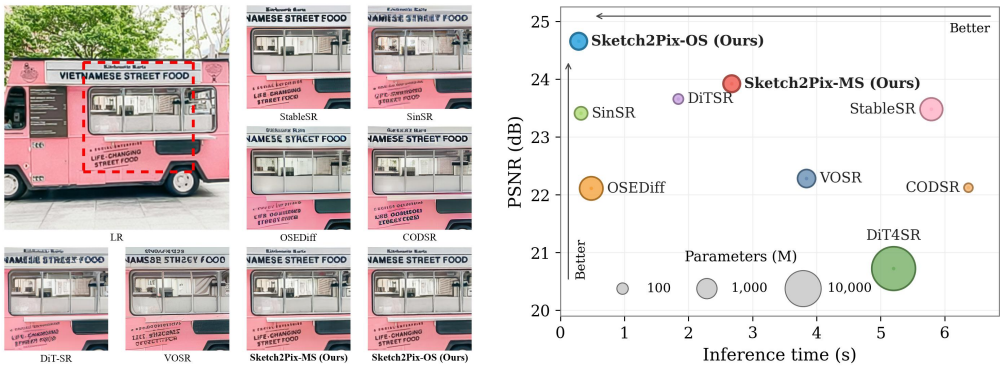
SKETCH BEFORE PIXEL: PIXEL-SPACE STRUCTURE FORCING FOR REAL-WORLD SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review



(a) Qualitative Comparisons (b) Quantitative–Efficiency Comparisons

Figure 1: **Qualitative, quantitative and efficiency comparisons on ScreenSR.** (a) Existing methods tend to produce blurred or distorted lettering and window details, whereas Sketch2Pix better preserves text shapes and window layouts while recovering clearer local details. (b) Our Sketch2Pix-MS achieves higher PSNR than the compared multi-step methods, while Sketch2Pix-OS further improves reconstruction fidelity with substantially faster inference.

ABSTRACT

Pixel-space diffusion model offers a promising alternative for Real-world Image Super-Resolution (Real-SR), enabling end-to-end restoration without lossy VAE compression. However, existing pixel-space generation methods typically couple structural reconstruction and detail generation along a single denoising trajectory, making joint optimization more challenging for Real-SR. To address these limitations, we propose **Sketch2Pix**, a VAE-free Pixel-Space Diffusion Transformer that restores images from structural sketches to fine-grained details. Specifically, we first discover that performing moderate structural restoration before image refinement can improve the restoration quality. Based on this discovery, we design a Dual-Time Joint Diffusion framework that leverages the low-dimensional manifold hypothesis of natural images to jointly predict clean low-frequency structure and image, thereby enabling end-to-end pixel-space restoration. Furthermore, to improve the quality of the recovery, we devise a Structure-Forcing Schedule strategy that reorders the structure and pixel denoising trajectories to explicitly control the recovery order, allowing structure to emerge earlier and guide subsequent pixel generation. Benefiting from the above designs, our Sketch2Pix preserves complete pixel information while transforming coupled restoration into an ordered structure-to-detail process. Extensive experiments on Real-SR benchmarks illustrate that our Sketch2Pix restores perceptually realistic details while faithfully preserving structural content and enabling efficient inference.

1 INTRODUCTION

Real-world Image Super-Resolution (Real-SR) aims to preserve faithful structures from observations degraded while generating realistic details (Dong et al., 2015; Wang et al., 2021). Recently, Latent Diffusion Models (LDMs) have emerged as the dominant paradigm for this task, combining

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.

powerful generative priors with efficient modeling in latent spaces (Rombach et al., 2022; Wu et al., 2024b). However, these methods rely on lossy VAEs that can discard fine textures and text strokes, while the separation of latent denoising and pixel reconstruction hinders joint optimization.

To overcome these limitations, recent studies have revisited pixel-space diffusion, enabling VAE-free image generation through improved DiT architectures and prediction parameterizations (Li & He, 2026; Yu et al., 2026; Sun et al., 2026; Ma et al., 2026). Although simply bypassing VAE compression avoids the loss of texture details, structural reconstruction and detail generation remain coupled along a single denoising trajectory, making their joint optimization more challenging (Baade et al., 2026). These limitations raise a natural question: *Can we explicitly organize the recovery order to better coordinate structural reconstruction and detail generation in pixel-space Real-SR?*

To address this issue, we analyze the denoising trajectories of pixel-space diffusion, as illustrated in Figure 2. Existing pixel-space restoration methods typically use a single time variable for the full-image denoising, without independently controlling the progress of structural recovery and detail generation (red curve) (Sun et al., 2026; Qiao et al., 2026). Since reliable spatial layouts and contours provide a basis for detail generation, we hypothesize that making structural cues available earlier can facilitate subsequent image refinement. To examine this hypothesis, we adopt clean-image prediction motivated by low-dimensional manifold theory and extract low-frequency components from the HR image as an explicit structural target, assigning it an independent time variable to control its denoising progress relative to the image. By comparing different denoising orders, we find that appropriately accelerating structural denoising improves restoration quality, highlighting the importance of recovery order in Real-SR. Inspired by this insight, we aim to explicitly reorganize the generation order in pixel-space through a *Sketch Before Pixel* paradigm, allowing structural recovery to precede and guide detail synthesis.

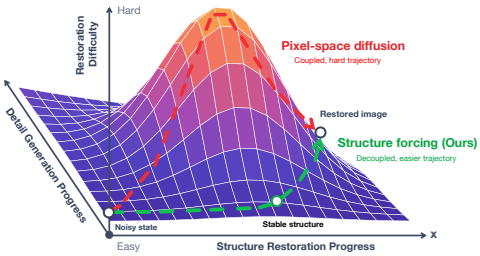


Figure 2: Conceptual comparison of Pixel-space diffusion and Structure forcing (Ours).

Based on the above analysis, we propose **Sketch2Pix**, a self-contained pixel-space diffusion framework that explicitly reorders the denoising trajectory for end-to-end Real-SR. Specifically, we propose the Dual-Time Joint Diffusion (DJD) that leverages low-dimensional manifold theory to jointly model low-frequency structure and the full image through independent diffusion times. Furthermore, we devise a Structure-Forcing Schedule (SFS) that allows structural recovery to lead image refinement along a joint denoising trajectory, realizing the *Sketch Before Pixel* paradigm. To reduce inference cost, we further distill the multi-step model into a one-step student for efficient restoration. Extensive quantitative and qualitative evaluations demonstrate that Sketch2Pix achieves superior fidelity and perceptual quality across Real-SR benchmarks, as illustrated in Figure 1.

Our contributions are summarized as follows:

- We propose a VAE-free, end-to-end pixel-space Diffusion Transformer for Real-SR, termed **Sketch2Pix**, which incorporates our DJD design to jointly model low-frequency structure and the full image without any external pre-trained encoders.
- To explicitly control the recovery order, we design an SFS that coordinates the relative denoising progress of structure and image, allowing structural cues to emerge earlier and guide subsequent image refinement.
- We further distill Sketch2Pix into a one-step model to accelerate inference while retaining competitive restoration quality. Extensive results show that our method achieves a favorable balance between performance and efficiency compared with state-of-the-art methods.

2 Related Work

2.1 Real-World Image Super-Resolution

Real-SR has progressed from reconstruction-oriented learning to generative modeling for perceptually realistic restoration. Early convolutional and Transformer-based approaches primarily optimize pixel-wise reconstruction objectives (Dong et al., 2015; Lim et al., 2017; Liang et al., 2021). Although pixel-level objectives can effectively reduce reconstruction errors, they typically yield results

that are excessively smooth and lack textural detail. To improve realistic quality, subsequent methods combine perceptual and adversarial learning with practical degradation modeling (Ledig et al., 2017; Wang et al., 2021). However, these methods may introduce artifacts or synthesize textures inconsistent with the underlying image content.

More recently, diffusion-based approaches have incorporated powerful generative priors into Real-SR to alleviate the mentioned issue (Rombach et al., 2022; Wang et al., 2024a; Lin et al., 2024; Wu et al., 2024b). Specifically, these methods explore restoration-specific conditioning, semantic guidance, and decoder adaptation (Wang et al., 2024a; Lin et al., 2024; Wu et al., 2024b). Despite these advances, reliance on latent representations can compromise fine-grained fidelity. Unlike the above methods, our Sketch2Pix performs Real-SR directly in pixel space, jointly restoring image structure and fine-grained details without VAE encoding or decoding.

2.2 Pixel-Space Generative Modeling

Early diffusion models generate images directly in pixel space through U-Net backbones that combine multiscale features with skip connections (Ho et al., 2020). To reduce computational costs, LDMs perform the denoising process in a VAE-compressed latent space (Rombach et al., 2022; Lipman et al., 2023). Recent advances have revisited VAE-free generation through improved Transformer architectures (Chen et al., 2026c; Ma et al., 2026; Yu et al., 2026) and prediction objectives (Li & He, 2026; Chen et al., 2026a). In addition, several studies exploit the frequency-domain properties of low-frequency structure and high-frequency details to design frequency-aware losses (Degeorge et al., 2026) and heterogeneous flow generation paths (Lin et al., 2026) for effective generation.

However, the above methods primarily focus on representation and optimization, without fundamentally decoupling reconstruction from generation. Based on this issue, Latent Forcing (Baade et al., 2026) uses independent diffusion times, allowing auxiliary latents to serve as a computational scratchpad before high-frequency pixel generation. However, constructing auxiliary representations requires a pre-trained vision encoder, which introduces additional feature extraction overhead and relies on external priors. In contrast, Sketch2Pix constructs self-contained structural targets through a fixed wavelet transform without external pre-trained encoders, and combines Dual-Time Joint Diffusion with a Structure-Forcing Schedule to let structural recovery lead and guide image refinement.

3 Methodology

3.1 Conditional Flow Matching

Let $\boldsymbol{x}_0$ denote an HR target image in pixel space and $\boldsymbol{y}$ its corresponding LR observation. We formulate Real-SR as conditional generation from $p(\boldsymbol{x}_0 \mid \boldsymbol{y})$ using flow matching (Lipman et al., 2023). Under a data-to-noise time convention, training states are constructed by linear interpolation:

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \quad t \in [0,1], \tag{1}$$

where $t = 0$ corresponds to the clean image and $t = 1$ to Gaussian noise. The corresponding target velocity is $\boldsymbol{v} = \boldsymbol{\epsilon} - \boldsymbol{x}_0$. Then, the conditional velocity field is learned by minimizing:

$$\mathcal{L}_{\mathrm{CFM}} = \mathbb{E}_{t,\,\boldsymbol{x}_0,\boldsymbol{y},\,\boldsymbol{\epsilon}} \left[\|\widehat{\boldsymbol{v}}_\theta - \boldsymbol{v}\|_2^2 \right], \quad \widehat{\boldsymbol{v}}_\theta = f_\theta(\boldsymbol{x}_t, t, \boldsymbol{y}). \tag{2}$$

However, predicting $\boldsymbol{v} = \boldsymbol{\epsilon} - \boldsymbol{x}_0$ requires modeling a full-dimensional noise component, which can be challenging in pixel space. Motivated by the low-dimensional manifold hypothesis of natural images (Li & He, 2026), $\boldsymbol{x}_0$-prediction instead directly estimates the clean target:

$$\widehat{\boldsymbol{x}}_0 = f_\theta(\boldsymbol{x}_t, t, \boldsymbol{y}), \qquad \widehat{\boldsymbol{v}}_\theta = \frac{\boldsymbol{x}_t - \widehat{\boldsymbol{x}}_0}{t}, \quad t > 0. \tag{3}$$

This formulation uses a single time variable for image denoising. We next extend it with an auxiliary structural state whose denoising progress can be controlled independently.

3.2 Dual-Time Joint Diffusion

Low-frequency components capture coarse spatial layouts and large-scale structures (Degeorge et al., 2026), making them a natural representation of image structure. To explicitly separate the

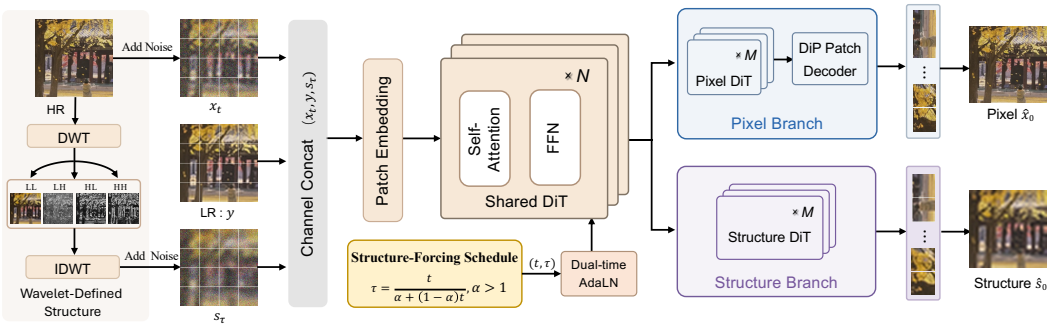


Figure 3: **Overview of our Sketch2Pix.** Our method jointly predicts clean structure and the full image through a wavelet-defined structure, a shared DiT backbone and two branches. Furthermore, the Structure-Forcing Schedule advances structural recovery to guide subsequent detail generation.

recovery progress of structure and details, we introduce Dual-Time Joint Diffusion (DJD), which jointly models both states with independent times via clean-image prediction guided by the low-dimensional image manifold. As illustrated in Figure 3, a shared DiT backbone and two prediction branches predict their clean targets.

Wavelet-Defined Structure. We construct the structural target directly from the HR image $\boldsymbol{x}_0$ using a fixed discrete wavelet transform (DWT), without external pre-trained encoders:

$$\boldsymbol{s}_0 = \mathrm{IDWT}\left(\mathrm{DWT}_{LL}(\boldsymbol{x}_0), \mathbf{0}, \mathbf{0}, \mathbf{0}\right). \tag{4}$$

The detail subbands are set to zero and applying the IDWT yields a low-frequency target spatially aligned with the full image.

Dual-Time Joint Modeling. Alongside the image state $\boldsymbol{x}_t$ defined in Eq. 1, we can construct a noisy structural state with an independent time variable τ:

$$\boldsymbol{s}_\tau = (1-\tau)\boldsymbol{s}_0 + \tau\boldsymbol{\epsilon}_s, \qquad \boldsymbol{\epsilon}_s \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \quad \tau \in [0, 1] \tag{5}$$

where $\boldsymbol{\epsilon}_s$ is independent of the image noise $\boldsymbol{\epsilon}$. The times t and τ specify the noise levels of the image and structural states, respectively. Given both noisy states and the LR observation, DJD jointly predicts the clean HR image and the structural target:

$$(\widehat{\boldsymbol{x}}_0, \widehat{\boldsymbol{s}}_0) = f_\theta\left(\boldsymbol{x}_t, \boldsymbol{s}_\tau, t, \tau, \boldsymbol{y}\right), \tag{6}$$

where f_θ denotes the joint prediction network. Following Eq. 3, the corresponding velocity estimates are $\widehat{\boldsymbol{v}}_\theta = \frac{\boldsymbol{x}_t - \widehat{\boldsymbol{x}}_0}{t}, \widehat{\boldsymbol{u}}_\theta = \frac{\boldsymbol{s}_\tau - \widehat{\boldsymbol{s}}_0}{\tau}, t, \tau > 0$. This formulation allows structural cues at different noise levels to participate in image prediction, while each state follows its own denoising schedule.

Shared DiT Backbone. To jointly process the image and structural states, we first resize the LR observation $\boldsymbol{y}$ to the target HR resolution, obtaining $\widetilde{\boldsymbol{y}}$. Then, we concatenate it with $\boldsymbol{x}_t$ and $\boldsymbol{s}_\tau$ along the channel dimension. The resulting input is embedded into a shared token sequence using non-overlapping $p \times p$ patches:

$$\boldsymbol{h}_0 = \mathrm{PatchEmbed}\left([\boldsymbol{x}_t, \boldsymbol{s}_\tau, \widetilde{\boldsymbol{y}}]\right), \qquad \widetilde{\boldsymbol{y}} = \mathrm{Resize}(\boldsymbol{y}). \tag{7}$$

Then, the tokens pass through N shared DiT blocks with dual-time AdaLN conditioning:

$$\boldsymbol{h}_{n+1} = \mathrm{DiT}_n\left(\boldsymbol{h}_n, t, \tau\right), \qquad n = 0, \ldots, N-1. \tag{8}$$

This backbone integrates the two noisy states and the LR observation, providing joint features for subsequent structure and image predictions.

State-Specific Prediction Branches. The shared features $\boldsymbol{h}_N$ are processed by two prediction branches, each containing M DiT blocks. The structure branch linearly projects its output tokens into structural patches and reshapes them into $\widehat{\boldsymbol{s}}_0$. The pixel branch instead uses a DiP patch decoder (Chen et al., 2026c) to produce the full-image prediction with fine-grained details:

$$\widehat{\boldsymbol{s}}_0 = \Phi_s(\boldsymbol{h}_N, t, \tau), \qquad \widehat{\boldsymbol{x}}_0 = \Phi_x(\boldsymbol{h}_N, t, \tau), \tag{9}$$

where Φ_s and Φ_x denote the complete structure and pixel branches, respectively.

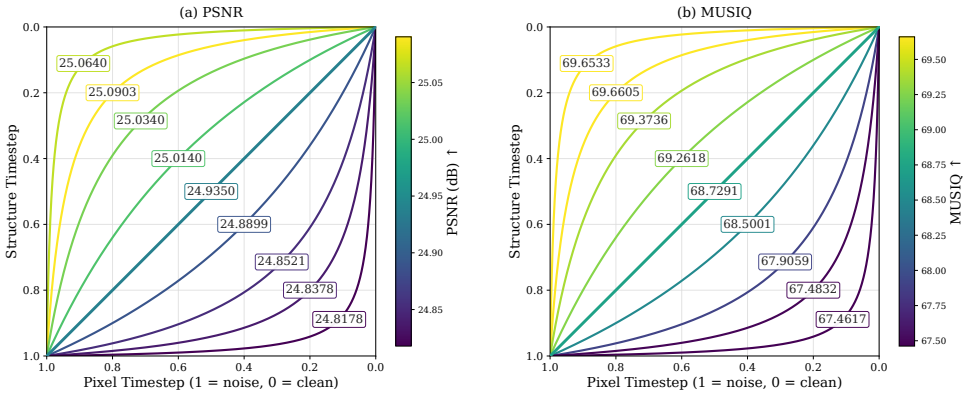


Figure 4: **Denoising trajectory analysis on RealSR.** Each curve represents a trajectory in the joint image–structure time plane, colored by its PSNR or MUSIQ averaged over five sampling seeds using the Dual-Time Exploration Model.

3.3 Structure-Forcing Schedule

Although DJD assigns independent time variables to structure and image, it does not explicitly define their relative denoising processes. We therefore compare joint denoising trajectories to identify an effective recovery order and develop the Structure-Forcing Schedule (SFS).

Parameterized Denoising Trajectories. We parameterize the joint denoising as a trajectory from $(1, 1)$ to $(0, 0)$ in the 2D time plane (t, τ). To control the relative denoising progress, we relate structural time τ to image time t through a time-shift mapping (Baade et al., 2026):

$$\tau = f_\alpha(t) = \frac{t}{\alpha + (1-\alpha)t}, \qquad \alpha > 0. \tag{10}$$

This mapping preserves both endpoints, while α controls the relative denoising progress:

$$\begin{cases} \alpha > 1: & \tau < t, & \text{structure-first,} \\ \alpha = 1: & \tau = t, & \text{synchronized,} \\ 0 < \alpha < 1: & \tau > t, & \text{image-first.} \end{cases} \qquad 0 < t < 1. \tag{11}$$

Thus, adjusting α controls the relative denoising progress of structure and image, allowing us to compare the effects of different recovery orders and degrees of structural precedence.

Analyzing Denoising Order. To analyze the effect of denoising order, we first train a Dual-Time Exploration Model covering different time pairs, allowing restoration along multiple trajectories. Then, we vary only α to change the inference trajectory while keeping the model weights fixed and initial noise samples, isolating the effect of relative denoising progress.

Figure 4 presents quantitative results for different trajectories using the Dual-Time Exploration Model on RealSR. Across the evaluated trajectories, structure-first denoising outperforms synchronized and image-first denoising in PSNR and MUSIQ. These results show that advancing structural denoising relative to image denoising improves restoration quality under the evaluated schedules.

Structure-Forcing Schedule. Based on the scheduling analysis, we adopt a suitable structure-first trajectory with $\alpha > 1$ and retrain the final model, Sketch2Pix-MS, along this fixed schedule. Specifically, we sample the image time t and set $\tau = f_\alpha(t)$, keeping all training time pairs on the selected trajectory. At inference, the model follows the same schedule, allowing progressively recovered structural cues to guide image refinement and realizing the *Sketch Before Pixel* paradigm.

Training Objective. Following Eqs. 2 and 3, we jointly optimize the image and structure predictions using a stabilized velocity-matching objective:

$$\mathcal{L}_{\text{joint}} = \mathbb{E}_{t,\tau,\boldsymbol{x}_0,\boldsymbol{y},\boldsymbol{\epsilon},\boldsymbol{\epsilon}_s} \left[\left\| \frac{\widehat{\boldsymbol{x}}_0 - \boldsymbol{x}_0}{\max(t,\delta)} \right\|_2^2 + \left\| \frac{\widehat{\boldsymbol{s}}_0 - \boldsymbol{s}_0}{\max(\tau,\delta)} \right\|_2^2 \right], \qquad t, \tau > 0. \tag{12}$$

where $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}_s$ are independent standard Gaussian noise. The choice $\delta = 1/3$ stabilizes training near the clean endpoint by downweighting low-noise states relative to exact velocity matching. The exploration model samples the two times independently, whereas the final model uses $\tau = f_\alpha(t)$.

Inference Procedure. Given an LR observation $\boldsymbol{y}$, we initialize both states with independent Gaussian noise and follow a decreasing time sequence $1 = t_0 > \cdots > t_K = 0$, with $\tau_k = f_\alpha(t_k)$. No HR-derived structure is provided at inference. At step k, the jointly predicted velocities give the Euler updates:

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + (t_{k+1} - t_k)\hat{\boldsymbol{v}}_\theta^k, \qquad \boldsymbol{s}^{k+1} = \boldsymbol{s}^k + (\tau_{k+1} - \tau_k)\hat{\boldsymbol{u}}_\theta^k. \tag{13}$$

where K denotes the number of sampling steps and $\boldsymbol{x}^K$ is the restored output.

3.4 One-Step Distillation for Efficient Inference

To accelerate inference, we distill Sketch2Pix-MS into the one-step student Sketch2Pix-OS by combining teacher velocity matching and shortcut consistency (Frans et al., 2025) with dual-anchor supervision from frozen teacher outputs and HR images (Qiao et al., 2026).

Because one-step inference cannot progressively recover structure, we provide a clean structural state at $\tau = 0$ and initialize the image state with Gaussian noise at $t = 1$. The frozen teacher uses the HR low-frequency structure $\boldsymbol{s}_0^{\mathrm{HR}}$, whereas the student uses $\boldsymbol{s}_0^{\mathrm{LR}}$ during both training and inference. We construct $\boldsymbol{s}_0^{\mathrm{LR}}$ by resizing the LR input to the HR resolution and applying Eq. (4), ensuring consistent spatial alignment and normalization. The one-step prediction is:

$$(\widehat{\boldsymbol{x}}_{\mathrm{S}}, \cdot) = f_\phi\left(\boldsymbol{\epsilon}, \boldsymbol{s}_0^{\mathrm{LR}}, t = 1, \tau = 0, \boldsymbol{y}\right), \tag{14}$$

where f_ϕ denotes the student network, $\widehat{\boldsymbol{x}}_{\mathrm{S}}$ is the restored image, and the structural output is unused.

The dual-anchor objective is:

$$\mathcal{L}_{\mathrm{anchor}} = \alpha_1 \left\|\widehat{\boldsymbol{x}}_{\mathrm{S}} - \widehat{\boldsymbol{x}}_{\mathrm{T}}\right\|_1 + \alpha_p \,\mathrm{LPIPS}\left(\widehat{\boldsymbol{x}}_{\mathrm{S}}, \widehat{\boldsymbol{x}}_{\mathrm{T}}\right) + \beta_1 \left\|\widehat{\boldsymbol{x}}_{\mathrm{S}} - \boldsymbol{x}_0\right\|_1 + \beta_p \,\mathrm{LPIPS}\left(\widehat{\boldsymbol{x}}_{\mathrm{S}}, \boldsymbol{x}_0\right), \tag{15}$$

where $\widehat{\boldsymbol{x}}_{\mathrm{T}}$ is the frozen teacher output. Together with velocity matching and shortcut consistency on both image and structural states, the overall objective is:

$$\mathcal{L}_{\mathrm{distill}} = \mathcal{L}_{\mathrm{vel}} + \mathcal{L}_{\mathrm{sc}} + \mathcal{L}_{\mathrm{anchor}}. \tag{16}$$

The detailed formulations and loss weights are provided in the **Appendix** A.2.

4 Experiments

4.1 Experimental Settings

Datasets. We train Sketch2Pix using publicly available datasets, including LSDIR (Li et al., 2023), DIV2K (Agustsson & Timofte, 2017), DIV8K (Gu et al., 2019), OutdoorSceneTraining (Wang et al., 2018), Flickr2K (Timofte et al., 2017), and the first 10K images from FFHQ (Karras et al., 2019). Following prior work (Cheng et al., 2025; Wu et al., 2026), we construct paired training samples by applying the Real-ESRGAN degradation pipeline (Wang et al., 2021) to HR images.

For synthetic evaluation, we use the LSDIR validation split (Li et al., 2023). We center-crop HR images to 512×512 and generate 128×128 LR inputs using the same Real-ESRGAN degradation pipeline as in training (Wang et al., 2021). For real-world evaluation, we use paired images from RealSR (Cai et al., 2019) and ScreenSR (Wu et al., 2026). On RealSR, we evaluate 512×512 HR center crops with corresponding 128×128 LR inputs. On ScreenSR, the supplied degraded inputs and GT images are both 512×512.

Evaluation Metrics. We evaluate all methods using complementary quality metrics. Specifically, PSNR and SSIM (Bakurov et al., 2022) measure reconstruction fidelity, while LPIPS (Zhang et al., 2018) and FID (Yu et al., 2021) assess perceptual and distribution-level similarity to reference images, respectively. Complementing these reference-based metrics, NIQE (Mittal et al., 2012), MUSIQ (Ke et al., 2021), MANIQA (Yang et al., 2022), and DeQA (You et al., 2025) assess perceptual quality without reference images.

Implementation Details. Sketch2Pix is a self-contained network trained end-to-end in pixel space without a pre-trained VAE. The shared DiT backbone contains $N = 20$ DiT blocks, followed by two prediction branches with $M = 8$ DiT blocks each. The pixel branch uses a DiP patch decoder (Chen et al., 2026c) with its default configuration. For the final model, we fix $\alpha = 16$ in the structure-forcing schedule during both training and inference, and use $K = 25$ sampling steps for the multi-step model. Further implementation details are provided in the **Appendix** A.1.

Table 1: Quantitative comparison on RealSR, LSDIR, and ScreenSR. The **best** and second-best results are highlighted within each dataset and inference setting based on unrounded values.

Dataset	Setting	Method	Metrics							
			PSNR↑	SSIM↑	LPIPS↓	NIQE↓	MUSIQ↑	MANIQA↑	DeQA↑	FID↓
RealSR	Multi-step	StableSR	24.39	0.6777	0.3194	6.26	47.87	0.2851	2.49	149.28
		DiT4SR	22.96	0.6537	0.3190	5.92	67.37	0.4559	3.40	121.22
		DiT-SR	**25.25**	**0.7328**	**0.2976**	5.57	65.83	0.4500	3.56	148.35
		VOSR	24.86	0.7024	0.3040	5.66	69.81	0.4944	3.58	135.65
		Sketch2Pix-MS	25.12	0.7098	0.3035	**4.51**	**70.09**	**0.5342**	**3.64**	**117.04**
	One-step	SinSR	25.54	**0.7217**	0.3181	6.28	60.84	0.4000	3.37	135.34
		OSEDiff	24.53	0.7166	0.2972	5.71	69.11	0.4710	**3.66**	123.12
		CODSR	24.73	0.7100	**0.2789**	5.24	**70.42**	0.4689	3.63	115.66
		Sketch2Pix-OS	**25.63**	0.7171	0.2843	**5.11**	68.34	**0.5157**	3.58	**107.01**
LSDIR	Multi-step	StableSR	20.49	0.5227	0.3193	4.60	62.72	0.3254	3.01	60.06
		DiT4SR	19.21	0.4804	0.2723	4.22	72.68	0.5067	3.87	59.03
		DiT-SR	**21.11**	**0.5490**	0.2457	4.30	69.57	0.4341	3.83	**41.79**
		VOSR	20.60	0.5111	0.2386	3.88	**72.95**	0.5194	**4.09**	46.49
		Sketch2Pix-MS	20.61	0.5112	**0.2339**	**3.74**	72.63	**0.5684**	4.00	45.38
	One-step	SinSR	20.92	0.5375	0.2268	4.66	69.64	0.4654	**3.93**	**39.65**
		OSEDiff	20.39	0.5215	0.2703	4.02	72.34	0.4586	3.84	59.54
		CODSR	20.33	0.5160	0.2499	**3.92**	**73.26**	0.4705	3.90	53.33
		Sketch2Pix-OS	**21.15**	**0.5399**	**0.2139**	4.25	70.89	**0.5398**	3.91	42.07
ScreenSR	Multi-step	StableSR	23.48	**0.7150**	0.2338	4.53	63.98	0.3629	3.24	72.43
		DiT4SR	20.72	0.5810	0.2617	4.17	71.15	0.4990	3.77	78.70
		DiT-SR	23.66	0.6751	0.2354	4.44	66.60	0.4099	3.52	71.94
		VOSR	22.29	0.6241	0.2167	4.22	**72.41**	**0.5190**	**4.00**	67.54
		Sketch2Pix-MS	**23.68**	0.6817	**0.2096**	**3.80**	71.11	0.5068	3.81	**62.53**
	One-step	SinSR	23.41	0.6619	0.2225	4.68	66.33	0.4281	3.66	65.82
		OSEDiff	22.12	0.6302	0.2302	4.08	71.43	0.4521	3.76	74.94
		CODSR	22.12	0.6253	0.2123	**4.06**	**72.44**	0.4705	**3.82**	67.57
		Sketch2Pix-OS	**24.66**	**0.7359**	**0.1703**	**4.06**	69.51	**0.4766**	3.78	**53.89**

4.2 Comparison with State-of-the-Art Methods

Compared Methods. We compare our Sketch2Pix with representative multi-step and one-step diffusion-based Real-SR methods. The multi-step baselines include StableSR (Wang et al., 2024a), DiT4SR (Duan et al., 2025), DiT-SR (Cheng et al., 2025), and VOSR (Wu et al., 2026). The one-step baselines include SinSR (Wang et al., 2024b), OSEDiff (Wu et al., 2024a), and CODSR (Chen et al., 2026b). All baselines are evaluated using their official checkpoints and inference settings. Unlike these latent-space approaches, Sketch2Pix performs restoration directly in RGB space within a fully self-contained network, requiring neither a pre-trained VAE nor an external encoder at inference.

Quantitative Results. Table 1 presents quantitative results on three benchmarks under multi-step and one-step settings. Overall, existing methods exhibit distinct strengths in reconstruction fidelity and perceptual realism. Specifically, DiT-SR and SinSR achieve strong reconstruction fidelity but are less competitive on several perceptual metrics, whereas VOSR and CODSR obtain high no-reference scores with comparatively weaker pixel-level fidelity. In contrast, our Sketch2Pix-MS maintains competitive fidelity while achieving strong perceptual results, thereby effectively alleviating the fidelity–realism seesaw effect. Meanwhile, the distilled Sketch2Pix-OS achieves strong reconstruction accuracy while retaining competitive perceptual quality, demonstrating that efficient one-step inference does not necessarily sacrifice restoration quality. These results further confirm that Sketch2Pix is a powerful pixel-space restoration framework that delivers faithful reconstruction and realistic generation in both multi-step and one-step settings.

Qualitative Results. Figure 5 provides qualitative comparisons on challenging real-world examples from RealSR and ScreenSR. In the text example, multi-step baselines tend to produce deformed letters and irregular logo shapes, whereas our Sketch2Pix-MS reconstructs their shapes more faithfully to the HR reference. Similarly, in the window example, competing one-step baselines exhibit

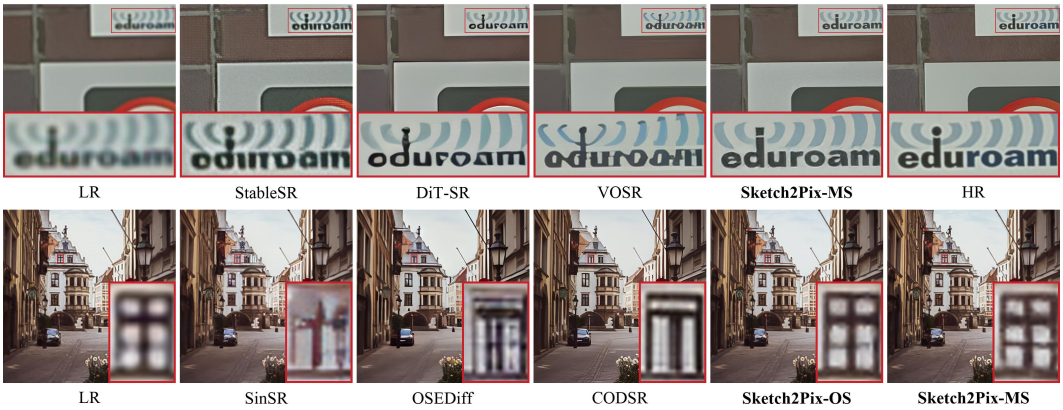


Figure 5: **Qualitative comparison on challenging real-world cases.** Sketch2Pix faithfully preserves input information while recovering more realistic details, whereas competing methods exhibit distorted characters, altered window divisions, or over-smoothed results. Best viewed zoomed in.

Table 2: **Quality and complexity comparisons on LSDIR.** Runtime is measured for 512×512 outputs on an H800 GPU with batch size 1. "NFE" denotes number of evaluations.

Method	NFE	PSNR ↑	MANIQA ↑	Time (s) ↓	Params (M) ↓
StableSR	200	20.49	0.3254	5.79	1540
DiT4SR	40	19.21	0.5067	5.20	21380
VOSR	25	20.60	0.5194	3.84	619.40
DiT-SR	15	**21.11**	0.4341	**1.84**	**61.00**
Sketch2Pix-MS	25	20.61	**0.5684**	2.67	475.00
SinSR	1	20.92	0.4654	0.32	174.00
OSEDiff	1	20.39	0.4586	0.48	1760
CODSR	1	20.33	0.4705	6.37	**35.79**
Sketch2Pix-OS	1	**21.15**	**0.5398**	**0.28**	476.00

Table 3: **Denoising schedule analysis on ScreenSR and LSDIR.** The exploration model uses varying α, while Sketch2Pix-MS is retrained with a fixed schedule at $\alpha = 16$.

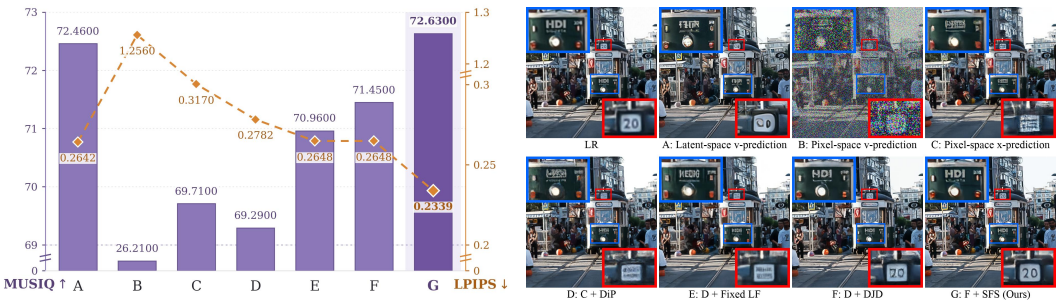
Model	α	ScreenSR		LSDIR	
		MUSIQ↑	MANIQA↑	MUSIQ↑	MANIQA↑
Dual-Time Exploration Model	1/64	69.43	0.4559	70.75	0.5109
	1/16	69.43	0.4496	70.78	0.5194
	1/4	69.47	0.4440	70.92	0.5176
	1	69.60	0.4516	71.38	0.5255
	4	70.05	0.4617	71.96	0.5349
	16	**70.43**	**0.4728**	**72.27**	**0.5425**
	64	70.31	0.4648	72.10	0.5370
Sketch2Pix-MS (fixed schedule)	16	**71.11**	**0.5068**	**72.63**	**0.5684**

inaccurate structures and overly smooth or blurred details, whereas Sketch2Pix-OS effectively preserves the observed layout and recovers clearer internal details. Despite performing distillation, it maintains visual quality comparable to its multi-step model. These results complement the quantitative results, demonstrating that our Sketch2Pix better achieves structural fidelity and perceptual realism through end-to-end pixel-space restoration.

Efficiency Comparison. Table 2 compares restoration quality, latency, and model size on LSDIR under a unified evaluation protocol. Sketch2Pix-MS achieves higher MANIQA than VOSR with fewer parameters and a $1.44\times$ inference speedup. DiT-SR achieves lower latency with fewer parameters, but its lower MANIQA indicates weaker perceptual quality. To further accelerate inference, we distill Sketch2Pix-MS into Sketch2Pix-OS, reducing runtime from 2.67 s to 0.28 s and achieving a $9.5\times$ speedup. Among the listed one-step methods, Sketch2Pix-OS achieves the highest PSNR and MANIQA and the lowest measured latency, running $1.14\times$, $1.71\times$, and $22.75\times$ faster than SinSR, OSEDiff, and CODSR, respectively. Although larger than SinSR and CODSR, it offers a favorable balance between restoration quality and inference speed.

4.3 Analysis of Dual-Time Structure Forcing

To examine the degree of structural precedence and the effect of fixed-schedule training, we further compare different schedules on ScreenSR and LSDIR in Table 3. Consistent with the trend observed on RealSR in Figure 4, structure-first trajectories ($\alpha > 1$) improve both MUSIQ and MANIQA over synchronized denoising ($\alpha = 1$), whereas image-first trajectories ($\alpha < 1$) generally perform worse. Increasing α from 16 to 64 yields little change in MUSIQ but lowers MANIQA on both datasets, indicating that a stronger structural lead is not always beneficial. Using $\alpha = 16$, selected from the RealSR analysis, we retrain the model along this fixed trajectory, further improving both metrics on ScreenSR and LSDIR. These results further support the benefit of moderate structural precedence and fixed-schedule training across datasets.



(a) Quantitative Ablation (b) Qualitative Ablation

Figure 6: **Ablation on key components.** (a) Staged quantitative ablation of representation space, prediction parameterization, decoder design, structural conditioning, and scheduling. (b) Qualitative comparisons showing the effects of these components on text fidelity and detail recovery.

4.4 Ablation Studies

Ablation on Key Components. Figure 6 presents a staged ablation from a latent-space v-prediction baseline to the complete Sketch2Pix-MS model. Specifically, variants A–C examine the representation space and prediction parameterization, while the C–D comparison evaluates the decoder. With the backbone configuration, pixel decoder, and other training settings held fixed, variants D–G compare image-only restoration, fixed LR-derived structural conditioning, DJD, and DJD with SFS. Directly transferring v-prediction to pixel space produces substantial residual noise and severely degrades performance. In contrast, switching to clean-image x_0-prediction substantially improves both metrics, supporting its effectiveness for pixel-space restoration. Adding the DiP decoder further improves LPIPS but slightly reduces MUSIQ. To assess the benefit of structural cues, we then add fixed LR-derived low-frequency conditioning, improving both metrics without predicting or denoising the structure. Replacing this conditioning with DJD introduces a dedicated branch for joint structure and image denoising, further improving MUSIQ while leaving LPIPS unchanged at the reported precision. Finally, SFS coordinates their relative denoising progress, reducing LPIPS and achieving the highest MUSIQ among the evaluated variants. The qualitative example also illustrates corresponding improvements in text fidelity and local detail clarity.

Ablation on Pixel Decoders. To evaluate the quality and efficiency of different pixel decoders, we compare PixelDiT's Transformer-based head (PiT) (Yu et al., 2026), DeCo's frequency-decoupled head (Ma et al., 2026), and DiP's lightweight convolutional U-Net (Chen et al., 2026c) under identical backbone and training settings. As shown in Table 4, DiP achieves the highest MUSIQ with the fewest parameters and lowest latency. Compared with PiT, it improves both SSIM and MUSIQ at substantially lower cost. Although DeCo achieves higher SSIM and fewer FLOPs, it yields poorer perceived quality and nearly twice the latency of DiP. Based on this quality–efficiency trade-off, we adopt DiP as our decoder head.

Table 4: **Pixel decoder comparisons on ScreenSR.** Inference latency and GFLOPs are measured with decoder-only inference for obtaining a 512 × 512 image on a single H800 GPU.

Metric	PiT	DeCo	**DiP (Ours)**
SSIM ↑	0.6273	**0.6984**	0.6817
MUSIQ ↑	70.90	69.52	**71.11**
Params (M) ↓	285.49	8.42	**6.91**
Latency (ms) ↓	34.53	13.02	**6.70**
FLOPs (G)↓	1340.43	**54.73**	277.29

Latency averaged over 100 runs after 20 warm-up runs.

5 Conclusion

In this paper, we propose Sketch2Pix, a self-contained, single-stage, end-to-end framework that explicitly coordinates structural recovery and detail generation for pixel-space Real-SR. Unlike methods that couple both processes along a single denoising trajectory, Sketch2Pix allows structural recovery to precede image refinement, realizing the *Sketch Before Pixel* paradigm without VAE compression. Specifically, we introduce Dual-Time Joint Diffusion to jointly model low-frequency structure and the full image, together with a Structure-Forcing Schedule to coordinate their relative denoising progress. To reduce inference cost, we further distill the multi-step model into a one-step student while retaining competitive restoration quality. Experimental results show that Sketch2Pix achieves a favorable balance between reconstruction fidelity, perceptual quality, and inference efficiency, supporting recovery-order coordination as a promising approach to pixel-space Real-SR.

AI USE STATEMENT

In this work, we used generative AI tools to polish the content. We did not use generative AI tools for data generation or concept design. We have reviewed all work completed with the assistance of AI. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The code will be released after the paper is accepted.

REFERENCES

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

Alan Baade, Eric Ryan Chan, Kyle Sargent, Changan Chen, Justin Johnson, Ehsan Adeli, and Li Fei-Fei. Latent forcing: Reordering the diffusion trajectory for pixel-space image generation. *arXiv preprint arXiv:2602.11401*, 2026.

Illya Bakurov, Marco Buzzelli, Raimondo Schettini, Mauro Castelli, and Leonardo Vanneschi. Structural similarity index (ssim) revisited: A data-driven approach. *Expert Systems with Applications*, 189:116087, 2022.

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

Hansheng Chen, Jan Ackermann, Minseo Kim, Gordon Wetzstein, and Leonidas Guibas. Asymmetric flow models. *arXiv preprint arXiv:2605.12964*, 2026a.

Hao Chen, Junyang Chen, Jinshan Pan, and Jiangxin Dong. Bridging fidelity-reality with controllable one-step diffusion for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30584–30594, 2026b.

Zhennan Chen, Junwei Zhu, Xu Chen, Jiangning Zhang, Xiaobin Hu, Hanzhen Zhao, Chengjie Wang, Jian Yang, and Ying Tai. Dip: Taming diffusion models in pixel space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36136–36146, 2026c.

Kun Cheng, Lei Yu, Zhijun Tu, Xiao He, Liyu Chen, Yong Guo, Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Hu. Effective diffusion transformer architecture for image super-resolution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 2455–2463, 2025.

Lucas Degeorge, Paul Couairon, Arijit Ghosh, Alexei A Efros, David Picard, and Vicky Kalogeiton. Balancing frequencies and pixels in flow matching. *arXiv preprint arXiv:2609.02748*, 2026.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *CoRR*, abs/1501.00092, 2015. URL http://arxiv.org/abs/1501.00092.

Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S Ren, Chunle Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18948–18958, 2025.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OlzB6LnXcS.

Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. Div8k: Diverse 8k resolution image dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3512–3516. IEEE, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

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

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

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 105–114. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.19. URL https://doi.org/10.1109/CVPR.2017.19.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

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

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, QC, Canada, October 11-17, 2021*, pp. 1833–1844. IEEE, 2021. doi: 10.1109/ICCVW54120.2021.00210. URL https://doi.org/10.1109/ICCVW54120.2021.00210.

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

Mingfeng Lin, Jiakun Chen, Liang Han, and Liqiang Nie. Frepix: Frequency-heterogeneous flow matching for pixel-space image generation. *arXiv preprint arXiv:2605.06421*, 2026.

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

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. Deco: Frequency-decoupled pixel diffusion for end-to-end image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43600–43610, 2026.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

Bingtian Qiao, Yue Shi, Yong Guo, Wenjun Zhang, and Jiezhang Cao. Pixelir: Fidelity-perception decoupling via pixel-space image-residual flow matching for efficient one-step real-world super-resolution, 2026. URL https://arxiv.org/abs/2608.30782.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Lingchen Sun, Rongyuan Wu, Xiangtao Kong, Jixin Zhao, Qiaosi Yi, Yujing Sun, Shuaizheng Liu, Zhengqiang Zhang, and Lei Zhang. Pixrestore: Unified image restoration via pixel diffusion transformer. *arXiv preprint arXiv: https://arxiv.org/pdf/2608.16793*, 2026.

Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 114–125, 2017.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. 2024a.

Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 606–615, 2018.

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*, pp. 1905–1914, 2021.

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

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37: 92529–92553, 2024a.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25456–25467, 2024b.

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*, pp. 16311–16321, 2026.

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

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14483–14494, 2025.

Yanchun Yu, Weibin Zhang, and Yun Deng. Frechet inception distance (fid) for evaluating gans. *China University of Mining Technology Beijing Graduate School*, 3(11), 2021.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. Pixeldit: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14273–14282, 2026.

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

A Appendix

A.1 More Implementation Details

Table 5 summarizes the architecture and training setup of Sketch2Pix-MS, including key model specifications and training hyperparameters for reproducibility.

Table 5: Architecture and training settings of the Sketch2Pix-MS model.

Hyperparameter	**Sketch2Pix-MS**
Architecture	
Restoration resolution	$512 \times 512 \times 3$
Concatenated input size	$512 \times 512 \times 9$
Shared DiT depth N	20
DiT depth per branch M	8
DiT hidden dimension	1024
Attention heads	16
Patch size p	16
MLP ratio	4
Positional encoding	RoPE
Time conditioning	Dual-time AdaLN
Structure output head	Linear projection
Pixel decoder	DiP (Chen et al., 2026c)
DiP base channels	64
External pre-trained encoder	None
Total parameters (M)	≈ 475
Structural Representation	
Structural target source	HR image $\boldsymbol{x}_0$
Wavelet transform	DWT and IDWT
Retained subband	LL
Structural target size	$512 \times 512 \times 3$
Dual-Time Training	
Prediction targets	Clean image and structure
Training objective	Velocity matching
Global-time distribution	$\mathcal{U}(0, 1)$
Training schedule	Fixed structure-first trajectory
Relative scheduling parameter α	16
Optimization	
Training steps	500K
Batch size per process	32
Learning rate	1×10^{-5}
Adam coefficients (β_1, β_2)	$(0.9, 0.95)$
Optimizer epsilon	1×10^{-8}
Weight decay	1×10^{-2}
Learning-rate schedule	Constant with warmup
Warmup steps	5,000
Gradient clipping	1.0
Mixed precision	BF16
EMA decay	0.9999
Training seed	42
Inference	
Sampler	Euler-based ODE solver
Joint sampling steps K	25
Relative scheduling parameter α	16
Restoration space	RGB pixel space

A.2 One-Step Distillation Details

We detail the velocity matching, shortcut consistency, and dual-anchor objectives in Eq. 16.

Teacher Velocity Matching. The teacher and student use HR-derived and LR-derived low-frequency structure, respectively. Their noisy structural states share the same Gaussian noise:

$$\boldsymbol{s}_\tau^{\mathrm{HR}} = (1-\tau)\boldsymbol{s}_0^{\mathrm{HR}} + \tau\boldsymbol{\epsilon}_s, \qquad \boldsymbol{s}_\tau^{\mathrm{LR}} = (1-\tau)\boldsymbol{s}_0^{\mathrm{LR}} + \tau\boldsymbol{\epsilon}_s, \tag{17}$$

where $\boldsymbol{s}_0^{\mathrm{HR}} = \boldsymbol{s}_0$. Both networks receive the same noisy image state $\boldsymbol{x}_t$, while their structural inputs are $\boldsymbol{s}_\tau^{\mathrm{HR}}$ and $\boldsymbol{s}_\tau^{\mathrm{LR}}$, respectively.

We convert the clean predictions into velocities Eq. 3 and match the student and frozen teacher velocities:

$$\mathcal{L}_{\mathrm{vel}} = \mathbb{E}\left[\|\widehat{\boldsymbol{v}}_\phi - \widehat{\boldsymbol{v}}_\theta\|_2^2 + \|\widehat{\boldsymbol{u}}_\phi - \widehat{\boldsymbol{u}}_\theta\|_2^2\right], \tag{18}$$

where ϕ and θ denote the student and frozen teacher, respectively.

Shortcut Consistency. Shortcut consistency encourages the student to reach the same state through either one denoising step or two smaller steps over the same time interval. The student additionally receives the destination structural time as a condition. For velocity matching, this condition equals the current structural time.

For global times $q > r$, we introduce the midpoint $m = (q+r)/2$. The corresponding local time pairs are (t_q, τ_q), (t_m, τ_m), and (t_r, τ_r), following the fixed dual-time schedule. Starting from $(\boldsymbol{x}_{t_q}, \boldsymbol{s}_{\tau_q}^{\mathrm{LR}})$, the student predicts $(\widehat{\boldsymbol{v}}_1, \widehat{\boldsymbol{u}}_1)$ for $q \to m$ and updates both states:

$$\widetilde{\boldsymbol{x}}_{t_m} = \boldsymbol{x}_{t_q} - (t_q - t_m)\widehat{\boldsymbol{v}}_1, \qquad \widetilde{\boldsymbol{s}}_{\tau_m} = \boldsymbol{s}_{\tau_q}^{\mathrm{LR}} - (\tau_q - \tau_m)\widehat{\boldsymbol{u}}_1. \tag{19}$$

From these intermediate states, it predicts $(\widehat{\boldsymbol{v}}_2, \widehat{\boldsymbol{u}}_2)$ for $m \to r$. We construct the two-step targets as:

$$\overline{\boldsymbol{v}} = \frac{(t_q - t_m)\widehat{\boldsymbol{v}}_1 + (t_m - t_r)\widehat{\boldsymbol{v}}_2}{\max(t_q - t_r, 10^{-6})}, \quad \overline{\boldsymbol{u}} = \frac{(\tau_q - \tau_m)\widehat{\boldsymbol{u}}_1 + (\tau_m - \tau_r)\widehat{\boldsymbol{u}}_2}{\max(\tau_q - \tau_r, 10^{-6})}. \tag{20}$$

The local time increments account for the nonlinear mapping between global time and the two branch times. The direct predictions for $q \to r$ are trained against these targets:

$$\mathcal{L}_{\mathrm{sc}} = \mathbb{E}\left[\left\|\widehat{\boldsymbol{v}}_\phi^{q\to r} - \mathrm{sg}[\overline{\boldsymbol{v}}]\right\|_2^2 + \left\|\widehat{\boldsymbol{u}}_\phi^{q\to r} - \mathrm{sg}[\overline{\boldsymbol{u}}]\right\|_2^2\right], \tag{21}$$

where sg denotes stop-gradient. Both shorter transitions are computed without gradient tracking.

Dual-Anchor Supervision. The endpoint objective follows Eq. 15. The teacher uses clean HR-derived structure to generate $\widehat{\boldsymbol{x}}_{\mathrm{T}}$, whereas the student predicts $\widehat{\boldsymbol{x}}_{\mathrm{S}}$ from Gaussian image noise and clean LR-derived structure. We set the teacher-anchor weights to $\alpha_1 = \alpha_p = 0.01$ and the HR-anchor weights to $\beta_1 = \beta_p = 0.001$. The complete objective is:

$$\mathcal{L}_{\mathrm{distill}} = \mathcal{L}_{\mathrm{vel}} + \mathcal{L}_{\mathrm{sc}} + \mathcal{L}_{\mathrm{anchor}}. \tag{22}$$

At inference, the destination-time condition is fixed at zero. The student restores the image with $t = 1$ and $\tau = 0$ in one network evaluation, and the structural output is unused.

A.3 Additional Qualitative Comparisons

We provide four additional visual comparisons with other methods on challenging examples in Figures 7 and 8. Specifically, Figure 7 presents two challenging text and symbol examples. In the street-sign example, competing methods often produce distorted symbols or irregular characters, whereas our Sketch2Pix-MS and Sketch2Pix-OS recover a more coherent symbol outline and a more legible "ZONE" inscription. For Chinese characters, DiT-SR and SinSR leave blurred or merged strokes, while other methods introduce malformed patterns. In contrast, Sketch2Pix produces more recognizable characters with better-separated strokes and more consistent layouts.



Figure 7: Additional visual comparisons on challenging text and symbol examples. Our Sketch2Pix-MS and Sketch2Pix-OS improve character legibility and stroke coherence. Zoom in for details.

Furthermore, Figure 8 compares results on feather textures and roof patterns. In the eagle example, several competing methods produce overly smooth feather surfaces or strongly sharpened boundaries, whereas Sketch2Pix-MS and Sketch2Pix-OS retain finer and more realistic textures. In the building example, DiT4SR, OSEDiff, and CODSR introduce irregular or densely textured roof patterns. Our methods produce more coherent tile arrangements while retaining the overall roof geometry and providing clearer local details.

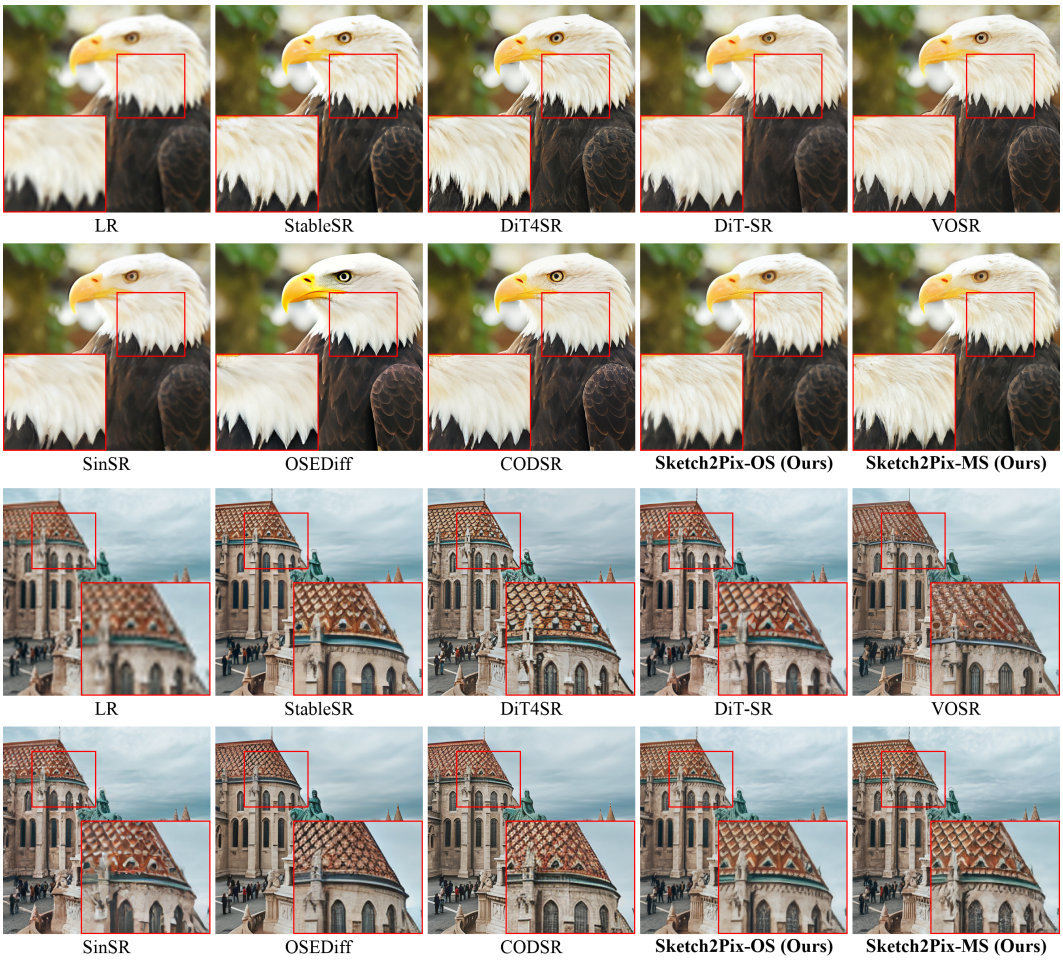


Figure 8: Additional visual comparisons with competing methods on challenging examples, highlighting fine-texture restoration and structural coherence. Zoom in for details.