Fill2SR: Repurposing Inpainting Diffusion Transformers for Real-World Super-Resolution

Xingfu Yi[1] and Xiaoxue Yu[2]

[1] Independent Researcher, Hangzhou, China
yixingfu.research@gmail.com
[2] Zhejiang University, Hangzhou, China

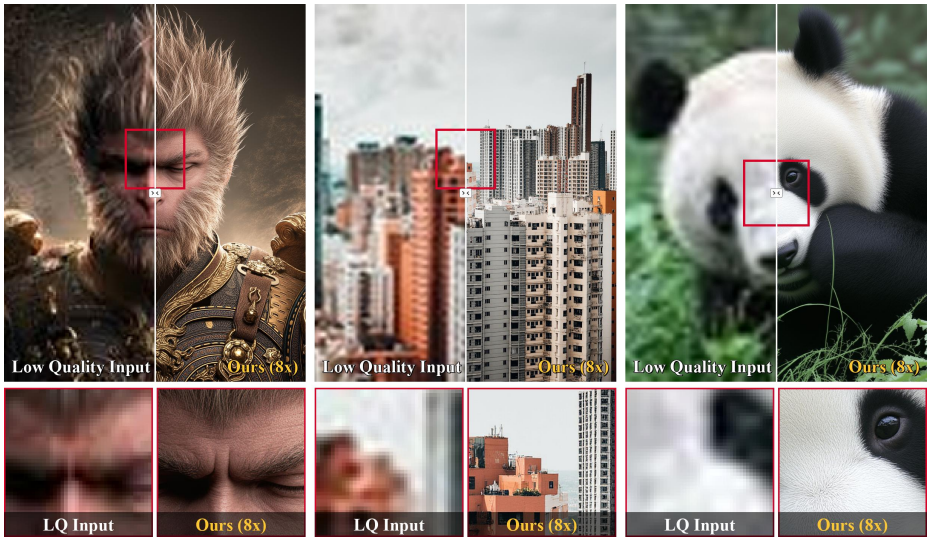

Fig. 1: Visual highlights on *RealLQ250* at 8× (real-world SR). Fill2SR produces artifact-suppressed, visually coherent high-resolution results across diverse content: non-rigid textures (left, Wukong), rigid urban structures (middle, Cityscapes), and fine biological details (right, Panda). *Bottom:* zoom-ins from the red boxes.

Abstract. Recent real-world image super-resolution (SR) methods often adapt text-to-image (T2I) backbones with ControlNet-style branches or spatial conditioning tokens, which increases memory and computes with resolution and often constrains training to a fixed scale. We propose **Fill2SR**, which repurposes a masked-inpainting Diffusion Transformer for SR without extra spatial branches. Our Inpainting-Interface Evidence Adapter (**IIEA**) writes the low-quality (LQ) observation into the native masked-image slot under a full-image mask, turning inpainting into a reverse-degradation conditional rectified flow trained with LoRA-only tuning. We further introduce **RCDT**, an offline pipeline that distills

degradation descriptors from unpaired real images and transfers them onto clean targets using frozen open-source models. Fill2SR supports mixed-resolution training up to QHD and yields stable performance across 512/1024/2048 outputs. On synthetic benchmarks, our base model with IIEA achieves the best LPIPS on DIV2K and LSDIR; adding RCDT trades a small LPIPS drop for consistently stronger no-reference quality on RealLQ250 and RealPhoto60. Fill2SR remains memory-predictable, running 1536^2 inference on a single 32GB GPU and extending to multi-megapixel outputs via tiled restoration.

Keywords: Real-World Super-Resolution · Image Restoration · Inpainting Diffusion Transformers · Rectified Flow Matching · Adapter Tuning

1 Introduction

Image super-resolution (SR) and restoration recover high-quality (HQ) images from low-quality (LQ) observations corrupted by blur, noise, compression, and complex real-world degradations. Despite strong discriminative restorers such as SwinIR [22], Restormer [47], and Uformer [39], distortion-driven training often follows the perception-distortion trade-off and produces over-smoothed details under severe degradations [4]. To train these models, previous approaches predominantly rely on artificially synthesized LQ-HQ data pairs. While synthetic degradation pipelines (e.g., BSRGAN [48], Real-ESRGAN [37]) offer scalable supervision for this paradigm, they fundamentally struggle to cover the long tail of complex corruptions in the wild.

Generative diffusion models [8,11,29] provide strong image priors and have become a leading approach for high-fidelity restoration. However, synthesis-oriented text-to-image (T2I) backbones are not optimized for pixel-aligned control, making them prone to geometric drift or spurious artifacts under severe degradations. Consequently, recent approaches typically train ControlNet-style branches [5,42,45] or introduce additional spatial conditioning tokens [2,7,9]. While these workarounds offer improvements, T2I-centric restoration pipelines face two structural bottlenecks. First, the memory/compute overhead induced by extra branches or tokens grows superlinearly with resolution (often quadratically in the number of tokens), strictly limiting the training resolution. Second, due to memory limits [5, 9, 45], the control module is typically trained at a specific resolution (e.g., 1024^2). It matches the backbone's prior at this specific resolution but fails to generalize. As a result, these methods achieve good restoration quality at their training resolution but suffer from severe hallucinations at different scales (see Tab. 2, Fig. 3, and Supplementary Material for per-degradation quantitative comparisons).

In this work, we advocate building real-world restoration on an interface that is natively pixel-aligned. We observe that modern mask-conditioned inpainting Diffusion Transformers (DiTs) [18] are trained to preserve known pixels while completing unknown regions. This induces a strong bias toward spatially consistent completion, making inpainting backbones a natural match for SR/restoration, since these tasks require respecting the LQ observation as strict pixel-level

evidence rather than weak high-level hints. We therefore present **Fill2SR**, a low-overhead adaptation framework that repurposes a pre-trained inpainting DiT for SR and real-world restoration. Our key component is IIEA (Inpainting-Interface Evidence Adapter), which repurposes the native masked-image slot as a dense, pixel-aligned evidence channel under a full-image mask, avoiding ControlNet-style branches or resolution-growing spatial conditioning token streams. We adapt the inpainting prior to the endpoint sampling of a reverse-degradation conditional rectified flow with LoRA-only [13] tuning. This flow starts from pure noise and ends at the HQ endpoint of an implicit degradation model, formulated via Rectified Flow Matching (RFM) [24, 26]. Because real-world degradations are diverse and paired LQ-HQ supervision is scarce, synthetic degradations alone may not cover the long tail and can introduce domain bias. To address this cost-effectively, we introduce **RCDT** (Reference-Conditioned Degradation Transfer), an offline paired-synthesis pipeline that distills degradation descriptors from an unpaired real-LQ pool and transfers them onto clean HQ targets using frozen open-source models [3, 40]. To further bound test-time conditioning cost, we replace heavyweight vision-language prompting [25] with a lightweight image embedder (Redux) [19], stabilizing semantic control with fixed-length visual-semantic tokens rather than computationally expensive test-time image prompting.

Extensive experiments show that Fill2SR achieves strong perceptual quality on synthetic benchmarks: our base model with IIEA and trained with synthetic only data [5, 37] attains the best LPIPS on DIV2K [1] and LSDIR [20]. Incorporating RCDT-generated pairs in the final 20% of training improves no-reference perceptual metrics [15, 35, 44] on RealLQ250 [2] and RealPhoto60 [5, 45], trading a small drop in synthetic LPIPS. In addition, the bounded conditioning cost of IIEA enables mixed-resolution, mixed-aspect-ratio training up to native QHD (2560×1440), while still fitting within a single 32GB GPU at 1536^2 for inference.

Contributions. Our main contributions are:

- **Structural shift from T2I with control to Inpainting with IIEA.** We repurpose mask-conditioned inpainting DiTs for SR/restoration by leveraging their natively pixel-aligned inpainting interface. Specifically, IIEA (Inpainting-Interface Evidence Adapter) repurposes the masked-image slot as a dense evidence channel under a full-image mask, avoiding ControlNet-style branches or resolution-growing spatial token streams while enabling stable mixed-resolution training across scales.
- **Reverse-degradation flow for domain shift.** We repurpose the backbone's native Rectified Flow Matching (RFM) objective. While originally trained for local region completion conditioned on preserved surroundings, we shift this formulation to a full-image reverse-degradation flow. This flow models the trajectory from pure noise to the clean HQ endpoint of an implicit degradation model.
- **Paired reverse-degradation supervision.** To efficiently learn this reverse-degradation flow under realistic degradations, we introduce RCDT, an offline

data generation pipeline that distills real-world degradation statistics from unpaired LQ images and creates "real-world" LQ-HQ training pairs from clean HQ images at a low cost.

2 Related Work

Discriminative and diffusion-based restoration. Classical methods map LQ to HQ images using CNNs or Transformers [22, 39, 47], often relying on synthetic degradations [21, 37, 48]. To overcome over-smoothing in deterministic models, diffusion priors are widely adopted for photo-realistic generation, evolving from iterative refinement [33] to highly efficient variants [38, 41, 46].

Adapting T2I priors for restoration. To tackle real-world degradations, recent methods repurpose large T2I backbones, whether UNet-based [23, 30, 31, 36, 45] or DiT-based [2, 9]. To enforce fidelity, they typically introduce extra conditioning stacks (*e.g.*, ControlNet [49]) or inject LQ guidance directly into the attention mechanism. However, these structural modifications inevitably inflate system complexity, introduce additional *spatial* token streams or control branches, and often remain brittle under severe degradations.

Inpainting models as a restoration foundation. Departing from complex control modules, diffusion inpainting naturally provides a pixel-aligned *(mask, context)* interface. While explored in UNet models [27, 32, 43] and early DiTs [50], repurposing massive state-of-the-art inpainting DiTs [18] for blind restoration remains underexplored. Our work capitalizes on this by repurposing the masked-image slot as an evidence channel. Combined with rectified-flow learning [6, 24, 26] and the conditioning design of fixed length tokens (IIEA and optional Redux [19]), we achieve memory-predictable, high-resolution restoration that bypasses the token-length inflation plaguing attention-modified DiTs.

3 Method

Problem setup. We formulate the problem and our method in the VAE (ε) [16,31] latent space for simplicity. Given an LQ image Y in normalized [-1, 1] RGB space and its latent $y = \varepsilon(Y) \in \mathbb{R}^{c \times \frac{h}{s} \times \frac{w}{s}}$, our goal is to get the HQ image latent $\mathbf{x} \in \mathbb{R}^{c \times h \times w}$ after an SR scale factor of s. We first upscale the LQ input so that $y \in \mathbb{R}^{c \times h \times w}$. We consider there exists a implicit degradation model $\mathcal{T}$ that y is generated by $\mathbf{y} = \mathcal{T}(\mathbf{x})$. Since we only have y as input and our goal is to get $\mathbf{x}$, our problem is to solve the function of the reverse degradation model so that

$$\mathbf{x} = \mathcal{T}^{-1}(y) \tag{1}$$

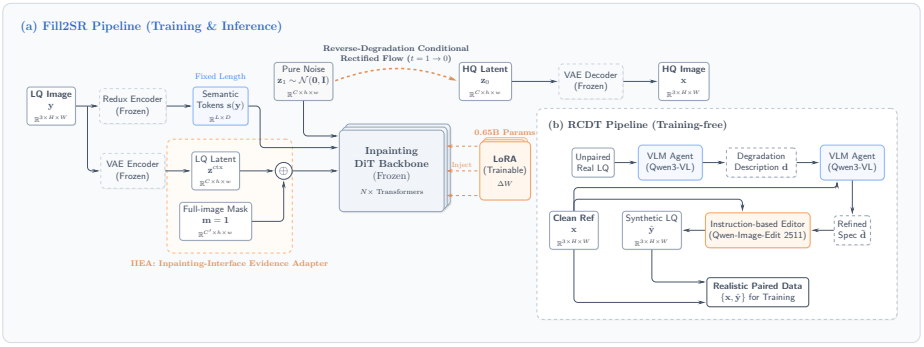

Fig. 2: Overview of the Fill2SR architecture. (a) The **Main Restoration Pipeline** utilizes the proposed IIEA to inject VAE-encoded LQ evidence into the native FLUX-Fill inpainting interface via a full-image mask ($\mathbf{m} = \mathbf{1}$). Sampling follows a conditional rectified flow trajectory from pure noise $\mathbf{z}_1$ to the restored HQ latent $\mathbf{z}_0$. (b) The **Offline RCDT Pipeline** leverages multi-modal agents (e.g., Qwen3-VL) to perceive real-world degradations and generate refined specifications, driving an instruction-based editor to synthesize realistic training pairs $\{\mathbf{x}, \hat{\mathbf{y}}\}$ with zero test-time overhead.

3.1 Reverse-Degradation Conditional Rectified Flow

Since directly modeling the real-world degradation operator $\mathcal{T}$ is intractable, we utilize Rectified Flow Matching (RFM) [24, 26] to learn an implicit reverse-degradation mapping. Concretely, we build a conditional rectified flow that transports pure Gaussian noise to the HQ latent $\mathbf{x}$, conditioned on the degraded observation $\mathbf{y} = \mathcal{T}(\mathbf{x})$. Following RFM, we sample an interpolation time $t \sim \mathrm{Unif}(0, 1)$ and form a linear path between $\mathbf{x}$ and $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\begin{aligned} \mathbf{z}_t &= (1-t)\mathbf{x} + t\mathbf{z}_1, \qquad \mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad t \sim \mathrm{Unif}(0, 1), \\ \frac{d\mathbf{z}_t}{dt} &= \mathbf{z}_1 - \mathbf{x} = \mathbf{v}_\theta(\mathbf{z}_t, \mathcal{T}(\mathbf{x}), t)\,. \end{aligned} \tag{2}$$

For synthetic degradation, we explicitly formulate $\mathcal{T}(\mathbf{x})$ as

$$\mathcal{T}(\mathbf{x}) = \mathrm{JPEG}_q\big((\mathbf{x} * \kappa_{\sigma_b}) \downarrow_s\big) + \eta, \quad \eta \sim \mathcal{N}(0, \sigma_n^2), \tag{3}$$

Here $*$ denotes convolution, κ_{σ_b} is a Gaussian blur kernel, and $\downarrow_s$ denotes bicubic downsampling by factor s followed by bicubic upsampling back to the target size (so that $\mathcal{T}(\mathbf{x})$ stays on the same spatial grid as $\mathbf{x}$). Therefore, we can get an explicit solution of the target velocity $\mathbf{z}_1 - \mathbf{x}$ and output velocity $\mathbf{v}_\theta(\mathbf{z}_t, \mathcal{T}(\mathbf{x}), t)$ with a single HQ image $\mathbf{x}$ and train the DiT backbone with RFM loss

$$\mathcal{L}_{\mathrm{RFM}} = \mathbb{E}\Big[\big\|\mathbf{v}_\theta(\mathbf{z}_t, \mathcal{T}(\mathbf{x}), t) - (\mathbf{z}_1 - \mathbf{x})\big\|_2^2\Big]. \tag{4}$$

However, for real-world degradation, $\mathcal{T}(\mathbf{x})$ is implicit, which means we can't calculate the loss with a single unpaired HQ image $\mathbf{x}$. So we introduce RCDT in Sec. 3.5 to get a numerical approximation of $\mathcal{T}(\mathbf{x})$ again.

Full generation vs. latent refinement (optional). Our default sampling performs full generation by initializing from pure noise and integrating Eq. 2 down to t=0. For completeness, we also consider a latent refinement variant that initializes at an intermediate time $\beta \in (0, 1]$ using the pixel-aligned evidence latent (defined in Sec. 3.2):

$$\mathbf{z}_\beta = (1-\beta)\,\mathbf{z}_\alpha^{\mathrm{ctx}}(\mathbf{y}) + \beta\,\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{5}$$

and then integrates Eq. 2 from t=β to t=0. β=1 recovers our default full-generation setting, while smaller β increases reliance on the evidence initialization. We analyze the fidelity-realism trade-off of this variant in Sec. 4.4.

3.2 Inpainting-Interface Evidence Adapter (IIEA)

Native FLUX-Fill conditioning. FLUX-Fill [18] is trained for inpainting with a pixel-aligned interface. Given HQ image latent $\mathbf{x} \in \mathbb{R}^{c\times h\times w}$, a zero image (zero in normed [-1, 1] RGB space) latent $\mathbf{x}' \in \mathbb{R}^{c\times h\times w}$, and a binary latent mask $\mathbf{m} \in \{0,1\}^{c'\times H\times W}$ (with $\mathbf{m}$=1 indicating the region to be inpainted), the masked-image latent is

$$\mathbf{z}^{\mathrm{masked}}(\mathbf{x}, \mathbf{m}) = (\mathbf{x}) \odot (\mathbf{1} - \mathbf{m}) + (\mathbf{x}') \odot \mathbf{m} \tag{6}$$

At noise level t, the DiT consumes aligned inputs $(\mathbf{z}_t, \mathbf{z}^{\mathrm{masked}}, \mathbf{m})$ on the same spatial grid and concatenated at the channel dimension, which are patchified into token grids without breaking pixel correspondence.

Full-image restoration via evidence injection. We convert inpainting into full-image restoration by setting a full-image mask $\mathbf{m} \equiv \mathbf{1}$. This choice is important: in the native inpainting interface, pixels under $\mathbf{m}$=0 are treated as preserved context and are therefore encouraged to remain unchanged. Such copy-preserve semantics are mismatched with SR/restoration, where every pixel may need refinement rather than verbatim copying. Under $\mathbf{m} \equiv \mathbf{1}$, the native construction in Eq. 6 collapses to an all-zero latent $\mathbf{z}^{\mathrm{masked}} = \mathbf{x}'$ and provides no observation. IIEA instead **repurposes** the masked-image slot as a pixel-aligned evidence channel: we keep $\mathbf{m} \equiv \mathbf{1}$ fixed and write the upsampled LQ latent y into $\mathbf{z}^{\mathrm{masked}}$. Notably, the mask is no longer used as a spatial indicator of known vs. unknown pixels; rather, it acts as a constant mode flag that activates the inpainting-conditioning pathway while the dense evidence is carried by the masked-image slot. Although this differs from the pre-training input distribution (masked regions are typically zeroed), LoRA fine-tuning adapts the conditioning pathway to interpret dense evidence. We optionally modulate the evidence strength with a scalar $\alpha \in [0, 1]$ in pixel space:

$$y(\alpha) = \varepsilon(\alpha \cdot Y), \qquad \alpha \in [0, 1], \tag{7}$$

and define the inpainting-style conditioning bundle

$$\mathbf{c}_{\mathrm{IIEA}}(\mathbf{y}; \alpha) = \{y(\alpha),\, \mathbf{m}{=}\mathbf{1}\}. \tag{8}$$

3.3 Lightweight Visual-Semantic Guidance

Fill2SR uses Redux [19] as an image embedder instead of a heavyweight VLM. Given an input image $\mathbf{y}$, Redux produces a fixed-length semantic token sequence

$$\mathbf{s}(\mathbf{y}) = \mathrm{Redux}(\mathbf{y}), \tag{9}$$

which converts images of arbitrary resolution into a 768-token representation. These tokens are injected through the model's existing context pathway to provide semantic stabilization while keeping the conditioning cost bounded.

3.4 Parameter-Efficient Fine-Tuning (LoRA)

We freeze the pre-trained DiT backbone and fine-tune only LoRA [13] on attention/FFN projections. For a linear weight $\mathbf{W}$, LoRA adds a rank-r update:

$$\mathbf{W}' = \mathbf{W} + \frac{\lambda}{r}\mathbf{B}\mathbf{A}, \tag{10}$$

with trainable factors $\mathbf{A}, \mathbf{B}$ and scaling λ (not the evidence strength α in Sec. 3.2). We train $\sim$0.65B parameters; the backbone, VAE, and text encoder are frozen. Target modules and hyperparameters are in the supplementary.

3.5 RCDT: Reference-Conditioned Degradation Transfer

Paired HQ/LQ supervision under unknown real-world degradations is scarce, yet our endpoint-flow objective (Eq. 4) fundamentally relies on paired samples $(\mathbf{x}, \mathbf{y} = \mathcal{T}(\mathbf{x}))$ under real-world degradation model $\mathcal{T}$. RCDT provides such paired endpoint supervision under in-the-wild degradation statistics via reference-anchored degradation transfer, without fitting an explicit parametric degradation model: we use a frozen VLM to distill degradation descriptors from unpaired real-LQ images and a frozen instruction-based image editor [40] to transfer these degradations onto clean targets, synthesizing paired observations $\hat{\mathbf{y}} \approx \mathcal{T}(\mathbf{x})$. Given an unpaired real-LQ set $\{\mathbf{y}_i^{\mathrm{real}}\}_{i=1}^{M}$, we mine one degradation description per image using Qwen3-VL [3] to form a prompt pool $\mathcal{P}$. For each clean training image $\mathbf{x}$ (used as the ground-truth target), we sample $\mathbf{d} \sim \mathrm{Unif}(\mathcal{P})$ and let Qwen3-VL produce a content-aware specification $\tilde{\mathbf{d}}$ describing how $\mathbf{d}$ would manifest on $\mathbf{x}$. Finally, an instruction-based editor [40] transfers the specified degradations onto $\mathbf{x}$ to synthesize the paired LQ observation $\hat{\mathbf{y}}$:

$$\mathbf{d}_i = \mathrm{Qwen3VL}(\mathbf{y}_i^{\mathrm{real}}),\ \tilde{\mathbf{d}} = \mathrm{Qwen3VL}(\mathbf{x}, \mathbf{d}),\ \hat{\mathbf{y}} = \mathrm{Edit}(\mathbf{x}; \tilde{\mathbf{d}}). \tag{11}$$

RCDT is an offline data-synthesis step with frozen models (no test-time overhead); we include its compute in end-to-end cost. To isolate architectural gains from extra supervision, we report results *with/without* RCDT (Tabs. 2, 5). As RCDT aims to transfer real-world degradation statistics, it primarily improves in-the-wild perceptual quality and may trade off distortion-oriented full-reference metrics on synthetic benchmarks. We therefore adopt a conservative construction and

usage protocol: (i) degradation-only prompting (no new objects/text and no intentional geometry edits), (ii) VLM-based verification to reject content drift or severe artifacts, (iii) leakage control by keeping the real-LQ pool disjoint from evaluation sets and screening near-duplicates, and (iv) **late-stage injection**: we mix RCDT pairs only in the final 20% of training to refine toward real-world statistics while preserving the synthetic fidelity learned earlier. More details are provided in the supplementary material.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We evaluate on synthetic (DIV2K [1], LSDIR [20], FFHQ [14]) and real-world benchmarks (RealLQ250 [2], RealPhoto60 [5,45]). Following FaithDiff [5], we construct three synthetic degradation levels (D1/D2/D3) of increasing severity by combining downsampling with stronger blur, additive noise, and JPEG compression, corresponding to 2×, 4×, and 8× inputs. On synthetic benchmarks we report PSNR↑, SSIM↑, LPIPS↓, and DINO feature similarity to audit semantic consistency (higher is better). On real benchmarks without ground truth, we report MUSIQ↑ [15], MANIQA↑ [44], and CLIPIQA↑ [35]; we additionally report AES on RealLQ250 (score and induced rank).
Baselines and Evaluation Protocol. We compare against Real-ESRGAN [37], StableSR [36], DiffBIR [23], OSEDiff [41], SeeSR [42], SUPIR [45], DreamClear [2], FaithDiff [5], and DiT4SR [9].
Backbone capacity note. These methods rely on backbones of different scales/modalities (e.g., SD2/SDXL, PixArt, and SD3.5). We evaluate each pipeline using its official checkpoint and recommended settings, and report memory/compute footprints (Tab. 4 and Supplementary) to contextualize capacity differences. For diffusion/flow baselines, we use their default sampling and (if available) default tiling configuration; for prompt-conditioned methods, we use their default automatic prompt generation without manual editing. For real-world SR, we further enforce a unified output-resolution budget by evaluating at long-side targets of 512/1024/2048 pixels, corresponding to $2\times/4\times/8\times$ on RealLQ250 and $2\times/4\times$ on RealPhoto60 (higher-resolution inputs).

4.2 Implementation Details

Backbone and Objective. Fill2SR is built upon FLUX-Fill [17,18]. We perform restoration in the latent space of a frozen VAE and learn a conditional rectified flow using the RFM objective in Eq. 4. Low-quality evidence is injected through our proposed inpainting-interface evidence adapter (Sec. 3.2), and Redux semantics (Sec. 3.3) are enabled by default. We freeze the DiT backbone and train only LoRA parameters [13] (rank 256) on attention and FFN projections. The total number of trainable parameters is 0.65B.
Training Schedule and Data Mixture. We train for 20k steps with a learning rate of 2.5×10^{-4}, a 2.5k-step warm-up, and a global batch size of 16. Rather than

adapting at a single fixed resolution, we use mixed-resolution, mixed-aspect-ratio crops: 512-1536 pixels in the early stage and 1024-2048 pixels in the later stage, with 1:1, 4:3, and 16:9 buckets up to native QHD (2560×1440). We pre-train on 300k synthetic pairs and inject 10k RCDT-generated pairs only in the final stage. This recipe, enabled by the bounded conditioning cost of IIEA, is a key reason for the stable behavior across 512/1024/2048 test scales. RCDT data synthesis is conducted offline and introduces no test-time overhead.

Compute and Memory Optimizations. Model training requires 4× RTX 5880 Ada (48GB) for 14 days (50.6 A100e GPU-days; 56 raw GPU-days; see Supplementary for normalization). RCDT data generation uses 4× RTX 5880 Ada for 5 days (18.1 A100e GPU-days; 20 raw GPU-days). To enable native QHD training within 48GB VRAM, we precompute and cache all frozen features (VAE latents, text-encoder embeddings when used, and Redux tokens), and enable gradient checkpointing in the DiT backbone.

4.3 Main Results

Quantitative comparison on synthetic benchmarks. Table 1 reports results averaged over three degradation levels (D1–D3) on DIV2K [1], LSDIR [20], and FFHQ [14]. As expected, distortion-oriented fidelity metrics (PSNR/SSIM) tend to favor regression-based pipelines. To isolate the effect of domain-transfer paired supervision, we report *Ours-Base* (w/o RCDT) and *Ours-Full* (+RCDT). *Ours-Base* achieves the best LPIPS on DIV2K and LSDIR and remains competitive on FFHQ-face. After injecting RCDT pairs, *Ours-Full* trades a small amount of synthetic-domain LPIPS for consistently stronger perceptual/no-reference metrics (MUSIQ/MANIQA/CLIPIQA), indicating improved recovery of photo-realistic textures on real degradations.

Quantitative comparison on real-world SR and the effect of RCDT. Table 2 reports results on RealLQ250 [2] and RealPhoto60 [45] under multiple upscaling factors, and includes DiT4SR [9] as a strong diffusion baseline. Consistent with the synthetic-to-real trade-off observed in Table 1, injecting RCDT incurs only a slight drop in synthetic-domain full-reference metrics (e.g., LPIPS) but yields substantial improvements on real-world perceptual/no-reference metrics. While DiT4SR attains strong results at the extreme 8× setting on RealLQ250, *Ours-Full* achieves the best or second-best performance on most entries across scales. These results support the fidelity-preserving control of IIEA and the effectiveness of RCDT for real-world domain transfer.

Aesthetic quality and semantic consistency audit. No-reference IQA metrics can sometimes reward over-sharpening or hallucinated details. To better assess whether perceptual gains come with semantic drift, we additionally report (i) AES on RealLQ250 (rank/score) and (ii) DINO feature consistency on paired benchmarks. Table 3 summarizes the AES score and ranking, and reports DINO consistency and robustness on LSDIR_VAL. Fill2SR achieves the best RealLQ250 AES score and maintains the second best DINO similarity.

Qualitative comparisons. Fig. 3 shows representative results on real-world inputs. Fill2SR improves text legibility and recovers natural micro-textures (e.g.,

Table 1: Quantitative comparison on synthetic benchmarks (*DIV2K*, *LSDIR*, and *FFHQ*). Results are averaged across three degradation levels (D1–D3). While regression-based methods obtain higher PSNR/SSIM, our generative approach performs strongly on perceptual metrics (e.g., LPIPS, CLIPIQA), indicating improved recovery of photo-realistic textures. **Ours-Base** is trained with standard synthetic paired data only (w/o RCDT), whereas **Ours-Full** additionally injects RCDT supervision (+RCDT) in the final stage, which may trade off synthetic-domain LPIPS while boosting perceptual/no-reference scores. Blue outlines mark **Ours-Full** on LPIPS for readability. The best and second-best results are highlighted in **bold** and underlined.

Datasets	Metrics	Methods								
		Real-ESRGAN	StableSR	DiffBIR	SeeSR	SUPIR	DreamClear	FaithDiff	**Ours-Base** (w/o RCDT)	**Ours-Full** (+RCDT)
DIV2K_VAL	PSNR↑	23.3520	22.9703	**23.4959**	22.5903	22.6610	21.7065	22.4013	22.4596	21.7831
	SSIM↑	**0.6414**	0.6085	0.6065	0.5807	0.5748	0.5557	0.5574	0.5872	0.5679
	LPIPS↓	0.4017	0.4380	0.4168	0.4165	0.4032	0.4153	0.4051	**0.3866**	0.3981
	MUSIQ↑	58.9699	49.0826	66.1573	**69.0464**	64.6142	65.9251	68.1122	65.9644	67.8893
	MANIQA↑	0.5252	0.4444	0.5709	0.6015	0.5858	0.5743	0.6184	0.6173	**0.6271**
	CLIPIQA↑	0.4875	0.3801	0.5737	0.6035	0.5051	0.5310	0.5763	0.6011	**0.6138**
LSDIR_VAL	PSNR↑	**20.7295**	20.3530	20.7116	20.0585	19.9397	19.4056	19.6345	19.6930	19.2354
	SSIM↑	**0.5670**	0.5185	0.5287	0.4900	0.4847	0.4923	0.4622	0.5019	0.4890
	LPIPS↓	0.4055	0.4442	0.4090	0.4126	0.4202	0.4100	0.4083	**0.3878**	0.3974
	MUSIQ↑	64.5601	52.8770	70.1140	72.6831	67.9508	70.3541	71.6307	72.3820	**73.1206**
	MANIQA↑	0.5674	0.4788	0.6153	0.6388	0.6234	0.6149	0.6653	0.6840	**0.6859**
	CLIPIQA↑	0.5424	0.4199	0.6309	0.6413	0.5704	0.6121	0.6307	0.7153	**0.7223**
FFHQ-face	PSNR↑	**28.3966**	27.3251	27.9839	27.3894	27.4845	26.6292	26.4585	26.5658	26.3979
	SSIM↑	**0.7777**	0.7190	0.7291	0.7305	0.7040	0.6902	0.6689	0.7081	0.7097
	LPIPS↓	0.3838	0.3879	0.4016	0.3731	**0.3682**	0.3837	0.3768	0.3827	0.3857
	MUSIQ↑	63.4199	66.7731	73.9354	73.3842	73.6046	70.7586	76.1941	75.6041	**76.4407**
	MANIQA↑	0.4920	0.5029	0.5925	0.5875	0.5970	0.5654	**0.6406**	0.6219	0.6286
	CLIPIQA↑	0.4297	0.4827	0.6283	0.5493	0.5300	0.4807	0.5719	0.6355	**0.6470**

frost crystals) without compromising structural details (e.g., facial geometry). Due to space, Fig. 3 focuses on representative diffusion/flow-based baselines (including DiT4SR [9]) and highlights robustness across scale factors (2×/4×/8×). Additional qualitative comparisons (including synthetic D3 cases and a broader set of baselines) are provided in the Supplementary Material.

Memory scalability and adaptation cost. As shown in Table 4, Fill2SR fits within a single 32GB GPU at 1536^2 while keeping the conditioning budget constant. Although we rely on a large frozen backbone and external models for offline RCDT synthesis, the end-to-end compute budget (training + data generation) remains substantially lower than prior diffusion restoration pipelines under the reported settings. For a more meaningful cross-hardware comparison, Table 4 reports A100-equivalent GPU-days (A100e) as the main value and raw GPU-days in parentheses; the Supplementary Material details the normalization.

Table 2: Quantitative comparison on real-world SR benchmarks (*RealPhoto60* and *RealLQ250*). We upscale real-world inputs without explicit degradation priors and evaluate at target output sizes of 512/1024/2048 pixels (*RealLQ250*: 2×/4×/8×; *RealPhoto60*: 2×/4×). **Ours-Full** (+RCDT; 10k pairs injected in the final stage) consistently improves over **Ours-Base** (w/o RCDT) on perceptual/no-reference metrics. The best and second-best results are highlighted in **bold** and underlined.

Datasets	Scale	Metrics	Methods										
			Real-ESRGAN	StableSR	DiffBIR	OSEDiff	SeeSR	SUPIR	DreamClear	FaithDiff	DiT4SR	**Ours-Base** (w/o RCDT)	**Ours-Full** (+RCDT)
RealPhoto60	2×	MUSIQ↑	59.0296	50.2670	61.6646	70.4686	71.8052	69.6326	68.8353	71.5853	72.5731	70.6389	**73.5949**
		MANIQA↑	0.4797	0.4554	0.5617	0.5891	0.6079	0.6116	0.5905	**0.6513**	0.6309	0.6307	0.6468
		CLIPIQA↑	0.5068	0.3632	0.5465	0.5726	0.5959	0.5972	0.5383	0.5718	0.6192	0.6780	**0.7137**
	4×	MUSIQ↑	50.9528	22.4571	49.8760	55.4632	58.1401	57.0462	62.4641	58.5360	63.7202	61.8281	**65.8413**
		MANIQA↑	0.4785	0.2960	0.5522	0.5321	0.5530	0.5269	0.5077	0.5657	0.5803	0.5957	**0.6042**
		CLIPIQA↑	0.4247	0.2354	0.4739	0.4595	0.5386	0.4542	0.3765	0.5007	**0.5879**	0.5207	0.5486
RealLQ250	2×	MUSIQ↑	67.7327	63.0388	71.5459	71.3742	72.1775	68.3565	69.4476	71.6023	71.7326	67.5721	**72.5417**
		MANIQA↑	0.6177	0.6116	0.6676	0.6567	0.6643	0.6463	0.6499	**0.6869**	0.6760	0.6454	0.6861
		CLIPIQA↑	0.5784	0.4904	0.6435	0.6116	0.6363	0.6196	0.6356	0.6299	0.6244	0.6373	**0.7158**
	4×	MUSIQ↑	62.5154	50.4880	65.9055	69.5570	70.3728	63.2153	66.2208	69.8047	70.5121	62.1914	**71.5498**
		MANIQA↑	0.5239	0.4500	0.5700	0.5782	0.5927	0.5812	0.5756	0.6373	0.6154	0.5751	**0.6387**
		CLIPIQA↑	0.4359	0.3332	0.5370	0.5282	0.5568	0.4542	0.5017	0.5269	0.5617	0.5384	**0.6287**
	8×	MUSIQ↑	41.9557	25.0598	50.9256	57.0877	58.9324	51.3011	49.7753	58.1177	**61.8378**	50.0516	59.3031
		MANIQA↑	0.4582	0.3119	0.5274	0.5196	0.5509	0.5173	0.4991	0.5653	0.5766	0.5516	**0.5876**
		CLIPIQA↑	0.3614	0.2425	0.4425	0.4435	0.5071	0.3849	0.3881	0.4829	**0.5403**	0.4293	0.4784

Metric	StableSR	DiffBIR	SUPIR	DreamClear	SeeSR	FaithDiff	DiT4SR	**Ours-Full**
RealLQ250 Score↑	4.9950	5.2653	5.2718	5.3254	5.3296	5.2487	5.5432	**5.5510**
RealLQ250 Rank↓	7.9160	5.9640	5.5760	5.0080	4.6800	5.7280	**2.7560**	2.9400
LSDIR DINO Avg↑	0.6667	0.7452	0.7397	**0.8089**	0.7676	0.7694	0.7604	0.8018

Table 3: Additional results on aesthetic quality and semantic consistency. RealLQ250 is evaluated on the 4x results. DINO Avg is computed by averaging DINO over LSDIR_VAL from D1 to D3. Best and second-best are marked with **bold** and underlined.

Method	Params (Train/Backbone, B)	Time (s)↓	32GB	Train (A100e)↓	Data (A100e)↓
SeeSR (SD2)	2.30 / 2.30	34	✓	N/R	N/A
SUPIR (SDXL)	1.30 / 1.30	26	✗	317.5 (640)	N/A
DreamClear (PixArt)	2.41 / 2.41	107	✗	224.0 (224)	592.8 (1360)
FaithDiff (SDXL)	2.40 / 2.40	35	✗	N/R	N/A
Fill2SR (FLUX-Fill)	**0.65 / 12.0**	75	✓	**50.6 (56)**	**18.1 (20)**

Table 4: Cost breakdown and 32GB feasibility at 1536^2. Time is measured with 28 inference steps (batch=1) under each method's recommended inference configuration (including its default tiling strategy and prompt pipeline when applicable). 32GB indicates whether the full pipeline runs without out-of-memory errors on a single 32GB GPU at 1536^2. "Train" and "Data" report A100-equivalent GPU-days (A100e) as the primary value; raw GPU-days are shown in parentheses. A100e uses peak dense BF16/FP16 tensor throughput for coarse cross-hardware normalization (see Supplementary for details). N/R and N/A denote not reported and not applicable.

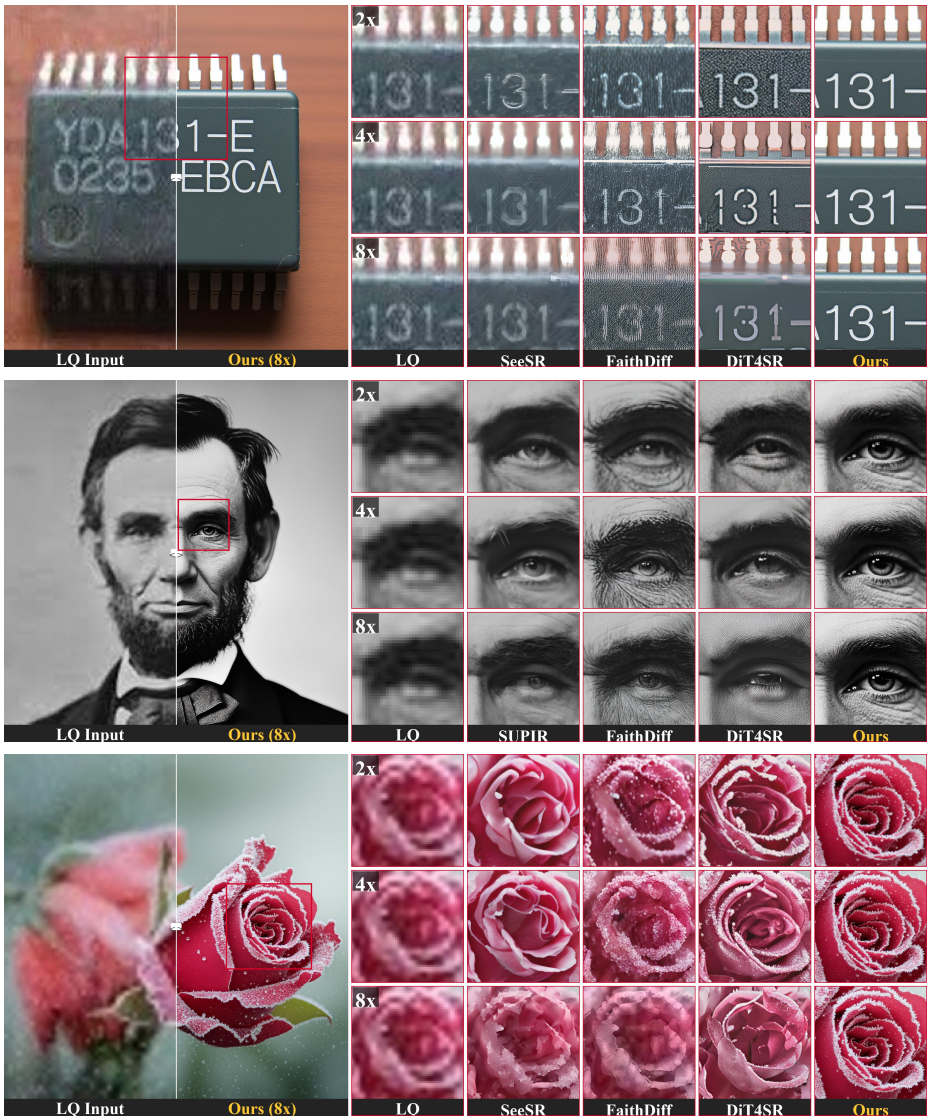

Fig. 3: Qualitative comparisons on real-world scenarios at up to 8x degradation. Fill2SR produces visually plausible restorations across diverse and challenging domains. **Row 1 (Industrial Text):** Fill2SR can improve text legibility and boundary sharpness, while some baselines introduce distracting structured artifacts. **Row 2 (Natural Textures):** Fill2SR tends to recover fine textures (e.g., frost-like micro-structures) without excessive over-smoothing. **Row 3 (Human Faces):** Fill2SR often preserves facial structure and eye highlights, while some baselines exhibit local geometric drift. The left split-screen intuitively highlights our global visual enhancement. More results with additional baselines are provided in the Supplementary Material.

4.4 Ablation Study

We ablate the key components of Fill2SR on synthetic (DIV2K [1], LSDIR [20]) and real-world (RealLQ250 [2]) sets, averaging scores over three degradation levels (D1–D3). Detailed per-level results are in the Supplementary Material.

Evidence strength α (IIEA). α controls how strongly the model uses pixel-aligned evidence without changing the conditioning budget. Table 5 shows that smaller α degrades both synthetic distortion and real-image no-reference scores. Differences between $\alpha \in \{0.9, 1.0\}$ are minor, making it a mild robustness knob rather than a dramatic controller. We default to α=1.

Redux visual-semantic tokens. Removing Redux tokens degrades no-reference metrics under severe degradations, indicating that bounded semantic guidance prevents category drift and stabilizes global structure. Alternative choices (zero/noise image tokens) underperform Redux, supporting our use of the image embedder and removal of the VLM.

RCDT paired synthesis. RCDT provides reference-anchored paired endpoint supervision complementing standard synthetic degradations. Removing it significantly degrades real-world no-reference metrics on RealLQ250, while slightly improving synthetic distortion metrics (*e.g.*, DIV2K_VAL PSNR/LPIPS). This indicates RCDT primarily aligns the training supervision to real degradation statistics rather than "gaming" synthetic metrics; we further report semantic/aesthetic audits to probe metric bias and hallucinations (Tab. 3). RCDT is offline only and introduces zero test-time overhead.

Full generation vs. latent refinement. Partial refinement (β<1) improves synthetic PSNR/SSIM but severely harms real-image perceptual quality. We hypothesize intermediate initializations fall off the learned flow manifold, conflicting with pixel-aligned evidence and yielding unstable trajectories [28]. Thus, we use full generation (β=1) by default.

4.5 Limitations and Discussion

Fill2SR is a generative restorer and may synthesize plausible yet incorrect details when the evidence is extremely weak or ambiguous; it should not be used in scenarios requiring pixel-level authenticity (*e.g.*, forensics, medical imaging, or legal evidence). Although the adaptation is parameter-efficient in *trainable* parameters, inference still requires running a large frozen backbone (*e.g.*, 12B for FLUX-Fill-dev [18]) and iterative sampling, which is slower than feed-forward regressors-distillation and faster solvers are important directions. IIEA repurposes the masked-image slot in a way that differs from the backbone's original inpainting pre-training; while LoRA fine-tuning mitigates this mismatch, a deeper analysis of the resulting conditioning behavior is left for future work. RCDT relies on frozen instruction-based editors and VLM verification, which may still miss subtle geometric drift; we adopt conservative prompting and filtering, but cannot fully guarantee pixel-perfect alignment for every synthesized pair. Finally, because sampling is stochastic, a fuller multi-seed statistical characterization of no-reference metrics would further strengthen the evaluation.

Fig. 4: Qualitative ablation on representative real-world inputs (RealLQ250). Columns show the LQ input, latent refinement ($\beta = 0.5$), w/o RCDT, w/o Redux, weakened evidence injection ($\alpha = 0.5$), and the default Fill2SR output. *Top:* industrial text. *Bottom:* natural texture.

Setting	DIV2K_VAL		LSDIR_VAL		RealLQ250		
Setting	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	CLIPIQA↑	MANIQA↑	MUSIQ↑
Default	21.7831	0.3981	19.2354	0.3974	**0.6076**	**0.6375**	**67.7982**
w/o RCDT data	22.4596	**0.3866**	19.6930	**0.3878**	0.5350	0.5907	59.9384
w/o Redux tokens	21.1379	0.4125	18.6203	0.4118	0.5963	0.6328	67.1587
Redux tokens (zero image)	21.4694	0.4067	18.9790	0.4081	0.6023	0.6336	67.4110
Redux tokens (noise image)	21.6229	0.4047	19.0723	0.4043	0.6059	0.6355	67.4399
Evidence α=0.5	20.9871	0.4105	18.5959	0.4040	0.6031	0.6333	67.5953
Evidence α=0.7	21.4020	0.4042	18.9017	0.4004	0.6037	0.6348	67.5593
Evidence α=0.9	21.7308	0.3994	19.1563	0.3981	0.6058	0.6364	67.6596
Refine β=0.5	23.7536	0.5568	21.1851	0.5731	0.3047	0.3987	28.2379
Refine β=0.7	**23.8974**	0.5261	**21.2508**	0.5501	0.3075	0.3972	28.8641
Refine β=0.9	23.8655	0.4433	21.2463	0.4599	0.3314	0.4027	34.7317

Table 5: Ablation results on DIV2K_VAL, LSDIR_VAL, and RealLQ250, averaged over three degradation levels (D1–D3). We compare removing RCDT training data, controlling Redux semantic tokens (disabled / zero-image / noise-image), and sweeping evidence/refine strengths.

5 Conclusion

We presented Fill2SR as a complementary inpainting-interface formulation for SR that keeps conditioning overhead bounded with resolution while making memory use predictable, without relying on ControlNet-style branches or spatial token streams. IIEA repurposes the native masked-image interface of inpainting diffusion transformers and writes VAE-encoded LQ evidence into this slot, enabling mixed-resolution training up to native QHD (2560×1440). With reverse-degradation rectified-flow training, bounded Redux guidance, and offline RCDT synthesis, Fill2SR achieves strong LPIPS performance on synthetic benchmarks and strong perceptual quality on real-world benchmarks while remaining feasible on a single 32GB GPU at 1536^2. Code and models will be released at `https://github.com/Xingfu-Yi/Fill2SR`.

References

1. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2017)
2. Ai, Y., Zhou, X., Huang, H., Han, X., Chen, Z., You, Q., Yang, H.: DreamClear: High-capacity real-world image restoration with privacy-safe dataset curation. Advances in Neural Information Processing Systems pp. 55443–55469 (2024)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
5. Chen, J., Pan, J., Dong, J.: Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28188–28197 (2025)
6. Chen, R.T.Q., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. In: Advances in Neural Information Processing Systems. vol. 31 (2018)
7. Deng, J., Wu, X., Yang, Y., Zhu, C., Wang, S., Wu, Z.: Acquire and then adapt: Squeezing out text-to-image model for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23195–23206 (2025)
8. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: Advances in Neural Information Processing Systems. pp. 8780–8794 (2021)
9. Duan, Z.P., Zhang, J., Jin, X., Zhang, Z., Xiong, Z., Zou, D., Ren, J.S., Guo, C., Li, C.: DiT4SR: Taming diffusion transformer for real-world image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18948–18958 (October 2025)
10. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (June 2020)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. pp. 6840–6851 (2020)
12. Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Transactions on Image Processing **29**, 4041–4056 (2020)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. In: Proceedings of the International Conference on Learning Representations (2022)
14. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
15. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)

16. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
17. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024)
18. Labs, B.F.: FLUX.1-Fill-dev. https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev (2024), model card; accessed 2026-02-04
19. Labs, B.F.: Redux. https://huggingface.co/black-forest-labs/FLUX.1-Redux-dev (2024)
20. Li, Y., Zhang, K., Liang, J., Cao, J., Liu, C., Gong, R., Zhang, Y., Tang, H., Liu, Y., Demandolx, D., Ranjan, R., Timofte, R., Van Gool, L.: LSDIR: A large scale dataset for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1775–1787 (2023)
21. Liang, J., Zeng, H., Zhang, L.: Efficient and degradation-adaptive network for real-world image super-resolution. In: Proceedings of the European Conference on Computer Vision. pp. 574–591 (2022)
22. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1833–1844 (2021)
23. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: DiffBIR: Toward blind image restoration with generative diffusion prior. In: Proceedings of the European Conference on Computer Vision. pp. 430–448 (2024)
24. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems. vol. 36, pp. 34892–34916 (2023)
26. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
27. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
28. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
29. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Proceedings of the International Conference on Machine Learning. pp. 8162–8171 (2021)
30. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
32. Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M.: Palette: Image-to-image diffusion models. arXiv preprint arXiv:2111.05826 (2021)
33. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2023)
34. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy,

S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: Advances in Neural Information Processing Systems. vol. 35, pp. 25278–25294 (2022)
35. Wang, J., Chan, K.C.K., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
36. Wang, J., Yue, Z., Zhou, S., Chan, K.C.K., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision **132**(12), 5929–5949 (2024)
37. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1905–1914 (2021)
38. Wang, Y., Yang, W., Chen, X., Wang, Y., Guo, L., Chau, L.P., Liu, Z., Qiao, Y., Kot, A.C., Wen, B.: SIN-SR: Diffusion-based image super-resolution in a single step. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25796–25805 (2024)
39. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17683–17693 (2022)
40. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-Image technical report (2025), https://arxiv.org/abs/2508.02324
41. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. Advances in Neural Information Processing Systems pp. 92529–92553 (2024)
42. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: SeeSR: Towards semantics-aware real-world image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25456–25467 (2024)
43. Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: SmartBrush: Text and shape guided object inpainting with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22428–22437 (June 2023)
44. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: MANIQA: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1191–1200 (2022)
45. Yu, F., Gu, J., Li, Z., Hu, J., Kong, X., Wang, X., He, J., Qiao, Y., Dong, C.: 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. pp. 25669–25680 (2024)
46. Yue, Z., Wang, J., Loy, C.C.: ResShift: Efficient diffusion model for image super-resolution by residual shifting. In: Advances in Neural Information Processing Systems. pp. 13294–13307 (2023)
47. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5728–5739 (2022)

48. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4791–4800 (2021)
49. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
50. Zhuang, J., Zeng, Y., Liu, W., Yuan, C., Chen, K.: A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In: Proceedings of the European Conference on Computer Vision. pp. 195–211 (2024)