FP16 mixed precision and gradient checkpointing. The per-GPU microbatch size is 1, with four gradient accumulation steps, yielding an effective batch size of 16. We use AdamW (Loshchilov & Hutter, 2017) with a constant learning rate of 5×10^{-5}, momentum parameters $(0.9, 0.999)$, weight decay of 0.01, numerical stability term 10^{-8}, gradient-norm clipping threshold of 1.0, and random seed 123. Training comprises 34,000 optimization steps, with the first 4,000 allocated to pixel warm-up. Pixel updates use image-space MSE and load-balancing losses (Fedus et al., 2022) with weights of 1.0 and 0.01, respectively. Semantic updates use MSE, LPIPS (Zhang et al., 2018b), CSD, and F-Ortho with weights of 1.0, 2.0, 1.0, and 0.01, respectively.

A.1.2 Model Implementation Details

Pre-trained backbone and conditioning. The restoration network is based on Stable Diffusion 2.1-base (Rombach et al., 2022). The pre-trained U-Net backbone, VAE, and text encoder are frozen, and only pixel experts, routers, and semantic adapters are optimized. Pixel and semantic adapters both have rank 4 and are inserted into selected convolutional and linear projection layers in the encoding, middle, and decoding stages of the U-Net. The restoration predictor uses a fixed timestep $t_\star = 1$. During training, the Recognize Anything Model (RAM) (Zhang et al., 2024) extracts tags from HR images, which are encoded as text conditions; the restoration predictor uses an empty text condition with probability 0.5. CSD (Yu et al., 2024) uses a frozen diffusion model with a guidance scale of 7.5, and noise timesteps are uniformly sampled from the $[0.02, 0.5)$ proportional interval of the total diffusion training timesteps. Text conditions enter through the backbone's original conditioning pathway; semantic LoRA does not receive separate text injection.

Local degradation descriptors. We construct a non-overlapping grid on the upsampled low-quality input and extract a seven-dimensional descriptor for each region: normalized mean magnitudes in four radial frequency bands, the low-to-high frequency magnitude ratio, the block-boundary gradient ratio, and the mean absolute Laplacian response. The final three statistics undergo $\log(1 + x)$ compression before the local and global descriptors are mixed with weights of 0.75 and 0.25. Descriptor extraction has no trainable parameters and is performed in FP32 for numerical stability in frequency-domain computations.

Expert structure and routing. Each pixel mixture-of-experts module contains four routed experts and no shared experts, with two experts selected at each position for response aggregation. The convolutional router uses a 1×1 convolution for content gating. The degradation-bias branch consists

of two 1×1 convolutions with channel dimensions $7 \to 14 \to 4$, a ReLU between them, and a zero-initialized final layer. Degradation biases are bilinearly resized to the current feature resolution and added to content-gating scores, followed by softmax normalization over the two selected experts. The down-projection of each convolutional expert uses the target layer's kernel size, stride, and padding; the up-projection uses a 1×1 convolution. The load-balancing term (Fedus et al., 2022) sums products of detached expert-assignment frequencies and mean pre-truncation routing probabilities at each routing layer, multiplies by the number of experts, and averages over routing layers.

Online subspace constraints. After pixel warm-up, SVD extracts the initial pixel restoration subspace with a cumulative squared-singular-value energy threshold of 0.95. Co-evolution performs a pixel refresh every eight optimization steps and updates semantic adapters at the remaining steps. Expert usage is aggregated across devices and tracked with an exponential moving average of momentum 0.95. The usage-weighted pixel subspace is re-extracted after every 50 pixel refreshes. Gradient projection acts on the semantic adapter output factor B_{sem}. At each subspace refresh, B_{sem} and its AdamW first moment are retracted simultaneously; the input factor A_{sem} and the second moment are not projected.

Functional orthogonality implementation. We select eight approximately evenly spaced layers, in module traversal order, from candidates containing both pixel and semantic adapters. Each layer uses four deterministic shared probes constructed from an equally spaced sequence over $[-1, 1]$ and its circular shifts, reshaped to match the adapter input. The pixel reference response is the arithmetic mean of the expert probe responses at that layer, rather than an image-conditioned Top-k routing output, and is detached from gradient propagation. The two branch responses are flattened and centered separately, and their squared normalized inner product is averaged over probes and layers, with a denominator lower bound of 10^{-12}. This regularizer is enabled only during semantic updates and is not used at inference.

A.1.3 Evaluation Protocol

We evaluate $4\times$ super-resolution on DIV2K-Val (Agustsson & Timofte, 2017), DRealSR (Wei et al., 2020), and RealSR (Cai et al., 2019). For DIV2K-Val, we randomly crop 512×512 HR patches from the original high-resolution images and generate corresponding 128×128 LR patches using the same Real-ESRGAN degradation pipeline as in training. For DRealSR and RealSR, we randomly crop 128×128 LR patches from real low-resolution images and extract the corresponding 512×512 HR regions for paired evaluation. PSNR, SSIM, LPIPS, and DISTS are computed between restored images and their corresponding HR references. FID is computed between the restored and HR reference image sets, whereas CLIPIQA, MUSIQ, and MANIQA directly assess the restored images.

A.2 Theoretical Analysis of OOCE

A.2.1 Fixed-Subspace Misalignment

During online co-evolution, pixel experts are continually updated, and their effective restoration directions change with expert parameters and routing usage. The pixel restoration subspace is therefore not static. Let $\mathcal{U}_t$ denote the pixel restoration subspace at time t, with orthonormal basis U_t and projection matrix:

$$P_t = U_t U_t^\top. \tag{11}$$

Suppose the latest reference subspace is constructed at time $\tau < t$. If subsequent semantic updates continue to use the fixed projection matrix P_τ, the semantic gradient G_t^{sem} is projected as:

$$\tilde{G}_{t,\tau}^{\mathrm{sem}} = (I - P_\tau) G_t^{\mathrm{sem}}. \tag{12}$$

By construction, $P_\tau \tilde{G}_{t,\tau}^{\mathrm{sem}} = 0$, so the projected gradient is orthogonal to the old pixel subspace. However, this orthogonality is no longer guaranteed when the pixel experts continue to evolve and form a new subspace P_t. **Proposition 1.** The residual component in the current pixel restoration subspace of a gradient projected using the old reference subspace P_τ satisfies:

$$\left\| P_t \tilde{G}_{t,\tau}^{\mathrm{sem}} \right\|_F \le \|P_t - P_\tau\|_2 \, \|G_t^{\mathrm{sem}}\|_F \,. \tag{13}$$

Proof. Since $P_\tau(I - P_\tau) = 0$, we have $P_t\tilde{G}^{\text{sem}}_{t,\tau} = P_t(I - P_\tau)G^{\text{sem}}_t = (P_t - P_\tau)(I - P_\tau)G^{\text{sem}}_t$. Eq. 13 follows from the submultiplicativity of matrix norms and $\|I - P_\tau\|_2 \leq 1$. This result shows that potential orthogonality misalignment caused by a fixed reference subspace is directly controlled by subspace drift. When $\mathcal{U}_t$ and $\mathcal{U}_\tau$ have the same dimension:

$$\| P_t - P_\tau \|_2 = \sin\theta_{\max}(\mathcal{U}_t, \mathcal{U}_\tau), \tag{14}$$

Here, $\theta_{\max}$ denotes the largest principal angle between the two subspaces. Thus, even a semantic gradient strictly orthogonal to the old reference subspace may have a nonzero component along current restoration directions when the pixel restoration subspace changes. This analysis shows that, when pixel experts are continually optimized, a fixed subspace represents restoration directions only at a historical point in time. Periodically refreshing the reference subspace allows the orthogonal constraint to evolve with pixel restoration capabilities, reducing semantic-update misalignment caused by subspace drift.

A.2.2 Gradient Projection and Weight Retraction

After the pixel restoration subspace is updated, projecting semantic gradients restricts subsequent updates from re-entering the current restoration subspace. However, this operation acts only on parameter increments and does not automatically remove existing overlap in semantic adapters.

Let B^{sem}_t denote the current semantic LoRA output factor. Given the current projection matrix P_t, an ideal gradient descent step using the projected gradient is:

$$B^{\text{sem}}_{t+1} = B^{\text{sem}}_t - \eta(I - P_t)G^{\text{sem}}_t. \tag{15}$$

Here, η is the learning rate. Since $P_t(I - P_t) = 0$, left-multiplying Eq. (15) by P_t gives:

$$P_t B^{\text{sem}}_{t+1} = P_t B^{\text{sem}}_t. \tag{16}$$

Eq. (16) shows that gradient projection prevents new update components from entering the current pixel restoration subspace but leaves existing parameter overlap unchanged. This issue is particularly relevant when the subspace evolves. Even if semantic parameters satisfy $P_{t-1}B^{\text{sem}}_t = 0$ under the old subspace, after the pixel restoration subspace changes to P_t, we have:

$$P_t B^{\text{sem}}_t = (P_t - P_{t-1})B^{\text{sem}}_t. \tag{17}$$

Thus, semantic parameters satisfying the old constraint may overlap with the updated pixel subspace, and projecting subsequent gradients cannot remove this historical component. To restore parameter feasibility under the current constraint, we retract semantic weights after each subspace refresh:

$$\widehat{B}^{\text{sem}}_t = (I - P_t)B^{\text{sem}}_t. \tag{18}$$

This operation satisfies $P_t\widehat{B}^{\text{sem}}_t = 0$ and also gives the feasible solution closest to the original parameters. Let $\mathcal{C}_t = \{X : U_t^\top X = 0\}$. Then:

$$\widehat{B}^{\text{sem}}_t = \arg\min_{X \in \mathcal{C}_t} \|X - B^{\text{sem}}_t\|_F^2 . \tag{19}$$

This follows directly from orthogonal projection. Any B^{sem}_t admits a unique decomposition into two orthogonal components in $\mathcal{U}_t$ and $\mathcal{U}_t^\perp$. Eq. (18) retains exactly the orthogonal-complement component, restoring feasibility under the current constraint with the minimum perturbation in Frobenius norm.

Gradient projection and weight retraction therefore play complementary roles. Gradient projection constrains subsequent semantic update directions, preventing new update components from entering the current pixel restoration subspace. Weight retraction removes overlap already accumulated in semantic parameters after the subspace changes. The former acts on future updates and the latter corrects historical parameters, jointly maintaining consistency between the semantic branch and dynamic orthogonal constraints.

Similarly, historical first moments in an adaptive optimizer may carry directions from an old subspace. We therefore simultaneously perform the following operation when refreshing the pixel restoration subspace:

$$M_t \leftarrow (I - P_t)M_t. \tag{20}$$

Here, M_t denotes the optimizer's first-moment state. Since Adam's element-wise adaptive preconditioning generally does not commute with P_t, we do not assume that semantic parameters remain strictly orthogonal after every optimization step. Instead, projected updates are combined with periodic weight retraction to repeatedly restore feasibility under the current dynamic subspace during training.

A.2.3 Parameter and Functional Orthogonality

Orthogonality constraints in parameter space do not guarantee complementary functional responses from the two branches in the full network. Let $\Delta\theta_{\mathrm{pix}}$ and $\Delta\theta_{\mathrm{sem}}$ denote local parameter changes in the pixel and semantic branches, respectively, and assume:

$$\Delta\theta_{\mathrm{pix}}^{\top}\Delta\theta_{\mathrm{sem}} = 0. \tag{21}$$

For a given input, first-order linearization around the current parameters approximates their functional responses as:

$$r_{\mathrm{pix}} \approx J_{\mathrm{pix}}\Delta\theta_{\mathrm{pix}}, \qquad r_{\mathrm{sem}} \approx J_{\mathrm{sem}}\Delta\theta_{\mathrm{sem}}. \tag{22}$$

It follows that:

$$r_{\mathrm{pix}}^{\top}r_{\mathrm{sem}} \approx \Delta\theta_{\mathrm{pix}}^{\top}J_{\mathrm{pix}}^{\top}J_{\mathrm{sem}}\Delta\theta_{\mathrm{sem}}. \tag{23}$$

Parameter-space orthogonality therefore does not directly imply orthogonal functional responses. In DynaOrtho-SR, which involves spatial routing and shared nonlinear propagation, this distinction may lead to residual correlation between the actual responses of the two branches.

We therefore directly constrain pixel and semantic branch responses on shared probes. Let $\mathcal{Q}$ denote the selected layer set and M the number of probes per layer. Following the main-text definition, let $\bar{p}_{\ell,m}$ and $\bar{s}_{\ell,m}$ denote the centered and flattened pixel and semantic responses, respectively. F-Ortho is defined as:

$$\mathcal{L}_{\mathrm{FO}} = \frac{1}{|\mathcal{Q}|M}\sum_{\ell\in\mathcal{Q}}\sum_{m=1}^{M}\frac{\langle\bar{p}_{\ell,m},\bar{s}_{\ell,m}\rangle^2}{\max\{\|\bar{p}_{\ell,m}\|_2^2\|\bar{s}_{\ell,m}\|_2^2,\epsilon\}}. \tag{24}$$

The pixel response is detached, allowing semantic adapters to learn complementary responses against a fixed reference. This constraint directly suppresses directional correlation between the functional responses of the two branches, complementing the dynamic parameter-space orthogonal constraint imposed by OOCE.

A.3 Training and Inference Algorithms

Algorithm 1 presents the complete training procedure of DynaOrtho-SR. The model first undergoes pixel restoration warm-up and constructs an initial pixel restoration subspace, then enters online orthogonal co-evolution with alternating updates of pixel experts and semantic adapters. During semantic updates, gradients are projected using the current pixel subspace, while F-Ortho suppresses branch response redundancy. Following pixel updates, the restoration subspace is periodically refreshed according to expert usage, and semantic adapter weights and the optimizer's first moments are retracted simultaneously. Algorithm 2 presents controllable inference, which adjusts the contributions of pixel restoration and semantic enhancement to obtain different fidelity-perception tradeoffs with the same model.

A.4 User Study

To further assess human-perceived visual quality, we conduct a user study comparing DynaOrtho-SR with nine representative Real-ISR methods. Test images are grouped into five categories, Animals, Buildings, Human Faces, Vegetation, and Characters, to cover diverse structures and textures. Participants select the candidate with the best visual quality based on content consistency with the reference image and the naturalness of textures and details. As shown in Fig. 6, DynaOrtho-SR receives the highest preference in every category, with an overall preference rate of 27.6%, clearly exceeding the runner-up, FaithDiff. Specifically, our method receives preference rates of 25.6%, 24.4%,

Algorithm 1 Training of DynaOrtho-SR

Input: Training set $\mathcal{S}$; total updates T; warm-up length T_{w}; pixel-update interval K; subspace-refresh interval R; energy threshold τ; usage momentum μ.
Output: Pixel parameters Θ_{pix} and semantic parameters Θ_{sem}.
All backbone parameters are frozen. Each iteration includes gradient accumulation.

Initialize adapters, routers, AdamW, and shared probes $\mathcal{V}$; $n \leftarrow 0$
Warm up Θ_{pix} for T_{w} updates using $\mathcal{L}_{\mathrm{pix}}$ with semantic adapters inactive
$\mathbf{U}_\ell \leftarrow \mathrm{SVD}_\tau([\Delta \mathbf{W}_{\ell,e}^{\mathrm{pix}}]_{e=1}^{E})$, for each constrained layer ℓ
for $t = T_{\mathrm{w}}, \ldots, T-1$ **do**
  Sample training batch from $\mathcal{S}$; $j \leftarrow t - T_{\mathrm{w}}$
  if $j > 0$ **and** $j \bmod K = 0$ **then**
    Activate pixel adapters only; update Θ_{pix} using $\mathcal{L}_{\mathrm{pix}}$ and AdamW
    Aggregate expert usage $\mathbf{q}$ across microbatches and GPUs; $n \leftarrow n+1$
    $\bar{\mathbf{q}} \leftarrow \mathbf{q}$ if $n = 1$, otherwise $\bar{\mathbf{q}} \leftarrow \mu\bar{\mathbf{q}} + (1-\mu)\mathbf{q}$
    if $n \bmod R = 0$ **then**
      for each constrained layer ℓ **do**
        $\mathbf{U}_\ell \leftarrow \mathrm{SVD}_\tau([\sqrt{\bar{q}_{\ell,e}}\,\mathbf{B}_{\ell,e}^{\mathrm{pix}}]_{e=1}^{E})$
        $\mathbf{P}_\ell^{\perp} \leftarrow \mathbf{I} - \mathbf{U}_\ell\mathbf{U}_\ell^{\top}$
        $(\mathbf{B}_\ell^{\mathrm{sem}}, \mathbf{m}_\ell) \leftarrow (\mathbf{P}_\ell^{\perp}\mathbf{B}_\ell^{\mathrm{sem}}, \mathbf{P}_\ell^{\perp}\mathbf{m}_\ell)$
      end for
    end if
  else
    Activate both adapter groups; keep Θ_{pix} fixed
    Evaluate $\mathcal{L}_{\mathrm{sem}}$, including $\mathcal{L}_{\mathrm{FO}}$ on $\mathcal{V}$ with detached mean pixel-expert responses
    Compute $\mathbf{g} \leftarrow \nabla_{\Theta_{\mathrm{sem}}}\mathcal{L}_{\mathrm{sem}}$
    Replace each constrained B-factor gradient by $\mathbf{g}_{B_\ell} \leftarrow (\mathbf{I} - \mathbf{U}_\ell\mathbf{U}_\ell^{\top})\mathbf{g}_{B_\ell}$
    Update Θ_{sem} using AdamW with $\mathbf{g}$
  end if
end for
$\mathbf{m}_\ell$ denotes the AdamW first moment; the second moment is unchanged. SVD_τ *retains the leading left singular vectors at squared-energy threshold τ.*

Algorithm 2 Controllable Inference of DynaOrtho-SR

Input: LR image I_{LR}; trained model; VAE $(\mathcal{E}, \mathcal{D})$; control scales $\lambda_{\mathrm{pix}}, \lambda_{\mathrm{sem}}$; fixed timestep $t_\star$ and condition c.
Output: Restored image $\widehat{I}_{\mathrm{SR}}$.

Upsample I_{LR} to target size, obtaining $\widetilde{I}_{\mathrm{LR}}$
$z_{\mathrm{LR}} \leftarrow \mathcal{E}(\widetilde{I}_{\mathrm{LR}})$; construct local degradation field $\mathbf{F}_{\mathrm{deg}}$
Condition the routers on $\mathbf{F}_{\mathrm{deg}}$
Predict $\widehat{r}_{\mathrm{all}}$ from $(z_{\mathrm{LR}}, t_\star, c)$ with both adapter groups active
if $\lambda_{\mathrm{pix}} = \lambda_{\mathrm{sem}} = 1$ **then**
  $\widehat{r} \leftarrow \widehat{r}_{\mathrm{all}}$
else
  Predict $\widehat{r}_{\mathrm{pix}}$ with pixel adapters only, using the same $(z_{\mathrm{LR}}, t_\star, c)$
  $\widehat{r} \leftarrow \lambda_{\mathrm{pix}}\widehat{r}_{\mathrm{pix}} + \lambda_{\mathrm{sem}}(\widehat{r}_{\mathrm{all}} - \widehat{r}_{\mathrm{pix}})$
end if
return $\widehat{I}_{\mathrm{SR}} = \mathcal{D}(z_{\mathrm{LR}} - \widehat{r})$

33.6%, 26.4%, and 28.0% for Animals, Buildings, Human Faces, Vegetation, and Characters, respectively, with the largest advantage on Human Faces. These results indicate that DynaOrtho-SR restores clear local details while effectively avoiding oversmoothing and unnatural texture synthesis, yielding more stable structural fidelity and perceptual realism across content types.

A.5 Additional Comparisons with GAN-Based Real-World ISR Methods

A.5.1 Quantitative Comparisons

We further compare with BSRGAN (Zhang et al., 2021) and Real-ESRGAN (Wang et al., 2021b) on DRealSR, RealSR, and DIV2K-Val, as reported in Table 4. GAN-based methods have advantages on full-reference metrics such as PSNR and SSIM, whereas DynaOrtho-SR achieves better FID,

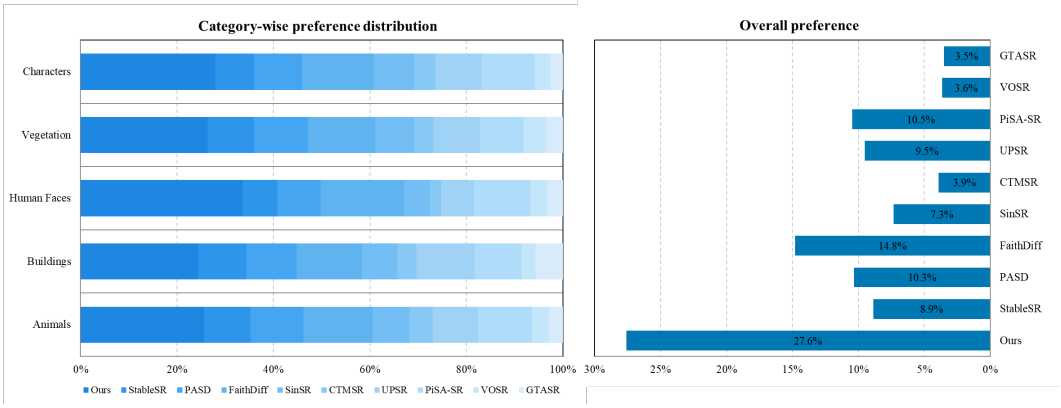

Figure 6: User study results. Left: Category-wise user preference distributions. DynaOrtho-SR receives the highest preference across all five categories. Right: Overall user preference rates. DynaOrtho-SR achieves the highest overall preference.

Table 4: Quantitative comparison of DynaOrtho-SR with GAN-based real-world image super-resolution methods on DRealSR, RealSR, and DIV2K-Val. The best results are highlighted in bold.

Datasets	Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
DRealSR	BSRGAN	**29.8991**	0.8223	0.2786	0.2149	74.0354	0.4933	55.0558	0.4750
	Real-ESRGAN	29.6028	**0.8251**	**0.2715**	**0.2056**	69.3719	0.4439	52.5118	0.4875
	Ours(sem1.2)	29.3248	0.7851	0.2970	0.2101	**58.1852**	**0.6858**	**62.6833**	**0.6115**
RealSR	BSRGAN	**26.7444**	**0.7842**	**0.2549**	0.1983	105.6003	0.5026	64.1254	0.5347
	Real-ESRGAN	25.9776	0.7786	0.2555	0.1915	96.5419	0.4597	62.2656	0.5586
	Ours(sem1.2)	25.9480	0.7436	0.2611	**0.1834**	**77.8680**	**0.6688**	**68.8955**	**0.6608**
DIV2K-Val	BSRGAN	22.6081	0.5955	0.3533	0.2356	99.2177	0.5378	62.1169	0.5061
	Real-ESRGAN	**22.7640**	**0.6180**	0.3172	0.2188	79.9873	0.5339	61.5593	0.5456
	Ours(sem1.2)	21.9406	0.5717	**0.2897**	**0.1986**	**63.2970**	**0.6904**	**68.0150**	**0.6398**

CLIPIQA, MUSIQ, and MANIQA on all three datasets while remaining competitive in DISTS. These results indicate that conventional distortion metrics emphasize pixel or structural consistency with reference images. For real-world degradation ISR, however, visually natural and plausible high-frequency details may not align exactly with the reference image, so full-reference metrics alone cannot fully characterize perceptual quality.

A.5.2 Qualitative Comparisons

Figure 7 presents visual comparisons between DynaOrtho-SR and GAN-based methods. BSRGAN and Real-ESRGAN recover the main structures but remain susceptible to oversmoothing, missing textures, or missing local details in complex regions such as feathers, clothing, and plants. In contrast, DynaOrtho-SR restores clearer, more natural fine-grained textures and richer local details while preserving the main structures, consistent with its performance on no-reference perceptual metrics and FID.

A.6 Visual Analysis of Fidelity-Perception Controllability

To further illustrate inference-time controllability, we fix $\lambda_{\text{pix}} = 1.0$ and gradually increase λ_{sem} from 1.0 to 1.6. As shown in Fig. 8, increasing the semantic branch weight progressively sharpens high-frequency details such as leaf textures, petal edges, feather structures, architectural ornaments, and scales. Smaller λ_{sem} emphasizes input structure preservation and smoother restoration, whereas moderate increases introduce richer natural textures. As λ_{sem} increases further, local detail enhancement saturates and may exhibit stronger sharpening or texture enhancement. These results show that a single DynaOrtho-SR model provides continuous, controllable visual changes between faithful

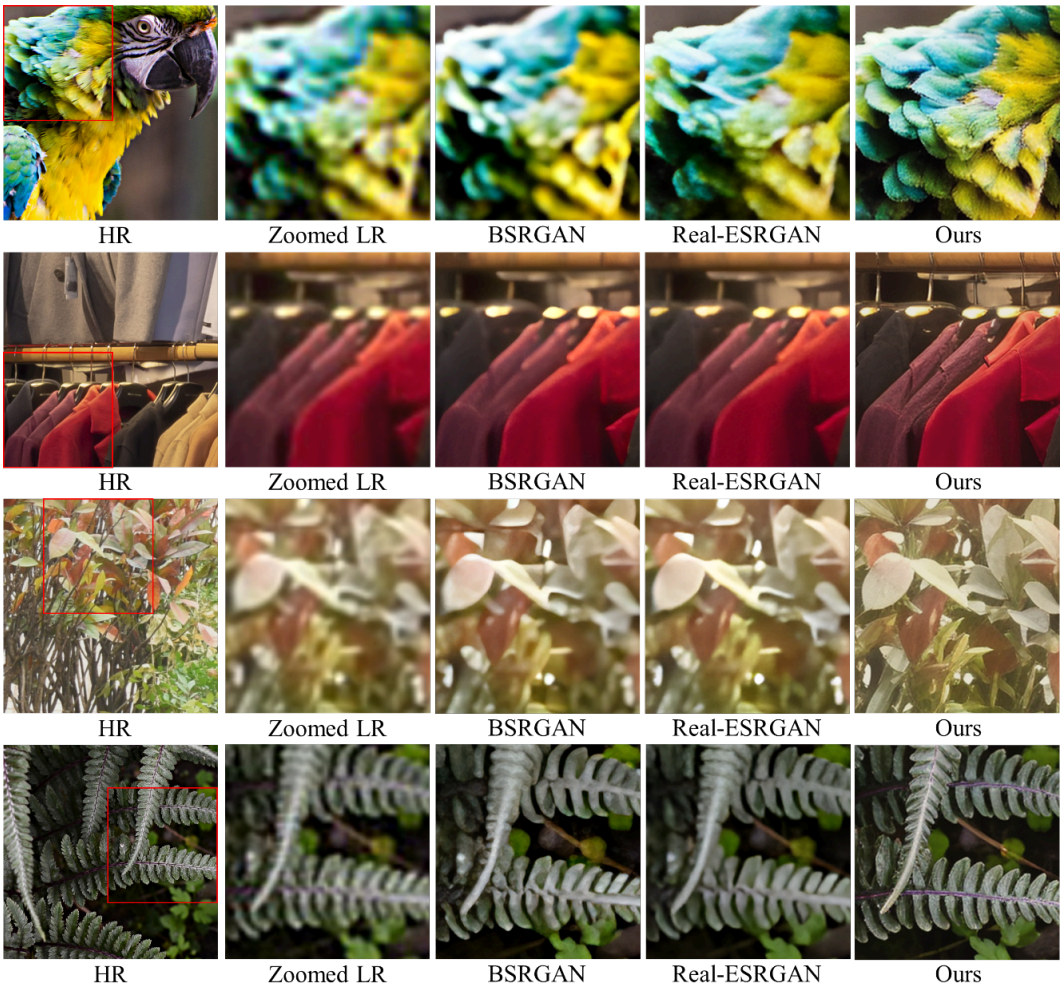

Figure 7: Visual comparisons of DynaOrtho-SR with GAN-based methods. Please zoom in for a better view.

restoration and perceptual enhancement by simply adjusting semantic branch strength, without retraining.

A.7 Additional Qualitative Comparisons

Figure 9 complements the main-text structural comparisons with texture-rich regions containing flowers, small fruits, and animal textures.

Figures 10–12 provide additional visual comparisons between DynaOrtho-SR and representative diffusion-based ISR methods. Figures 10 and 11 focus on structural and content fidelity: DynaOrtho-SR better preserves original structures in repetitive edges, regular lines, text, and object contours, while reducing local deformations and inconsistent restoration. Figure 12 further illustrates texture and detail recovery, where our method generates clearer, more natural local textures in fine-grained regions such as petals and food surfaces. Overall, these results further validate the stable restoration capability of DynaOrtho-SR in balancing structural fidelity and detail synthesis.

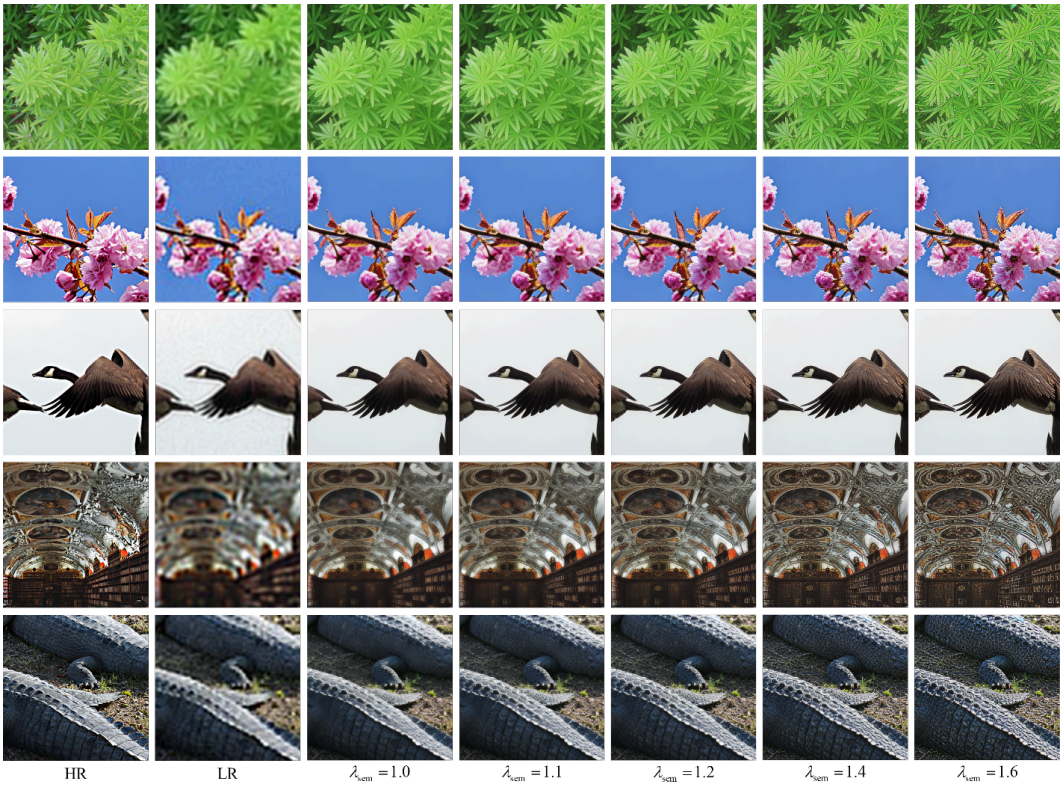

Figure 8: Visual results of DynaOrtho-SR under different semantic control coefficients. We fix $\lambda_{\text{pix}} = 1.0$ and progressively increase λ_{sem} to obtain different fidelity-perception operating points.

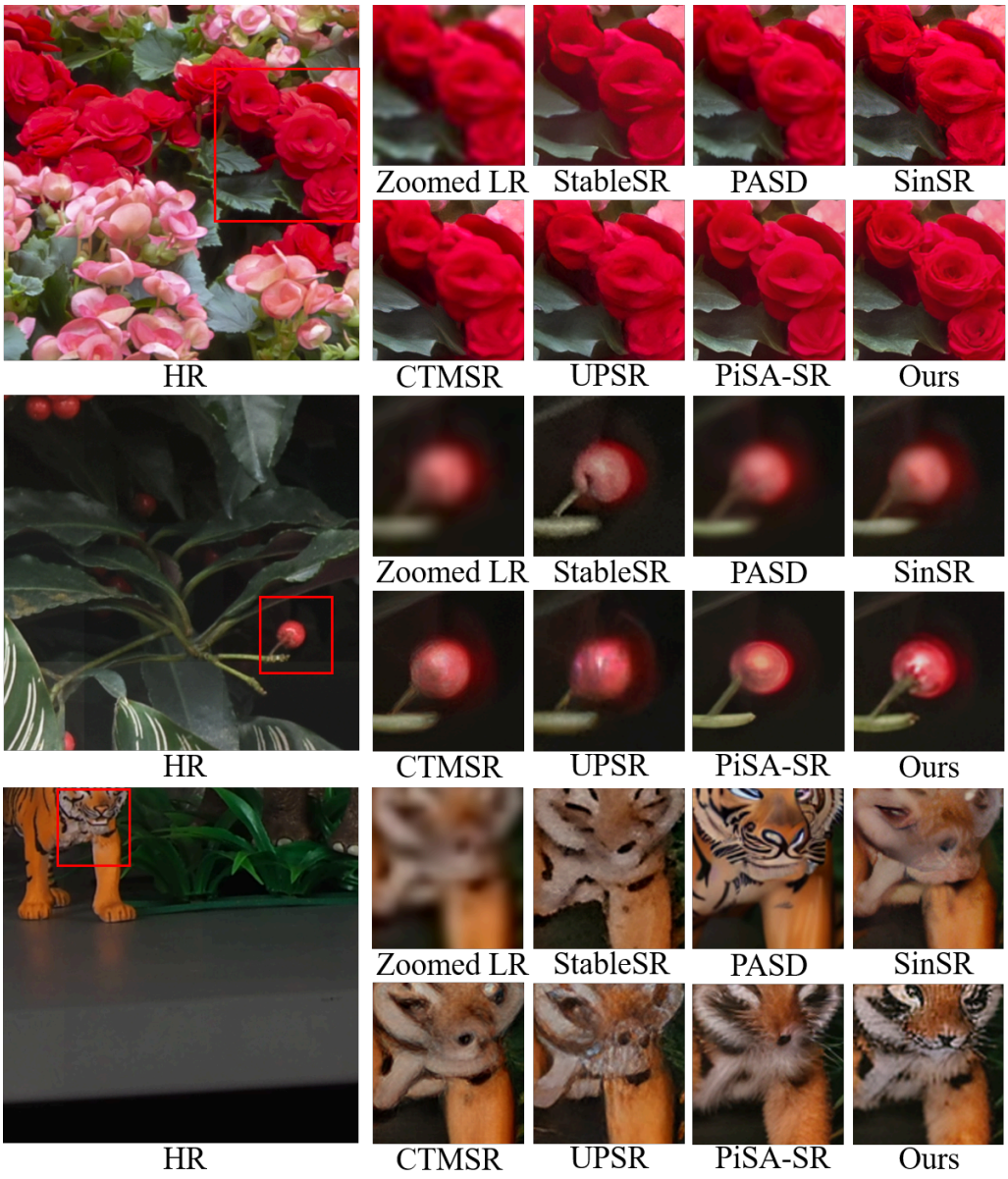

Figure 9: Visual comparisons of different Real-ISR methods on natural texture-rich regions. Please zoom in for a better view.

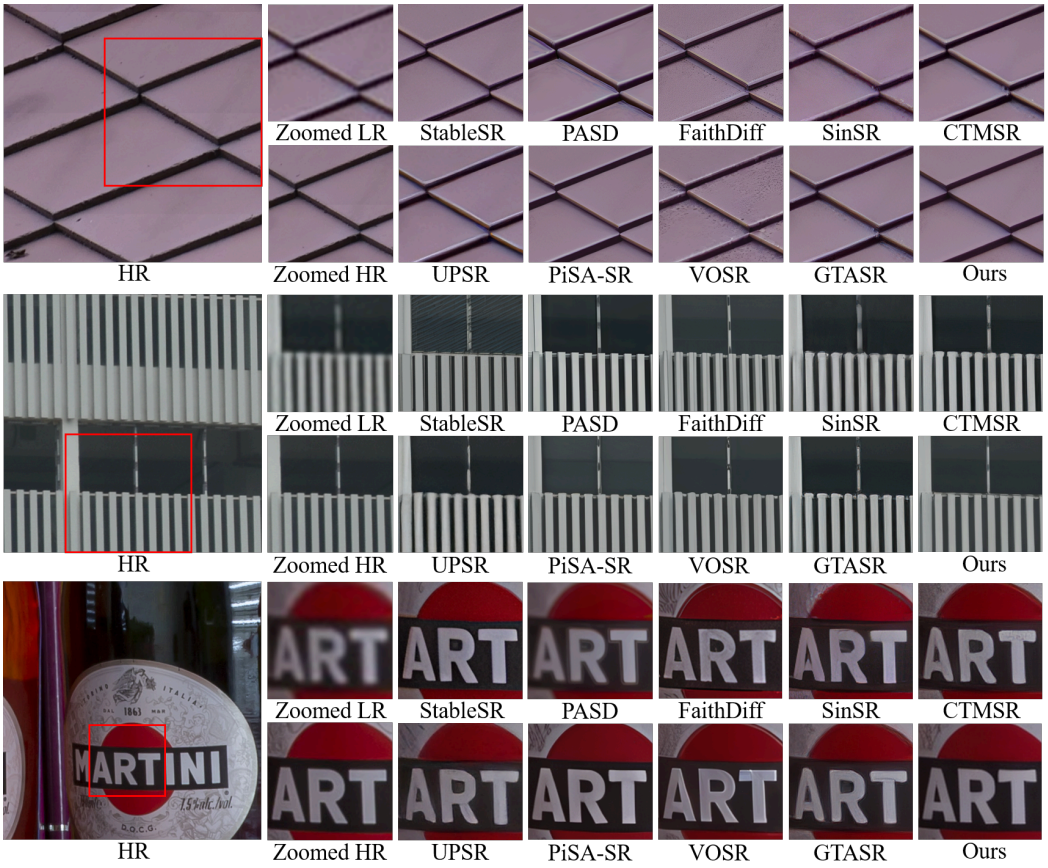

Figure 10: Additional visual comparisons of DynaOrtho-SR with representative diffusion-based super-resolution methods on structure-sensitive regions. DynaOrtho-SR better preserves regular edges and local structures. Please zoom in for a better view.

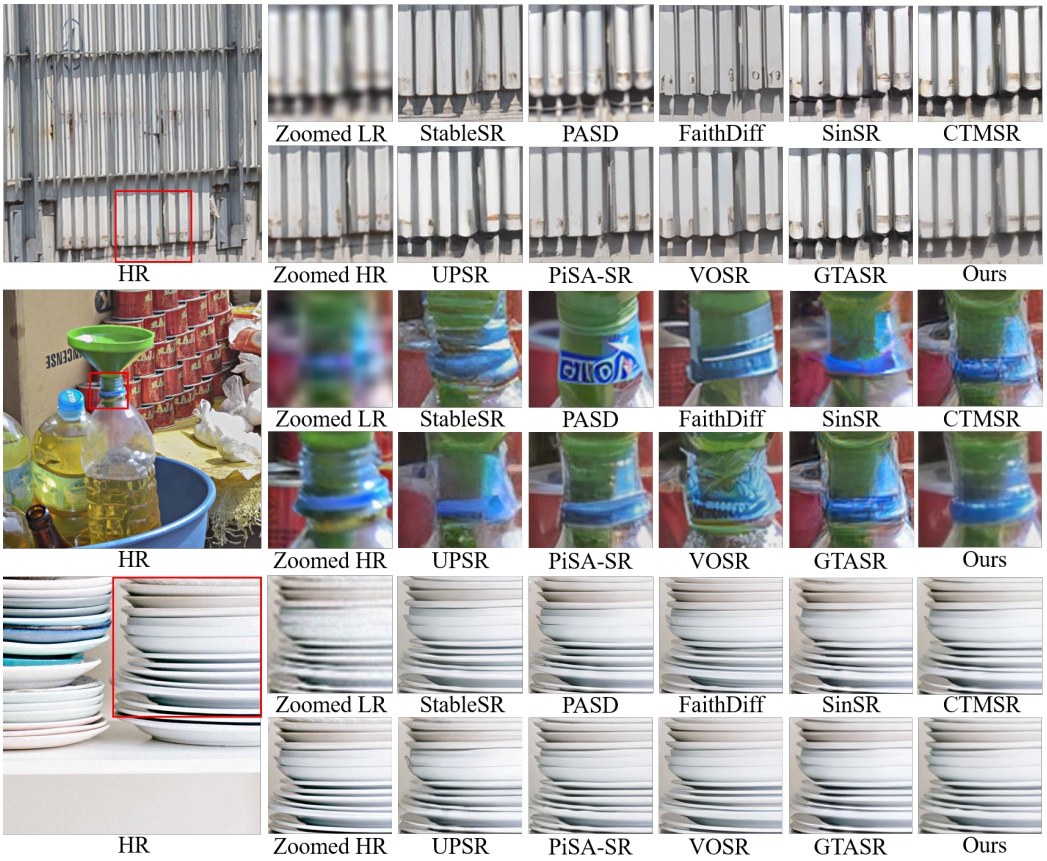

Figure 11: Additional visual comparisons of DynaOrtho-SR with representative diffusion-based super-resolution methods on fidelity-sensitive regions. DynaOrtho-SR exhibits more stable structure preservation in repetitive lines, text, and object contours. Please zoom in for a better view.

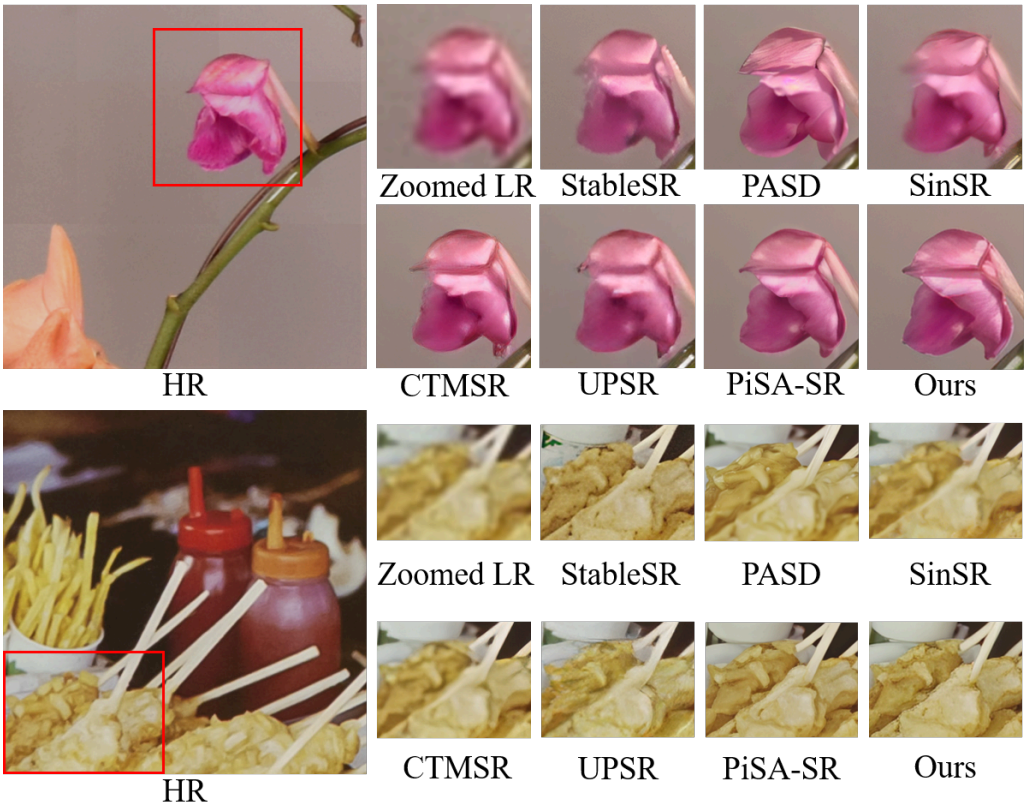

Figure 12: Additional visual comparisons of DynaOrtho-SR with representative diffusion-based super-resolution methods on texture-rich regions. DynaOrtho-SR recovers clearer and more natural fine-grained textures. Please zoom in for a better view.