PANOFORCING: ENFORCING MEMORY-SCALABLE FULL-CONTEXT FOR BLIND PANORAMIC SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

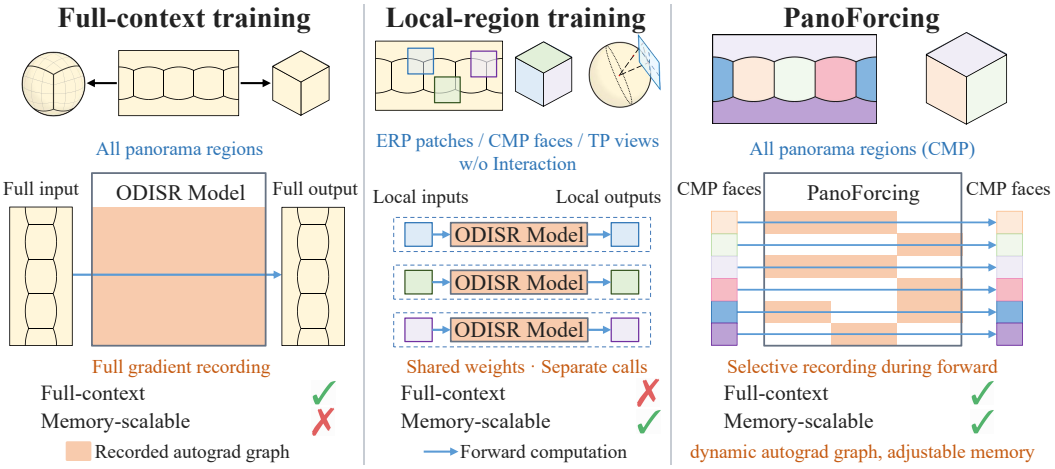


Figure 1: Prior full-context training incurs substantial memory costs, while local-region training uses less memory but fragments panoramic context. In contrast, PanoForcing preserves full-context forward computation with scalable memory.

ABSTRACT

Blind panoramic super-resolution requires realistic details that remain coherent across the full scene and faithful to the input. Such scene-wide coherence depends on full panoramic context, which is costly to preserve: processing isolated regions fragments this context, while even a one-step diffusion baseline requires over 80 GB of GPU memory for full-panorama fine-tuning. We propose PanoForcing, a one-step diffusion framework that preserves full panoramic context during the forward pass under configurable training and inference memory budgets. PanoForcing selectively constructs the training-memory-related autograd graph while maintaining full-context forward computation, and incorporates lightweight observation guidance within a single network pass to constrain generated details. Experimental results demonstrate high-quality full-context panoramic reconstruction across a broad range of controllable memory budgets. PanoForcing achieves the best WS-PSNR, LPIPS, and FID among the compared methods. For memory scalability, it freely controls peak training memory from 8.7 to 61.7 GB and dynamically controls peak inference memory from 3.5 to 7.9 GB according to device conditions.

1 INTRODUCTION

Omnidirectional images capture a complete $180° \times 360°$ field of view, representing the surrounding scene within a single image. This broad visual coverage benefits autonomous driving, embodied intelligence, robot navigation, and immersive visual experiences. However, panoramic content is typically viewed through narrow perspective viewports, which require very high spatial resolution to preserve local detail (Ai et al., 2022). Capturing panoramas at such resolutions substantially

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.

increases sensing, storage, and processing costs. Omnidirectional image super-resolution (ODISR) therefore offers a practical alternative by reconstructing high-resolution (HR) panoramas from more readily available low-resolution (LR) observations.

ODISR has evolved along two main directions. Regression-based methods adopt CNN- or Transformer-based models that account for spherical distortion and exploit complementary panoramic representations (Deng et al., 2021; Yu et al., 2023; Yoon et al., 2022; Wang et al., 2024a). Generative methods instead introduce pretrained diffusion priors from general image SR (Wang et al., 2024b) into ODISR (Li et al., 2024; Liu et al., 2025; 2026). These priors capture strong natural-image statistics and generate fine details. They are especially useful when high-resolution panoramic training data are scarce. To avoid the latency of iterative denoising, one-step diffusion methods have been developed for general image SR (Wang et al., 2024c; Wu et al., 2024b; Zhang et al., 2024; Dong et al., 2025), and RealOSR (Sheng et al., 2024) further brings this paradigm to ODISR. However, most of these methods reconstruct local views or patches separately, which makes it difficult to perceive the entire panorama, and maintaining consistency with the LR observation often requires extra computation, such as iterative correction (Li et al., 2024).

Different view directions depict the same scene: a local region shares semantics, illumination, and structures with the rest of the panorama, and its reconstruction must remain consistent across view boundaries. Therefore, a generative model should see the entire panorama when reconstructing any local region. We call this complete panoramic input *full-context*. However, keeping the full context available throughout reconstruction incurs heavy memory costs, even for compact models. In the full-panorama training setting, BPOSR (Wang et al., 2024a) exceeds 24 GB of GPU memory despite having only 2.19M parameters. A large diffusion backbone makes the cost even more restrictive: as shown in Fig. 5, even with gradient checkpointing, LoRA fine-tuning of OSEDiff (Wu et al., 2024b) on a full equirectangular panorama requires more than 80 GB. Inference avoids storing activations for backward, but jointly processing high-resolution panoramic features still costs substantial memory. The challenge is therefore to preserve complete scene context while making both training and inference memory adjustable to the available hardware.

Existing approaches reduce memory in two main ways. The first samples local views, such as equirectangular projection (ERP) patches, cubemap projection (CMP) faces, or tangent-plane (TP) views. This reduces the amount of data processed together. However, training on isolated regions restricts the context available to each prediction. The second optimizes how computation is executed. Gradient checkpointing recomputes intermediate activations and offloads them to storage. Efficient attention lowers the cost of token interaction. Nevertheless, these techniques still record autograd graph through all participating inputs. As a result, they cannot independently control how much panoramic context enters the forward pass, nor how much computation is recorded for backward. This gap motivates a framework that preserves full-context computation while exposing explicit memory controls for training and inference.

We propose PanoForcing, a one-step diffusion framework for blind panoramic super-resolution under configurable memory budgets. It combines two complementary principles: full-context forcing and observation-consistency forcing. Full-context forcing keeps the entire panorama visible to the model when reconstructing each local region. The key insight is that preserving full context during the forward pass does not require recording all computation for backward. As illustrated in Fig. 1, PanoForcing selectively constructs the autograd graph without restricting the input context. It also supports adjustable memory use during both training and inference to accommodate different device constraints. Observation-consistency forcing addresses a separate limitation: full context alone does not ensure fidelity to the LR input. PanoForcing therefore uses the LR observation to guide detail generation within a single network pass. Together, these principles support coherent and input-faithful panoramic reconstruction under controllable memory budgets.

Our main contributions are as follows.

❑ (1) We propose PanoForcing, a one-step diffusion framework for blind panoramic super-resolution. It unifies full-context forcing and observation-consistency forcing under configurable memory budgets.

❑ (2) We introduce Spherical Subgraph Routing (SSR) to realize full-context forcing. SSR selectively retains autograd subgraphs while preserving complete panoramic context in the forward pass. This enables configurable training memory without restricting the model's view of the scene.

❑ (3) We develop Fourier Pseudoinverse Guidance (FPG) to realize observation-consistency forcing. Grounded in a frequency-domain analysis of re-degradation, FPG uses the LR observation to guide detail generation within a single network pass. This promotes input-faithful reconstruction without iterative denoising.

❑ (4) Experiments on ODISR and SUN 360 show the best WS-PSNR, LPIPS, and FID among the compared methods. Further analyses demonstrate adjustable training and inference memory with full panoramic context and assess reconstruction quality across training budgets.

2 Related Work

2.1 Omnidirectional Image Super-Resolution

ODISR methods exploit spherical geometry (Cai et al., 2024) and address non-uniform ERP sampling through latitude-adaptive or distortion-aware modeling (Deng et al., 2021; Yu et al., 2023; Sun et al., 2023). SphereSR (Yoon et al., 2022) learns a continuous spherical representation, while BPOSR (Wang et al., 2024a) fuses ERP and CMP features. Comparative studies examine ERP, cubemap, and tangent-plane representations (Pi et al., 2023). OmniSSR (Li et al., 2024) combines diffusion-based tangent-view reconstruction with iterative fusion and observation correction. Recent diffusion models also use latitude-aware conditioning in pixel and latent domains (Liu et al., 2025; 2026).

PanoForcing complements these methods with full-context reconstruction under configurable memory budgets.

2.2 Memory-Efficient and Memory-Scalable Computation

Memory-efficient computation reduces memory use by optimizing token interactions and managing intermediate states. Windowed attention restricts token interactions to local windows (Liu et al., 2021). FlashAttention (Dao et al., 2022) instead optimizes memory access without materializing the full attention matrix. To reduce activation memory, checkpointing recomputes intermediate states during backward (Chen et al., 2016). Offloading transfers these states to CPU memory (Ren et al., 2021). In standard full-gradient training, these methods still differentiate all participating branches.

PanoForcing instead selectively records autograd paths to control training memory while preserving full-context forward computation.

2.3 Generative Super-Resolution and Observation Guidance

Diffusion-based super-resolution uses pretrained natural-image priors to reconstruct realistic details (Wang et al., 2024b; Yang et al., 2024a; Yu et al., 2024). To reduce inference cost, SinSR (Wang et al., 2024c), OSEDiff (Wu et al., 2024b), and S3Diff (Zhang et al., 2024) replace iterative sampling with one-step reconstruction.

Observation guidance constrains iterative diffusion sampling (Song et al., 2023; Chung et al., 2022). DDRM (Kawar et al., 2022) and DDNM (Wang et al., 2023) use known linear operators; DPS (Chung et al., 2023) handles noisy linear or nonlinear measurements. For unknown degradations, GDP (Fei et al., 2023) estimates degradation parameters, while Diff-DKP (Yang et al., 2024b) models kernels dynamically. PSLD (Rout et al., 2023) extends posterior sampling to latent diffusion, and OmniSSR (Li et al., 2024) guides RGB and latent panoramic estimates. Unlike these iterative methods, RealOSR (Sheng et al., 2024) provides multi-depth latent guidance for one-step reconstruction under unknown degradation.

PanoForcing also targets observation-guided one-step reconstruction, with an emphasis on computationally efficient guidance.

3 METHOD

3.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

Blind panoramic super-resolution aims to recover an HR panorama from an LR observation $\mathbf{y}$. The observation is modeled as

$$\mathbf{y} = \mathcal{D}(\mathbf{x}_{\mathrm{GT}}) + \mathbf{n}, \tag{1}$$

where $\mathbf{x}_{\mathrm{GT}}$ is the ground-truth HR panorama, $\mathcal{D}$ is an unknown degradation operator, and $\mathbf{n}$ denotes additive noise. We represent $\mathbf{y}$ as six cubemap faces $\{\mathbf{y}_i\}_{i=1}^{6}$ and jointly reconstruct their HR counterparts. Rather than processing each face in isolation, full-context reconstruction uses all six observed faces to predict each output face:

$$\hat{\mathbf{x}}_i = F_i(\mathbf{y}_1, \ldots, \mathbf{y}_6), \qquad i \in \{1, \ldots, 6\}, \tag{2}$$

where F_i denotes the reconstruction mapping for face i and $\hat{\mathbf{x}}_i$ is its HR prediction.

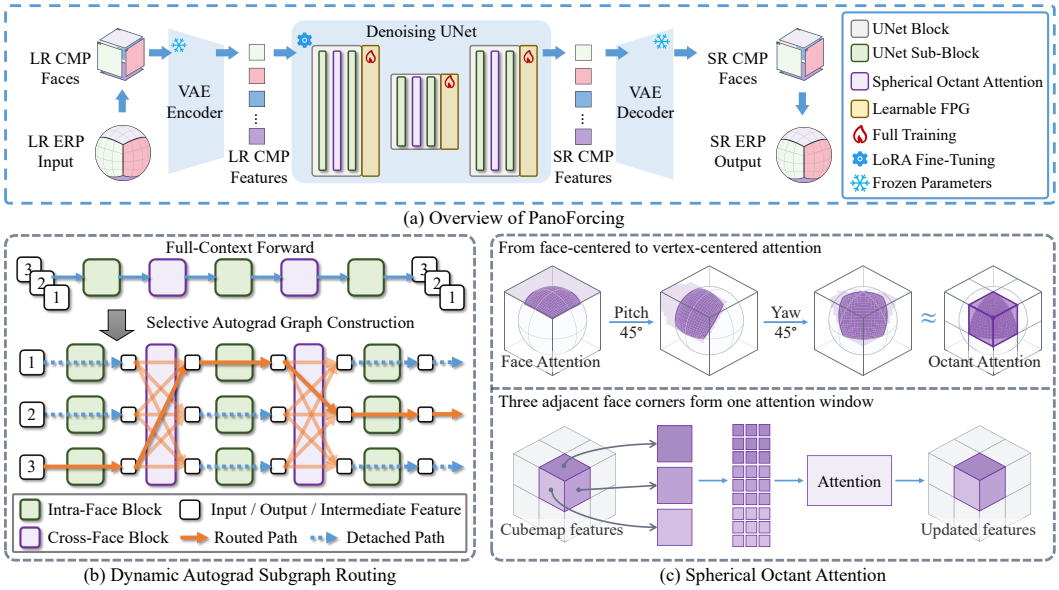


Figure 2: (a) PanoForcing combines full-context modeling across six cubemap faces with multi-depth observation guidance in a single U-Net pass. (b) Selective autograd recording enables controllable training memory without restricting panoramic context. (c) Spherical Octant Attention enables efficient cross-face interaction without interpolation or additional attention parameters.

Fig. 2(a) presents the overall PanoForcing framework, which combines full-context forcing and observation-consistency forcing in a single denoising pass. The six cubemap faces are encoded into latent features and jointly refined. The refined features are decoded into RGB faces and reprojected to ERP. For full-context forcing, Spherical Subgraph Routing (SSR) keeps all six faces in the forward pass. It selectively record autograd histories to control training memory. At the same time, cross-face attention shares information across faces to guide local reconstruction. For observation-consistency forcing, Fourier Pseudoinverse Guidance (FPG) uses the LR observation to constrain reconstruction. The learnable FPG modules marked in Fig. 2(a) correct intermediate features at multiple network depths. Training-free FPG further corrects the reconstructed panorama in RGB space.

PanoForcing provides separate memory controls for training and inference. During training, activation memory is controlled by how many faces are randomly selected for recording autograd graph at each depth for intra-face computation. Inference requires no autograd routing. Instead, peak memory is controlled by how many faces are processed concurrently, while full-context computation is preserved. The following sections introduce SSR and FPG. Training settings and cubemap sampling details are provided in Appendix A.

3.2 SPHERICAL SUBGRAPH ROUTING

SSR keeps all six faces in the forward pass but retains autograd histories for only K intra-face paths (Fig. 2(b)). Within SSR, Dynamic Autograd Subgraph Routing controls training memory, and Spherical Octant Attention enables efficient cross-face interaction.

Dynamic Autograd Subgraph Routing Fig. 2(b) illustrates selective autograd recording during the forward pass. All six faces are processed, but only K faces retain autograd subgraphs within each intra-face block. We organize the diffusion denoiser into *intra-face blocks* and *cross-face blocks*. An intra-face block processes each face independently through separate invocations. A cross-face block jointly processes multiple faces to exchange information. To reuse the pretrained denoiser, we select a subset of Transformer blocks and convert them into cross-face blocks, while the remaining Transformer and ResNet blocks act as intra-face blocks.

Each intra-face block records autograd subgraphs for a selected set $\mathcal{R}$ with K faces, while faces outside $\mathcal{R}$ are processed without gradient recording. The subsequent cross-face block mixes features across faces, connecting previously unselected faces to the retained subgraphs. These connections allow gradients to propagate between faces selected at different depths. After each cross-face block G_ℓ, we randomly select K faces again as the next set $\mathcal{R}'$. Each selected face must be reachable from $\mathcal{R}$ through G_ℓ, preserving its connection to the retained autograd subgraphs. This produces the changing orange paths in Fig. 2(b).

Implementing this subgraph routing is very simple. It only requires controlling the gradient recording during the forward pass. Standard backpropagation, on the other hand, does not need to be modified. During inference, gradient recording is disabled for all faces. We control peak memory by processing at most B faces concurrently within each intra-face block. Smaller batches reduce temporary activation memory, while cross-face blocks still receive features from all six faces to preserve full context. Algorithms 1 and 2 summarize training and inference, respectively.

Algorithm 1: Memory-scalable full-context training (Dynamic Autograd Subgraph Routing)

Input: Face features $\{\mathbf{h}_i\}_{i=1}^{6}$, memory budget K, intra-face blocks P_ℓ, cross-face blocks G_ℓ

Select an initial retained set $\mathcal{R}$ of K faces
for *each stage* ℓ **do**
  for $i = 1$ **to** 6 **do**
    $\mathbf{u}_i \leftarrow P_\ell(\mathbf{h}_i)$ // record gradients if $i \in \mathcal{R}$
  $\mathbf{h}_{1:6} \leftarrow G_\ell(\mathbf{u}_{1:6})$ // always record gradients
  Select K faces as $\mathcal{R}$ for the next stage

Algorithm 2: Memory-scalable full-context inference

Input: Face features $\{\mathbf{h}_i\}_{i=1}^{6}$, parallelism budget B, intra-face blocks P_ℓ, cross-face blocks G_ℓ

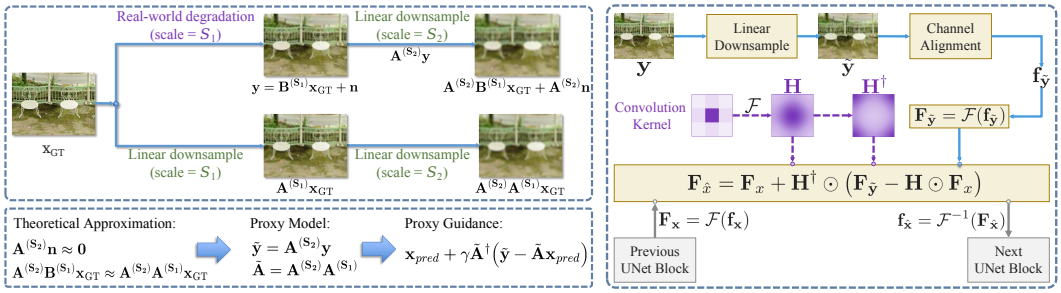
for *each stage* ℓ **do**
  for $j = 1$ **to** $\lceil 6/B \rceil$ **do**
    $s \leftarrow (j-1)B + 1$, $e \leftarrow \min(jB, 6)$
    $\mathbf{u}_{s:e} \leftarrow P_\ell(\mathbf{h}_{s:e})$
  $\mathbf{h}_{1:6} \leftarrow G_\ell(\mathbf{u}_{1:6})$

Spherical Octant Attention Dynamic Autograd Subgraph Routing limits graph construction within intra-face blocks, but cross-face attention still processes features from all six faces. Dense attention over high-resolution features in shallow U-Net layers can exhaust GPU memory. This motivates a more efficient spherical attention mechanism.

Inspired by sliding-window attention in planar images, we move windows on the sphere through rotation rather than translation. The upper part of Fig. 2(c) illustrates a face-centered window rotated by 45° in pitch and then 45° in yaw. The rotated window lies near a cube vertex and approximately overlaps the three adjacent face corners. Instead of resampling this window, we group the three corner patches into a discrete window corresponding to a spherical octant. We apply attention within the eight resulting windows and refer to this mechanism as *Spherical Octant Attention*.

This design offers three advantages. First, it avoids resampling loss. As illustrated in the lower part of Fig. 2(c), tokens are gathered directly from the cubemap grids. After attention, the updated tokens are returned to their original face positions. No interpolation-based resampling is performed throughout this process. Second, octant windows reduce the quadratic cost of attention. For six $H \times W$ feature maps, dense attention computes $(6HW)^2 = 36H^2W^2$ pairwise scores. In comparison, each octant window contains three $H/2 \times W/2$ corner patches. The eight windows therefore require $8(3HW/4)^2 = 4.5H^2W^2$ scores, an eightfold reduction. Third, because it reuses pretrained Transformer blocks by only regrouping tokens, no additional parameters are introduced.

3.3 Fourier Pseudoinverse Guidance



(a) Frequency-domain analysis of re-degradation (b) Learnable multi-depth Fourier guidance

Figure 3: Fourier Pseudoinverse Guidance. (a) Re-degradation motivates the proxy guidance. (b) The proxy operator $\mathbf{H}$ and its regularized pseudoinverse $\mathbf{H}^{\dagger}$ correct intermediate U-Net features using re-degraded observations.

Frequency-Domain Analysis of Re-Degradation. FPG complements SSR with observation-consistency forcing through a proxy degradation. To analyze re-degradation, we approximate the real degradation using an simplified degradation model:

$$\mathbf{y} = \mathbf{B}^{(S_1)}\mathbf{x}_{\mathrm{GT}} + \mathbf{n}, \tag{3}$$

where S_1 is the downsampling factor, $\mathbf{B}^{(S_1)}$ is an unknown downsampling operator, and $\mathbf{n}$ denotes unknown additive noise. As illustrated in Fig. 3(a), we re-degrade $\mathbf{y}$ with a known linear downsampler $\mathbf{A}^{(S_2)}$. The result approximates two successive downsampling operations:

$$\begin{aligned} \tilde{\mathbf{y}} &= \mathbf{A}^{(S_2)}\mathbf{y} \\ &= \mathbf{A}^{(S_2)}\mathbf{B}^{(S_1)}\mathbf{x}_{\mathrm{GT}} + \mathbf{A}^{(S_2)}\mathbf{n} \\ &\approx \underbrace{\mathbf{A}^{(S_2)}\mathbf{A}^{(S_1)}}_{\tilde{\mathbf{A}}}\mathbf{x}_{\mathrm{GT}}. \end{aligned} \tag{4}$$

This approximation requires $\mathbf{A}^{(S_2)}\mathbf{n} \approx \mathbf{0}$ and $\mathbf{A}^{(S_2)}\mathbf{B}^{(S_1)}\mathbf{x}_{\mathrm{GT}} \approx \mathbf{A}^{(S_2)}\mathbf{A}^{(S_1)}\mathbf{x}_{\mathrm{GT}}$. Please see Appendix C for theoretical analysis. Under these conditions, $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{A}}$ provide proxy guidance, forming the basis of learnable Fourier Pseudoinverse Guidance (FPG) and its training-free variant.

Learnable Fourier Pseudoinverse Guidance. Learnable FPG is inserted after each U-Net block to provide observation guidance in a single forward pass (Fig. 2(a)). Fig. 3(b) shows its operation at one depth. The observation $\mathbf{y}$ is linearly resized to obtain $\tilde{\mathbf{y}}$, then channel-aligned to match the current feature $\mathbf{f}_{\mathbf{x}}$, yielding $\mathbf{f}_{\tilde{\mathbf{y}}}$. Their Fourier representations are $\mathbf{F}_{\mathbf{x}} = \mathcal{F}(\mathbf{f}_{\mathbf{x}})$ and $\mathbf{F}_{\tilde{\mathbf{y}}} = \mathcal{F}(\mathbf{f}_{\tilde{\mathbf{y}}})$, where $\mathcal{F}$ denotes the spatial Fourier transform. At each forward pass, a learnable spatial-domain depth-wise convolution kernel is Fourier-transformed to obtain $\mathbf{H}(\boldsymbol{\omega})$. Its pseudoinverse $\mathbf{H}^{\dagger}$ is computed using Tikhonov regularization (Tikhonov, 1963) rather than learned independently:

$$\mathbf{H}^{\dagger}(\boldsymbol{\omega}) = \frac{\overline{\mathbf{H}(\boldsymbol{\omega})}}{|\mathbf{H}(\boldsymbol{\omega})|^2 + \epsilon^2}, \tag{5}$$

where $\boldsymbol{\omega}$ indexes frequency-domain components, the overline denotes complex conjugation, and $\epsilon > 0$ stabilizes inversion when $|\mathbf{H}(\boldsymbol{\omega})|$ is small. The resulting operator pair shares one learned kernel and guides the features in the Fourier domain:

$$\mathbf{F}_{\hat{\mathbf{x}}} = \mathbf{F}_{\mathbf{x}} + \mathbf{H}^{\dagger} \odot \big(\mathbf{F}_{\tilde{\mathbf{y}}} - \mathbf{H} \odot \mathbf{F}_{\mathbf{x}}\big), \qquad \mathbf{f}_{\hat{\mathbf{x}}} = \mathcal{F}^{-1}(\mathbf{F}_{\hat{\mathbf{x}}}), \tag{6}$$

where $\odot$ denotes element-wise multiplication. Each FPG module has its own operator pair and aligned observation features.

Training-Free Fourier Pseudoinverse Guidance FPG also provides training-free observation guidance directly in RGB space. Let $\mathbf{x}_{\mathrm{pred}}$ denote the predicted HR image. Training-free FPG is formulated as:

$$\hat{\mathbf{x}} = \mathbf{x}_{\mathrm{pred}} + \gamma\tilde{\mathbf{A}}^{\dagger}\big(\tilde{\mathbf{y}} - \tilde{\mathbf{A}}\mathbf{x}_{\mathrm{pred}}\big), \tag{7}$$

where γ controls the guidance strength and $\tilde{\mathbf{A}}$ is the proxy linear downsampler. The corresponding pseudoinverse $\tilde{\mathbf{A}}^{\dagger}$ can be computed following prior methods (Wang et al., 2023; Kawar et al., 2022).

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and degradation. Evaluation covers $\times 4$ blind panoramic super-resolution on ODISR from LAU-Net (Deng et al., 2021) and SUN 360 (Xiao et al., 2012). Real-ESRGAN degradations (Wang et al., 2021) are applied in the dual-fisheye domain to produce the training and testing sets.

Metrics. Full-reference metrics include WS-PSNR (Sun et al., 2017), WS-SSIM (Zhou et al., 2018), LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2020), and AFINE-FR (Chen et al., 2025), while FID (Heusel et al., 2017) assesses distribution-level similarity. No-reference metrics include AFINE-NR (Chen et al., 2025), Q-Align (Wu et al., 2024a), VQ-R1 (Wu et al., 2025), and OIQAND (Yan et al., 2025). As some metrics are designed for planar images, we also evaluate them on CMP faces to reduce the influence of ERP distortion.

Training settings. PanoForcing uses SD-Turbo (stabilityai; Sauer et al., 2024) with rank-4 U-Net LoRA (Hu et al., 2022). Training uses AdamW on NVIDIA A800 GPUs for 20k iterations, with total batch size 16 and learning rate 5×10^{-5}. Pretrained weights use FP16 while trainable LoRA parameters use FP32. Cubemap faces are processed at 512×512 resolution, and SSR selects $K = 2$ faces for autograd recording. Baselines share the fine-tuning data, degradation, and iteration budget. RGB L_2, LPIPS (Zhang et al., 2018), and CSD (Sun et al., 2025) losses have weights 1, 2, and 1, respectively.

Table 1: Quantitative comparison on the ODISR and SUN 360 datasets. "Proj." denotes the projection for metric evaluation. Best results are shown in **red bold**, and second-best results in blue.

Dataset	Proj.	Metric	StableSR	DiffBIR	SeeSR	OmniSSR	OSEDiff	S3Diff	InvSR	VOSR	PanoForcing
ODISR	ERP	WS-PSNR↑	22.09	22.20	22.58	22.53	22.33	21.58	22.40	21.16	**22.70**
		WS-SSIM↑	0.5696	0.5485	**0.6000**	0.5263	0.5836	0.5560	0.5961	0.5530	0.5976
		Q-Align↑	3.860	3.922	4.017	2.519	4.023	4.010	3.717	3.986	**4.171**
		VQ-R1↑	4.124	4.154	4.310	2.184	4.313	4.276	3.631	4.128	**4.385**
		OIQAND↑	3.5443	3.5604	3.5086	1.8312	3.5820	**3.6371**	3.1254	3.4394	3.5978
	CMP	LPIPS↓	0.2831	0.3251	0.3240	0.6595	0.2677	0.2636	0.3975	0.3109	**0.2530**
		DISTS↓	0.1693	0.1926	0.1899	0.3169	0.1702	**0.1617**	0.2471	0.2056	0.1630
		FID↓	57.63	72.12	62.80	117.85	63.81	55.15	105.97	60.81	**49.79**
		AFINE-FR↓	-0.702	-0.770	-0.757	1.515	-0.650	-1.146	-0.096	-0.527	**-1.249**
		AFINE-NR↓	-0.782	-0.831	-0.821	-0.305	-0.769	-0.874	-0.723	-0.781	**-0.879**
		Q-Align↑	3.335	**3.653**	3.506	1.269	3.434	3.616	2.706	3.253	3.453
		VQ-R1↑	3.786	3.857	3.919	0.954	3.974	3.973	3.031	3.706	**4.010**
SUN 360	ERP	WS-PSNR↑	22.55	22.36	22.86	22.39	22.53	22.01	22.51	21.40	**22.96**
		WS-SSIM↑	0.5975	0.5573	**0.6321**	0.5220	0.5970	0.5695	0.6191	0.5801	0.6282
		Q-Align↑	3.776	3.899	4.015	2.364	3.903	3.976	3.721	3.902	**4.134**
		VQ-R1↑	3.940	4.072	4.303	1.905	4.029	4.131	3.631	4.022	**4.316**
		OIQAND↑	3.4954	3.5184	3.4671	1.7261	3.6027	**3.6638**	3.1569	3.4834	3.6071
	CMP	LPIPS↓	0.3044	0.3462	0.3248	0.6812	0.3004	0.2784	0.4046	0.3272	**0.2623**
		DISTS↓	0.1739	0.1909	0.1855	0.3306	0.1815	0.1633	0.2394	0.1966	**0.1629**
		FID↓	65.82	75.13	66.36	136.11	70.27	60.83	101.17	71.56	**55.50**
		AFINE-FR↓	-0.457	-0.462	-0.732	1.680	-0.237	-0.870	0.049	-0.490	**-1.099**
		AFINE-NR↓	-0.874	-0.897	-0.971	-0.319	-0.833	-0.937	-0.830	-0.907	**-1.016**
		Q-Align↑	3.397	**3.849**	3.764	1.256	3.403	3.672	2.892	3.507	3.520
		VQ-R1↑	3.736	3.923	4.011	0.927	3.818	3.933	3.128	3.724	**4.045**

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

PanoForcing is compared with three groups of methods: multi-step diffusion SR models StableSR (Wang et al., 2024b), DiffBIR (Lin et al., 2024), and SeeSR (Wu et al., 2024c); one-step diffusion SR models OSEDiff (Wu et al., 2024b), S3Diff (Zhang et al., 2024), InvSR (Yue et al., 2025), and VOSR (Wu et al., 2026); and the diffusion-based ODISR model OmniSSR (Li et al., 2024).

Quantitative Comparison. Tab. 1 shows that PanoForcing leads on several full-reference and non-reference metrics across both datasets. On ODISR, it exceeds S3Diff by 1.12 dB in WS-PSNR and reduces FID by 5.36, while achieving the highest ERP Q-Align (4.171) and VQ-R1 (4.385)

scores. As shown in Tab. 2, PanoForcing has the fewest parameters and uses only 53.4% of OSEDiff's GMACs. These results show strong reconstruction fidelity and perceptual quality at a lower computational cost.

Qualitative Comparison. Fig. 4 compares PanoForcing with competing methods. In image 0000, PanoForcing reconstructs clear chair contours while preserving fine details in the background plants. In image 0008, it restores the contours and numbers of helicopter clearly.

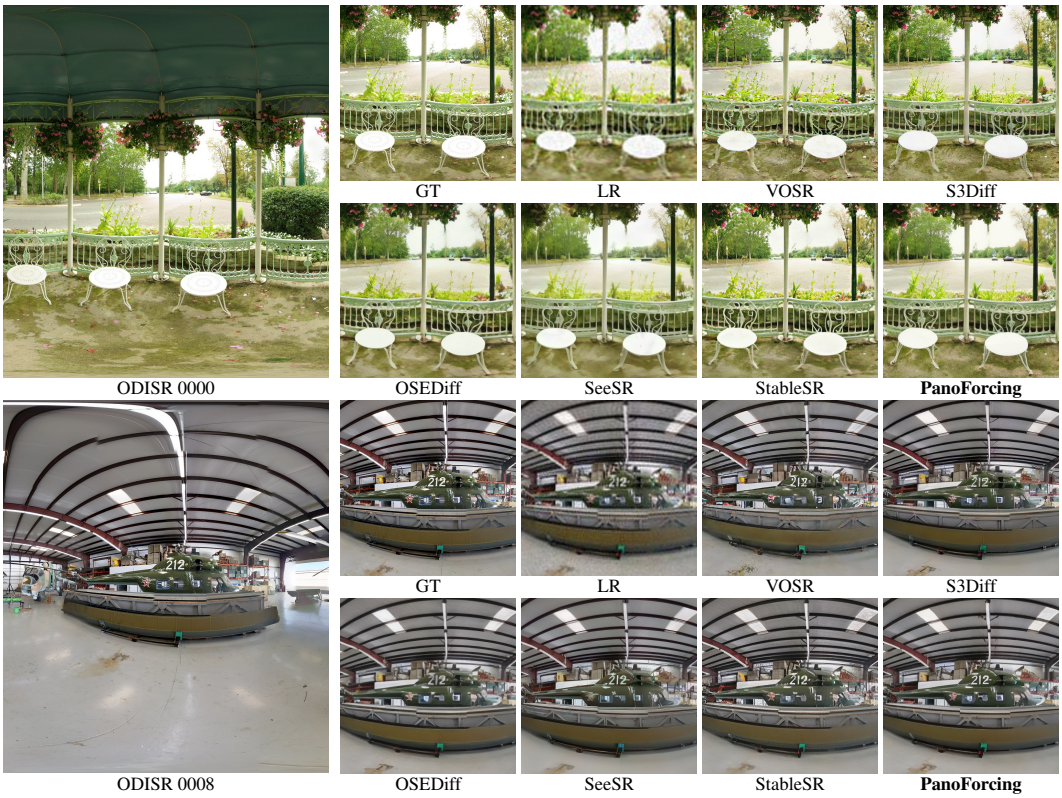


Figure 4: Qualitative comparison on ODISR scenes 0000 and 0008. PanoForcing achieves superior SR results compared to other generative methods.

4.3 Ablation and Discussion

Effects of SSR and Learnable FPG. Tab. 3 shows that learnable FPG improves reconstruction fidelity, while adding SSR further enhances overall quality. Together, observation-consistency forcing (FPG) and full-context forcing (SSR) provide complementary benefits, achieving the best results on most metrics.

Model Complexity and Reconstruction Quality. PanoForcing has the fewest parameters and the lowest computational cost among the compared methods (Tab. 2). Spherical Octant Attention reuses pretrained attention parameters, while learnable FPG shares a kernel between its paired operators, keeping parameter and computation overhead low. The visual comparison in Fig. 6 highlights PanoForcing's advantages in both model complexity and super-resolution quality.

Training and Inference Memory Scalability. Fig. 5 demonstrates two independent memory controls. During training, the retained-face budget K determines how many intra-face paths record autograd histories. During inference, the number of concurrently processed faces controls temporary activation memory. Both settings preserve all six faces in the forward pass. Across the evaluated configurations, peak training memory ranges from 8.7 to 61.7 GiB, while peak inference memory ranges from 3.5 to 7.9 GiB.

Table 2: Model inference complexity in terms of parameter count and MACs. Best results shown in **red**.

Method	Params (M)↓	TMACs↓
StableSR	1554.64	814.17
DiffBIR	1683.45	1167.23
SeeSR	2510.76	1410.68
OmniSSR	1554.64	8046.97
OSEDiff	1765.30	25.40
S3Diff	1326.77	35.08
VOSR	1745.67	27.03
PanoForcing	**1305.99**	**13.57**

Table 3: Ablation study of SSR and learnable FPG. Best and second-best results shown in **red** and blue, respectively.

SSR	✗	✗	✓
Learnable FPG	✗	✓	✓
WS-PSNR↑	22.89	22.97	**23.01**
WS-SSIM↑	0.6293	**0.6318**	0.6315
FID↓	56.93	56.54	**55.91**
VQ-R1↑	4.232	4.225	**4.259**
Q-Align↑	3.083	3.122	**3.477**
OIQAND↑	3.5009	3.5270	**3.5995**

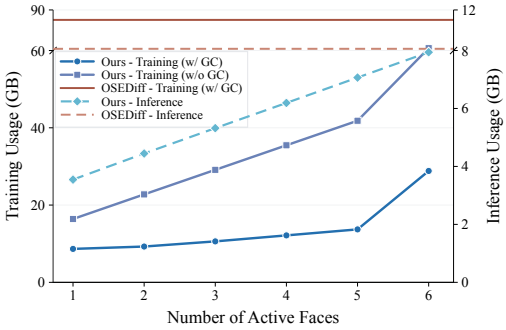


Figure 5: PanoForcing can dynamically adjust training and inference memory usage. (GC: Gradient Checkpointing)

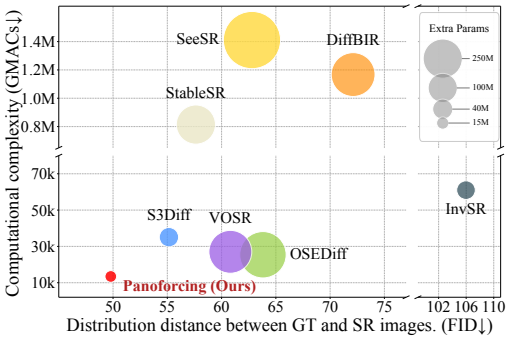


Figure 6: PanoForcing achieves better reconstruction quality with lower computational cost and fewer parameters.

Number of Faces Selected in SSR. Tab. 4 compares different numbers of faces selected for autograd recording in SSR. Increasing K from 1 to 2 improves all reported metrics, with notable gains in LPIPS and FID. Further increases yield smaller overall gains and some variation across metrics. This suggests that selecting two or more faces may already provide sufficient gradient connections for effective learning in the cross-face blocks. Based on these observations, $K = 2$ is recommended to balance training memory and reconstruction quality.

Table 4: Effect of the number of faces selected in SSR on SUN 360. Increasing K from 1 to 2 improves all metrics, with limited further gains at larger K.

K	WS-PSNR↑	WS-SSIM↑	LPIPS↓	FID↓	AFINE-NR↓	Q-Align↑	VQ-R1↑
1	22.85	0.6224	0.3135	62.57	-0.841	3.861	3.905
2	23.01	0.6315	0.2715	55.91	-0.977	4.055	4.259
3	23.08	0.6338	0.2712	55.69	-0.998	4.084	4.306
4	23.01	0.6306	0.2680	55.19	-1.004	4.089	4.322

5 Conclusion

PanoForcing unifies full-context forcing and observation-consistency forcing for one-step blind panoramic super-resolution. SSR preserves complete panoramic context while selectively recording autograd subgraphs to control training memory. Face-wise parallelism separately adjusts inference memory, while FPG guides reconstruction using the LR observation within a single network pass. Experiments on ODISR and SUN 360 demonstrate strong reconstruction fidelity and perceptual quality with lower computational cost and fewer parameters than the compared methods. Together, these results show that full-context panoramic reconstruction can be achieved under configurable training and inference memory budgets.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, polish writing, and draft sections of the paper from author-provided material. The mathematical proofs and interpretations of experimental results were provided by the authors in Chinese, and AI tools assisted with their English presentation and notation. We also used these tools for retrieval and discovery, including finding related work, checking bibliographic information, and formatting references. Additional assistance included manuscript organization, figure preparation, code review, formatting experimental records, and LaTeX editing and compilation checks. These tools were not used to run model experiments. We have reviewed all AI-assisted work, checked the edited mathematical text against the authors' derivations, checked result descriptions and table entries against the original analyses and experimental records, and verified the references. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Hao Ai, Zidong Cao, Jinjing Zhu, Haotian Bai, Yucheng Chen, and Ling Wang. Deep learning for omnidirectional vision: A survey and new perspectives. *arXiv preprint arXiv:2205.10468*, 2022.

Qing Cai, Mu Li, Dongwei Ren, Jun Lyu, Haiyong Zheng, Junyu Dong, and Yee-Hong Yang. Spherical pseudo-cylindrical representation for omnidirectional image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang. Toward generalized image quality assessment: Relaxing the perfect reference quality assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Xin Deng, Hao Wang, Mai Xu, Yichen Guo, Yuhang Song, and Li Yang. LAU-Net: Latitude adaptive upscaling network for omnidirectional image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 (TPAMI)*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Runyi Li, Xuhan Sheng, Weiqi Li, and Jian Zhang. Omnissr: Zero-shot omnidirectional image super-resolution using stable diffusion model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Leiming Liu, Ting Luo, Gangyi Jiang, Yeyao Chen, Haiyong Xu, Renzhi Hu, and Zhouyan He. Diffosr: Latitude-aware conditional diffusion probabilistic model for omnidirectional image super-resolution. *Knowledge-Based Systems (KBS)*, 2025.

Leiming Liu, Ting Luo, Yeyao Chen, Gangyi Jiang, Haiyong Xu, Renzhi Hu, and Zhouyan He. Osrlldm: Omnidirectional image super-resolution with latitude-aware latent diffusion models. *Information Fusion (IF)*, 2026.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Huicheng Pi, Senmao Tian, Ming Lu, Jiaming Liu, Yandong Guo, and Shunli Zhang. A comprehensive comparison of projections in omnidirectional super-resolution. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Min Zhang, Dong Li, and Yuxiong He. ZeRO-Offload: Democratizing billion-scale model training. In *Proceedings of the USENIX Annual Technical Conference (USENIX ATC)*, 2021.

Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alex Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Xuhan Sheng, Runyi Li, Bin Chen, Weiqi Li, Xu Jiang, and Jian Zhang. RealOSR: Latent guidance boosts diffusion-based real-world omnidirectional image super-resolutions. *arXiv preprint arXiv:2412.09646*, 2024.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

stabilityai. Sd-turbo. URL `https://huggingface.co/stabilityai/sd-turbo`.

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

Xiaopeng Sun, Weiqi Li, Zhenyu Zhang, Qiufang Ma, Xuhan Sheng, Ming Cheng, Haoyu Ma, Shijie Zhao, Jian Zhang, Junlin Li, and Li Zhang. Opdn: Omnidirectional position-aware deformable network for omnidirectional image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023.

Yule Sun, Ang Lu, and Lu Yu. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE Signal Processing Letters (SPL)*, 2017.

A. N. Tikhonov. On the solution of ill-posed problems and the method of regularization. *Doklady Akademii Nauk SSSR (DAN SSSR)*, 1963.

Jiangang Wang, Yuning Cui, Yawen Li, Wenqi Ren, and Xiaochun Cao. Omnidirectional image super-resolution via bi-projection fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024a.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision (IJCV)*, 2024b.

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

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C. Kot, and Bihan Wen. SinSR: Diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024c.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024a.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024c.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. VisualQuality-R1: Reasoning-induced image quality assessment via reinforcement learning to rank. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jianxiong Xiao, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Jiebin Yan, Jiale Rao, Xuelin Liu, Yuming Fang, Yifan Zuo, and Weide Liu. Subjective and objective quality assessment of non-uniformly distorted omnidirectional images. *IEEE Transactions on Multimedia (TMM)*, 2025.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024a.

Zhixiong Yang, Jingyuan Xia, Shengxi Li, Xinghua Huang, Shuanghui Zhang, Zhen Liu, Yaowen Fu, and Yongxiang Liu. A dynamic kernel prior model for unsupervised blind image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Youngho Yoon, Inchul Chung, Lin Wang, and Kuk-Jin Yoon. SphereSR: 360° image super-resolution with arbitrary projection via continuous spherical image representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Fanghua Yu, Xintao Wang, Mingdeng Cao, Gen Li, Ying Shan, and Chao Dong. OSRT: Omnidirectional image super-resolution with distortion-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yufeng Zhou, Mei Yu, Hualin Ma, Hua Shao, and Gangyi Jiang. Weighted-to-spherically-uniform ssim objective quality evaluation for panoramic video. In *Proceedings of the IEEE International Conference on Signal Processing (ICSP)*, 2018.

Appendix

A Detailed Experimental Settings

A.1 Data and Optimization

Experiments evaluate $\times 4$ blind panoramic super-resolution on ODISR and SUN 360. Following RealOSR (Sheng et al., 2024), Real-ESRGAN degradations (Wang et al., 2021) are applied in the dual-fisheye domain. Aligned LR and HR panoramas are projected onto six cubemap faces for training.

PanoForcing uses SD-Turbo (stabilityai) with rank-4 LoRA adapters (Hu et al., 2022) in the U-Net. Tab. S1 summarizes the settings used for the main quantitative comparison. Baselines share the fine-tuning data, degradation pipeline, and iteration budget. The loss combines RGB L_2, LPIPS (Zhang et al., 2018), and CSD (Sun et al., 2025) with weights 1, 2, and 1, respectively.

Table S1: Training settings for the main quantitative comparison.

Setting	Value
Backbone	SD-Turbo
LoRA rank	4
Optimizer	AdamW
Learning rate	5×10^{-5}, fixed
Iterations	20,000
Total batch size	16
Precision	Mixed: FP16 weights, FP32 LoRA
Face resolution	512×512
Retained faces K	2
Hardware	NVIDIA A800
Loss weights (L_2, LPIPS, CSD)	1, 2, 1

The reported model uses the best checkpoint from the 20k-iteration training run. Per-device batch size and gradient accumulation can be adjusted to match the stated total batch size.

A.2 Ablation Settings

The component ablation and face-selection study use 10k training iterations for rapid configuration search, rather than the 20k iterations used for the main comparison. Their precision settings also differ as described below. The resulting scores should therefore be compared within each ablation table rather than directly with the main results.

Number of faces selected in SSR. The face-selection study in Tab. 4 uses FP32 for both pretrained model weights and LoRA parameters across all values of K. With FP16 pretrained weights and FP32 LoRA parameters, training at $K = 1$ exhibited numerical overflow, producing NaN outputs and contaminating the gradients. FP32 weights were therefore used throughout this sweep to compare different values of K under the same precision setting.

SSR and learnable FPG. The component ablation in Tab. 3 uses FP16 pretrained weights and FP32 LoRA parameters, with $K = 2$ when SSR is enabled. The overflow observed at $K = 1$ did not occur at $K = 2$ in these experiments, allowing mixed-precision training for this study.

A.3 Cubemap Projection Details

The CMP representation in the main paper uses the equi-angular cubemap (EAC) variant. Unlike standard CMP, which samples uniformly on each cube face, EAC samples uniformly in angular coordinates. This reduces stretching near face boundaries while preserving the six-face structure. Each EAC face has a resolution of 512×512 pixels.

B Training and Inference Memory

Tab. S2 reports the complete memory sweeps, including training with and without gradient checkpointing.

Table S2: Peak memory in GB for separate sweeps over the retained-face count K during training and the parallel face count B during inference. GC denotes gradient checkpointing.

K	Training, no GC	Training, GC	B	Inference
1	16.395	8.657	1	3.546
2	22.766	9.265	2	4.446
3	29.118	10.601	3	5.321
4	35.474	12.153	4	6.191
5	41.824	13.705	5	7.066
6	61.741	28.829	6	7.941

At the recommended $K = 2$, gradient checkpointing reduces peak training memory from 22.766 to 9.265 GB. Cross-face blocks and shared features also consume memory, so the total does not scale simply as $K/6$ of the full-recording configuration.

All values report peak allocated GPU memory. Measurements use NVIDIA A800 GPUs with the mixed-precision setting in Tab. S1 and 512×512 faces.

C Theoretical Analysis of FPG

C.1 Setup and Re-Degradation

Following the main paper, the unknown degradation is approximated by

$$\mathbf{y} = \mathbf{B}^{(S_1)}\mathbf{x}_{\mathrm{GT}} + \mathbf{n}. \tag{S1}$$

To simplify the noise analysis, the additive noise is assumed to be independent and Gaussian, $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Here $\mathbf{B}^{(S_1)}$ is an unknown linear downsampling operator, while $\mathbf{A}^{(S_1)}$ and $\mathbf{A}^{(S_2)}$ are known linear downsamplers. The factors S_1 and S_2 correspond to the original degradation and re-degradation, respectively. Images are vectorized when used in matrix expressions.

The re-degraded observation and composite proxy operator are

$$\tilde{\mathbf{y}} = \mathbf{A}^{(S_2)}\mathbf{y}, \qquad \tilde{\mathbf{A}} = \mathbf{A}^{(S_2)}\mathbf{A}^{(S_1)}. \tag{S2}$$

Their discrepancy separates into degradation mismatch and filtered noise:

$$\begin{aligned} \tilde{\mathbf{y}} - \tilde{\mathbf{A}}\mathbf{x}_{\mathrm{GT}} &= \boldsymbol{\Delta} + \mathbf{A}^{(S_2)}\mathbf{n}, \\ \boldsymbol{\Delta} &= \mathbf{A}^{(S_2)}\big(\mathbf{B}^{(S_1)} - \mathbf{A}^{(S_1)}\big)\mathbf{x}_{\mathrm{GT}}. \end{aligned} \tag{S3}$$

The following analysis explains how re-degradation can reduce both terms.

C.2 Noise Attenuation by Averaging

To simplify the noise analysis, consider average downsampling as the re-degradation operator. Each output pixel averages M input pixels, so its noise component satisfies

$$\bar{n} = \frac{1}{M}\sum_{i=1}^{M} n_i, \qquad \mathbb{E}[\bar{n}] = 0, \qquad \mathrm{Var}(\bar{n}) = \frac{1}{M^2}\sum_{i=1}^{M}\sigma^2 = \frac{\sigma^2}{M}. \tag{S4}$$

Thus, $\bar{n} \sim \mathcal{N}(0, \sigma^2/M)$, so averaging reduces the noise standard deviation from σ to $\sigma/\sqrt{M}$. For one-dimensional average downsampling, $M = S_2$. A two-dimensional $S_2 \times S_2$ averaging window instead has $M = S_2^2$, giving variance σ^2/S_2^2.

More generally, a weighted average $\bar{n} = \sum_i w_i n_i$ has variance $\sigma^2 \sum_i w_i^2$. Normalized nonnegative weights spread across multiple pixels therefore attenuate independent noise by an amount determined by the kernel.

C.3 Frequency-Domain Analysis of Degradation Mismatch

Downsampling model. The linear downsamplers are modeled as low-pass filtering followed by decimation,

$$\mathbf{A}^{(S)} = D_S H_{A,S}, \qquad \mathbf{B}^{(S)} = D_S H_{B,S}, \tag{S5}$$

where $H_{A,S}$ and $H_{B,S}$ are convolution operators and D_S retains samples at stride S. The two degradation paths share sampling coordinates but use different filters. To simplify the frequency-domain analysis, the derivation starts with one-dimensional signals and then extends to two-dimensional images.

One-dimensional derivation. Let $x \in \mathbb{R}^N$ denote a one-dimensional signal and define its unitary DFT as

$$X[k] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n] e^{-\mathrm{i}2\pi kn/N}, \qquad k = 0, \dots, N-1. \tag{S6}$$

Parseval's identity gives $\|x\|_2^2 = \sum_k |X[k]|^2$. To analyze the filtering effect on a common grid, lift the second-stage kernel to the original grid by inserting $S_1 - 1$ zeros between its taps. Denote this convolution operator by $\bar{H}_{A,S_2}$. For integer-aligned decimation and circular convolution,

$$H_{A,S_2} D_{S_1} = D_{S_1} \bar{H}_{A,S_2}. \tag{S7}$$

Consequently, the full mismatch and its pre-decimation counterpart obey

$$\Delta = D_{S_1 S_2} \Delta_{\mathrm{lp}}, \qquad \Delta_{\mathrm{lp}} = \bar{H}_{A,S_2}(H_{B,S_1} - H_{A,S_1})x. \tag{S8}$$

Lifting the second-stage kernel from the first-stage sampling grid places both filters on common frequency coordinates.

Let $a_1[k]$, $b_1[k]$, and $\bar{a}_2[k]$ be the convolution eigenvalues of these filters. Under the unitary DFT convention, each eigenvalue equals $\sqrt{N}$ times the unitary DFT of the corresponding padded kernel. Convolution therefore yields

$$\widehat{\Delta}_{\mathrm{lp}}[k] = \bar{a}_2[k]\big(b_1[k] - a_1[k]\big) X[k]. \tag{S9}$$

Applying Parseval's identity gives the exact pre-decimation error energy:

$$\|\Delta_{\mathrm{lp}}\|_2^2 = \sum_{k=0}^{N-1} |\bar{a}_2[k]|^2 |b_1[k] - a_1[k]|^2 |X[k]|^2. \tag{S10}$$

Thus, the filter mismatch is weighted by both the re-degradation filter and the image spectrum. Since ordinary decimation selects a subset of samples, $\|\Delta\|_2 \le \|\Delta_{\mathrm{lp}}\|_2$.

Low- and high-frequency bands. Partition the frequency indices into disjoint bands Ω_L and Ω_H. For $q \in \{L, H\}$, define

$$\alpha_q = \max_{k \in \Omega_q} |b_1[k] - a_1[k]|, \quad \beta_q = \max_{k \in \Omega_q} |\bar{a}_2[k]|, \quad E_q = \sum_{k \in \Omega_q} |X[k]|^2. \tag{S11}$$

Splitting the sum in Eq. S10 and bounding each factor within its band gives

$$\|\Delta\|_2^2 \le \|\Delta_{\mathrm{lp}}\|_2^2 \le \beta_L^2 \alpha_L^2 E_L + \beta_H^2 \alpha_H^2 E_H. \tag{S12}$$

In a shared low-frequency passband, normalized downsampling filters can have similar responses near unity, making α_L small. Beyond this band, the second filter can suppress part of the mismatch, while spectral decay in natural images reduces E_H. Because the lifted filter has periodic spectral replicas, its high-frequency suppression depends on the first-stage anti-aliasing filters and signal energy, as captured by the bound.

Extension to two-dimensional images. For an $H \times W$ image, replace k by (u, v) and use the unitary two-dimensional DFT. The corresponding identity is

$$\|\Delta_{\mathrm{lp}}\|_2^2 = \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} |\bar{a}_2[u,v]|^2 |b_1[u,v] - a_1[u,v]|^2 |X[u,v]|^2. \tag{S13}$$

The band decomposition and bound in Eq. S12 follow identically, with energies summed over two-dimensional frequency bands and channels. These expressions assume aligned integer sampling and convolution-compatible boundary handling. Interpolation at noninteger coordinates requires accounting for the corresponding resampling operator.

C.4 Implication for Proxy Guidance

Combining noise attenuation with the mismatch analysis gives

$$\|\tilde{\mathbf{y}} - \tilde{\mathbf{A}}\mathbf{x}_{\mathrm{GT}}\|_2 \le \sqrt{\beta_L^2\alpha_L^2 E_L + \beta_H^2\alpha_H^2 E_H} + \|\mathbf{A}^{(S_2)}\mathbf{n}\|_2. \tag{S14}$$

When the filtered noise and spectrally weighted mismatch are small, $\tilde{\mathbf{y}} \approx \tilde{\mathbf{A}}\mathbf{x}_{\mathrm{GT}}$. The re-degraded observation $\tilde{\mathbf{y}}$ and known composite operator $\tilde{\mathbf{A}}$ can therefore provide proxy guidance without perfectly matching the unknown degradation. This motivates both learnable FPG in feature space and its training-free RGB variant.

D Training-Free FPG Experiments

Training-free FPG is evaluated on OSEDiff and PanoForcing to compare its effect on a planar baseline and a model that already includes learnable observation guidance.

D.1 Training-Free FPG on Planar Super-Resolution

We evaluate training-free FPG on OSEDiff, a planar baseline also included in our main comparisons. Tab. S3 varies the re-degradation scale s and guidance strength γ using bicubic operators on DIV2K-val, DRealSR, and RealSR. The original OSEDiff output serves as the no-guidance reference ($\gamma = 0$). No model is fine-tuned.

Table S3: Training-free FPG on OSEDiff with bicubic re-degradation at different scales s and guidance strengths γ. Moderate guidance improves several full-reference metrics, while stronger guidance can reduce perceptual quality. The original output is shown at $\gamma = 0$.

Dataset	s	γ	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑	FID↓
DIV2K-val	–	0.00	23.72	0.6109	0.2942	0.1975	67.96	26.34
	4	0.25	23.95	0.6117	0.2927	0.1959	67.79	25.93
	4	0.50	24.12	0.6122	0.2925	0.1950	67.56	26.09
	4	0.75	24.23	0.6125	0.2942	0.1947	67.23	27.00
	4	1.00	24.27	0.6124	0.2974	0.1953	66.84	28.66
	8	0.50	24.06	0.6119	0.2922	0.1956	67.72	26.04
DRealSR	–	0.00	27.92	0.7836	0.2966	0.2163	64.69	135.41
	4	0.25	28.25	0.7841	0.2954	0.2145	64.51	133.49
	4	0.50	28.50	0.7845	0.2948	0.2135	64.36	132.63
	4	1.00	28.64	0.7843	0.2969	0.2128	63.91	135.35
RealSR	–	0.00	25.15	0.7341	0.2921	0.2128	69.09	123.49
	2	0.50	25.77	0.7364	0.2885	0.2088	68.85	116.60
	2	1.00	25.92	0.7345	0.2893	0.2119	68.42	116.68
	4	0.25	25.38	0.7350	0.2906	0.2112	68.95	121.64
	4	0.50	25.54	0.7357	0.2896	0.2104	68.83	120.96
	4	1.00	25.59	0.7359	0.2896	0.2104	68.54	121.84
	8	0.25	25.34	0.7347	0.2910	0.2118	68.97	122.67
	8	0.50	25.47	0.7352	0.2902	0.2111	68.89	121.97
	8	0.75	25.61	0.7360	0.2893	0.2101	68.70	120.94
	8	1.00	25.51	0.7355	0.2898	0.2108	68.77	123.23

Compared with the original OSEDiff outputs, training-free FPG improves several full-reference metrics across the tested scales and guidance strengths. These gains suggest that generating realistic high-frequency detail does not necessarily preserve the underlying low-frequency image content. Guidance from the re-degraded observation can help reduce such inconsistencies.

Stronger guidance, however, can worsen perceptual quality. LPIPS and FID deteriorate in some configurations, while MUSIQ decreases as γ increases. This behavior is consistent with errors in the proxy degradation relation. Excessive guidance can amplify these errors and push the reconstruction

away from the desired image. A moderate guidance strength therefore provides a better balance between fidelity and perceptual quality.

D.2 Incremental Effect on PanoForcing

Tab. S4 evaluates additional training-free FPG on PanoForcing using the SUN 360 test set, checkpoint 10001, and bicubic re-degradation with $s = 4$. RGB guidance is applied to the reconstructed ERP image using the LR ERP observation. DISTS is evaluated on CMP faces, while VQ-R1 and OIQAND are evaluated on ERP images.

Table S4: Additional training-free FPG on PanoForcing with RGB guidance applied to the reconstructed ERP image.

γ	WS-PSNR↑	WS-SSIM↑	DISTS↓	VQ-R1↑	OIQAND↑
0.00	23.06	0.6357	0.1658	4.280	3.5847
0.25	23.12	0.6356	0.1655	4.218	3.5779
0.50	23.15	0.6350	0.1685	4.140	3.5524
0.75	23.14	0.6341	0.1731	4.010	3.5027
1.00	23.10	0.6328	0.1787	3.882	3.4326

In contrast to OSEDiff, PanoForcing shows little additional benefit from training-free FPG. Small improvements in WS-PSNR and DISTS are accompanied by lower WS-SSIM and poorer scores on other perceptual metrics, yielding no overall quality improvement.

PanoForcing already incorporates learnable FPG at multiple network depths. Its learned observation guidance may leave little room for improvement from an additional training-free RGB-space update, consistent with the limited gains observed here. The main comparisons therefore use learnable FPG without the optional training-free guidance.