Accelerating Diffusion-based Super-Resolution with Dynamic Time-Spatial Sampling

Rui Qin[1,†], **Qijie Wang**[1,†], **Ming Sun**[2], **Haowei Zhu**[1], **Chao Zhou**[2] and **Bin Wang**[1,✉]
[1]School of Software, Tsinghua University, Beijing, China
[2]Kuaishou Technology, Beijing, China
qr20@mails.tsinghua.edu.cn, wqj24@mails.tsinghua.edu.cn, sunming03@kuaishou.com, zhuhw23@mails.tsinghua.edu.cn, zhouchao@kuaishou.com, wangbins@tsinghua.edu.cn

Abstract

Diffusion models have gained attention for their success in modeling complex distributions, achieving impressive perceptual quality in SR tasks. However, existing diffusion-based SR methods often suffer from high computational costs, requiring numerous iterative steps for training and inference. Existing acceleration techniques, such as distillation and solver optimization, are generally task-agnostic and do not fully leverage the specific characteristics of low-level tasks like super-resolution (SR). In this study, we analyze the frequency- and spatial-domain properties of diffusion-based SR methods, revealing key insights into the temporal and spatial dependencies of high-frequency signal recovery. Specifically, high-frequency details benefit from concentrated optimization during early and late diffusion iterations, while spatially textured regions demand adaptive denoising strategies. Building on these observations, we propose the Time-Spatial-aware Sampling strategy (TSS) for the acceleration of Diffusion SR without any extra training cost. TSS combines Time Dynamic Sampling (TDS), which allocates more iterations to refining textures, and Spatial Dynamic Sampling (SDS), which dynamically adjusts strategies based on image content. Extensive evaluations across multiple benchmarks demonstrate that TSS achieves state-of-the-art (SOTA) performance with significantly fewer iterations, improving MUSIQ scores by $0.2 \sim 3.0$ and outperforming the current acceleration methods with only half the number of steps.

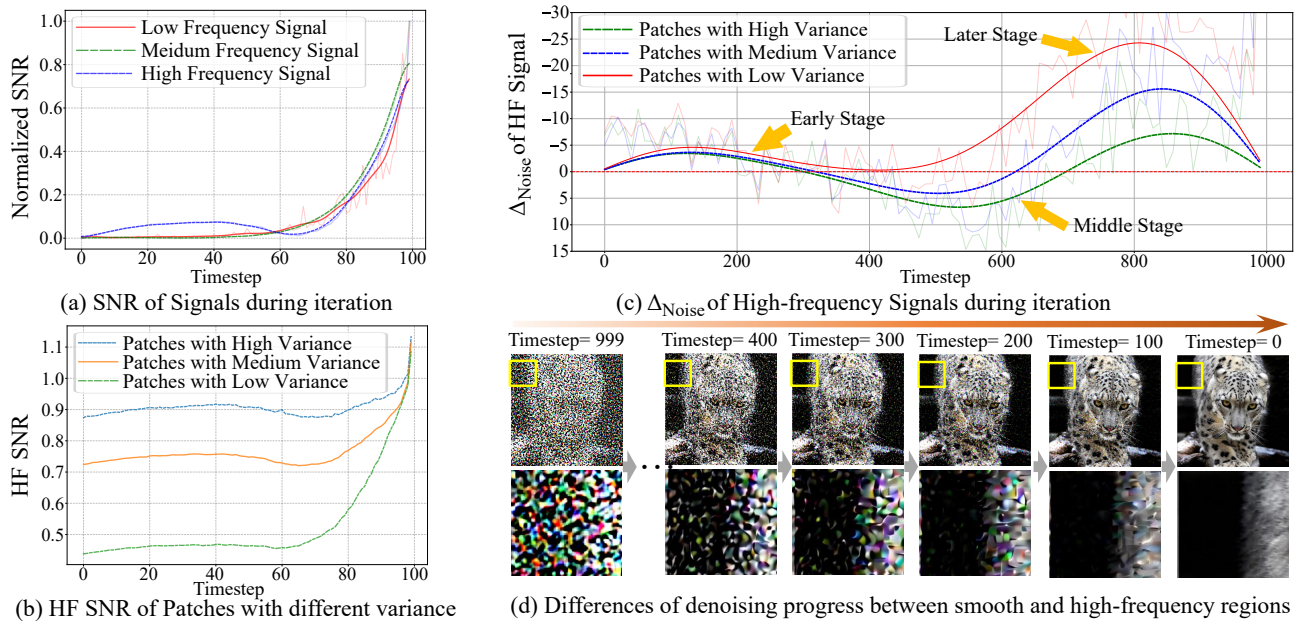
1 Introduction

Image super-resolution (SR) [Wang *et al.*, 2021b; Zhang *et al.*, 2021; Liang *et al.*, 2021; Qin *et al.*, 2023; Liu *et al.*, 2023; Zhao *et al.*, 2023; Qin *et al.*, 2024; Bao *et al.*, 2025] aims to reconstruct high-resolution (HR) images from low-resolution (LR) inputs. Recently, diffusion models [Ho *et al.*, 2020; Song *et al.*, 2020] have gained attention for their ability to model complex distributions, achieving notable success in SR [Chen *et al.*, 2024; Wang *et al.*, 2024b; Yang *et al.*, 2025; Yu *et al.*, 2024; Wu *et al.*, 2024b; Qu *et al.*, 2025], particularly in perceptual quality [Wang *et al.*, 2023; Wu *et al.*, 2023]. Diffusion-based image super-resolution methods take two primary approaches: integrating the low-resolution image into a task-specific denoiser [Saharia *et al.*, 2022; Wang *et al.*, 2024a] or adapting the reverse diffusion process of pre-trained models [Wu *et al.*, 2024b; Yang *et al.*, 2025; Yu *et al.*, 2024]. These methods are computationally intensive, requiring 1000 steps for training and several, such as 20 (PASD [Yang *et al.*, 2025]), 50 (SUPIR [Yu *et al.*, 2024]), or more steps (StableSR [Wang *et al.*, 2024a]) during testing.

Efforts to accelerate denoising generation focus on sampler acceleration and distillation, achieving results in 10 or fewer steps [Yue *et al.*, 2024; Wang *et al.*, 2024b]. Most Diffusion SR methods [Wang *et al.*, 2024a; Yu *et al.*, 2024; Yang *et al.*, 2025] adopt these general strategies without considering the unique frequency characteristics of low-level vision tasks. However, in fact, recent studies like STAR [Xie *et al.*, 2025] have highlighted the diverse recovery of diffusion-based SR across frequency domains, suggesting the potential to learn low- and high-frequency information at different training stages. Despite these insights, these works primarily focus on modifying the training process and optimization losses. Given the availability of many large-scale open source and pre-trained diffusion SR models [Yang *et al.*, 2025; Yu *et al.*, 2024; Wang *et al.*, 2024a], we aim to develop a tailored training-free acceleration strategy by leveraging the characteristics of Diffusion SR methods with the spatial and frequency information, seeking to enhance the performance of these existing models at minimal cost.

To analyze frequency-based disparities in the denoising process, we conduct a tiny experiment, using SUPIR, one of the latest typical state-of-the-art Diffusion SR methods, on the RealPhoto60 dataset [Yu *et al.*, 2024]. RealPhoto60 comprises 60 real-world images from common benchmarks. To explore the frequency characteristics, we applied Fourier transformation [Cochran *et al.*, 1967] to categorize spectra into low, medium, and high-frequency signals. To observe the time domain dynamics, we recorded the signal-to-noise ratio (SNR) of intermediate and final outputs over the 100-step inference. As shown in Fig. 1.a, SNR improvements were most pronounced in the later stages across all frequency

[1]The full version is available at https://arxiv.org/abs/2505.12048.
[2]✉ indicates the corresponding author of the paper.



(a) SNR of Signals during iteration

(b) HF SNR of Patches with different variance

(c) Δ_{Noise} of High-frequency Signals during iteration

(d) Differences of denoising progress between smooth and high-frequency regions

Figure 1: (a) SNR of different frequency components in SUPIR denoising, where high-frequency signals show a unique two-stage pattern. (b) SNR of high-frequency signals in different spatial regions during the denoising process of SUPIR. (c) Noise amplitude in high-frequency regions of SUPIR denoising: higher variance shortens the positive optimization phase. (d) Denoising visualization of a RealPhoto60 sample.

bands. Notably, **unlike low and medium frequency components, high-frequency components uniquely exhibited visible SNR gains in the early stages**, indicating the critical role of early denoising in restoring high-frequency details.

Furthermore, we analyzed spatial variations in high-frequency recovery by cropping RealPhoto60 samples into 128-sized patches and categorizing them as smooth, medium-textured, or highly textured based on variance. We recorded the SNR of high-frequency signals for each category throughout the iterations (Fig. 1.b), and analyzed the noise magnitude changes (Fig. 1.c). For an intuitive comparison, Fig. 1.d illustrates the denoising process for a typical sample with both low- and high-frequency regions, showing that smooth areas recover early, while high-frequency regions, such as fur, recover significantly in the final $0 \sim 100$ steps. **Spatially, regions with more high-frequency information concentrate recovery in the initial and final stages, whereas content variation influences denoising dynamics.**

Based on the above observations, efficiently generating high-frequency details requires leveraging their unique temporal optimization, which is concentrated in the early and late iterations. Therefore, we propose a Time Dynamic Sampling (TDS) strategy that prioritizes high-frequency signal recovery and enhances texture perception by allocating more denoising steps to diffusion stages critical for refining high-frequency details. Furthermore, from a spatial perspective, the sampling strategy must adapt to variations in image content. To achieve this, we introduce Spatial Dynamic Sampling (SDS), which dynamically adjusts the sampling frequency based on spatial content, ensuring alignment with the characteristics of different image regions. By integrating these two strategies, we propose Time-Spatial-Aware Sampling (TSS), a novel framework to accelerate existing diffusion-based SR without additional training costs. As both strategies require no additional training and only minimal code modifications, TSS offers an efficient and broadly applicable solution.

Evaluations on six BSR benchmarks across various metrics illustrate that TSS significantly improves the performance of various diffusion SR methods within a few iterations, without incurring additional training costs. TSS consistently achieves an increase of $0.2 \sim 3.0$ of MUSIQ in diverse SR diffusion frameworks and datasets. Remarkably, TSS outperforms the current state-of-the-art method while using only half the number of steps. Our main contributions are as follows:

1. We identified the temporal and spatial dynamics of diffusion-based methods in high-frequency detail recovery of image super-resolution tasks.
2. Based on the observations, we propose the Time-Spatial-aware Sampling strategy (TSS) to achieve training-free acceleration for diffusion-based image super-resolution.
3. Comprehensive evaluations across multiple real-world SR benchmarks show that TSS achieves state-of-the-art performance with fewer denoising iterations.

2 Related Work

2.1 Diffusion Model in Super-Resolution

Due to their exceptional ability to generate high-quality images, diffusion-based super-resolution methods have garnered widespread attention. Early approaches leveraged low-resolution images as guidance by training a conditional DDPM [Ho *et al.*, 2020; Kawar *et al.*, 2022; Saharia *et al.*, 2022], or conditionally steering a pre-trained DDPM [Choi *et al.*, 2021] to perform super-resolution tasks. Recently,

several studies have utilized pre-trained text-to-image (T2I) models, such as Stable Diffusion (SD) [Rombach *et al.*, 2022; Podell *et al.*, 2023], harnessing learned priors to address super-resolution challenges and achieve high-quality image enhancements. These methods either train a ControlNet [Zhang *et al.*, 2023; Yang *et al.*, 2025; Yu *et al.*, 2024; Chen *et al.*, 2024] or an additional encoder that encodes guidance features [Wu *et al.*, 2024b; Wang *et al.*, 2024a], both of which have demonstrated outstanding performance. Yet, as mentioned earlier, these methods require a large number of diffusion steps, resulting in high computational costs.

2.2 Diffusion Model Acceleration

Recent state-of-the-art super-resolution methods still require dozens or even hundreds of diffusion steps, even with acceleration techniques like DDIM [Song *et al.*, 2020], leading to significant time overhead. Reducing diffusion steps often degrades output quality. Various pruning [Zhu *et al.*, 2024], caching [Ma *et al.*, 2024], and distillation [Yin *et al.*, 2024] techniques have accelerated general T2I diffusion models while preserving generation quality. In super-resolution, ResShift [Yue *et al.*, 2024] reduces diffusion steps by modeling a Residual Shifting Markov chain between high-resolution and low-resolution images but requires a considerable cost of 500K iterations training from scratch. Distillation approaches [Wu *et al.*, 2024a; He *et al.*, 2024] have achieved one-step diffusion in super-resolution tasks, but their training costs remain high, and their performance is constrained by the teacher models. Therefore, given the training costs and limitations of existing methods, exploring a training-free acceleration framework for diffusion-based super-resolution is valuable and necessary.

3 Method

3.1 Revisiting Diffusion in Super-Resolution

To investigate the relationship between frequency signals, spatial characteristics, and temporal steps in denoising, we conducted two demo experiments using the representative SOTA method SUPIR [2024], a classic framework based on pre-trained SD [2022] with fine-tuned ControlNet [2023]. The experiments are conducted on the RealPhoto60 dataset containing 60 real-world images from benchmarks like RealSR [2020], DRealSR [2020], and web sources.

Frequency Signal Analysis We analyze the relationship between denoising stages and frequency-specific signal recovery by computing the signal-to-noise ratio (SNR) over 100-step inference using the Fourier Transform, with the final denoised result as a noise-free reference and intermediate results decoded from feature representations. As shown in Fig. 1.a, SNR increases significantly in the late stages (400 $\sim$ 0 steps). Notably, high-frequency (HF) components, unlike low and medium frequencies, also exhibit visible SNR improvement in the early stages (1000 $\sim$ 700 steps). To further investigate HF signals crucial for SR, Fig. 1.c presents the noise delta throughout denoising, where noise represents the residual with the final result, and delta denotes stepwise changes. Noise decreases in the early (1000 $\sim$ 700) and late (400 $\sim$ 0) stages but rises in the middle, mirroring the SNR trend in Fig. 1.a. This suggests that, unlike low and medium frequencies, early denoising stages are also crucial for HF recovery, while intermediate stages may hinder optimization.

Spatial Dynamics Analysis To examine the spatial dynamics of denoising across regions with varying content, we analyzed signals in image patches of different textures. RealPhoto60 images were divided into non-overlapping 128×128 patches and classified as smooth, medium textured, or highly textured based on variance, applying Frequency Signal Analysis above to each category. As shown in Fig. 1.c, higher variance patches have a narrower range of time steps for effective optimization in the early and late stages. Fig. 1.d further illustrates that smooth regions undergo visible denoising earlier than fur-textured areas during late iterations. This suggests that denoising effectiveness varies across textures, with high-texture patches containing more high-frequency signals exhibiting shorter, more concentrated optimization stages in both early and late iterations.

In summary, the restoration of high-frequency (HF) signals is unevenly distributed across both the denoising process and spatial regions within an image. However, the widely used traditional uniform, data-independent sampling methods haven't taken the spatiotemporal dynamics of HF signal denoising into account, which are critical for super-resolution tasks. Additionally, acceleration strategies relying on distillation or pruning often incur additional training time. Consequently, we propose that an optimal acceleration strategy should meet the following criteria:

1. **Cost-Effectiveness.** The strategy should achieve superior performance with fewer iterations while keeping additional training costs minimal.
2. **Generality.** The strategy should exhibit robust generalizability, enabling flexible integration across a wide range of established super-resolution frameworks.

3.2 Time-Spatial-aware Sampling

To address these issues and meet the requirements, we propose Time-Spatial-aware Sampling (TSS), a training-free content-adaptive accelerated sampling strategy (in Fig. 2). TSS integrates two core strategies: Time Dynamic Sampling and Spatial Dynamic Sampling, which are elaborated below.

Time Dynamics Sampling (TDS)

As discussed in Sec. 3.1, high-frequency signal optimization is concentrated in the early and late stages. Based on this observation, we propose the Time Dynamic Sampling (TDS) strategy with a non-uniform, adaptive timestep allocation. The non-uniform sampling design should meet two key criteria: 1) Higher sampling density in the early and late stages.

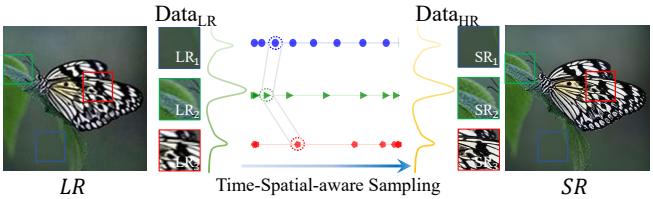


Figure 2: Overview of the proposed Time-Spatial-aware Sampling.

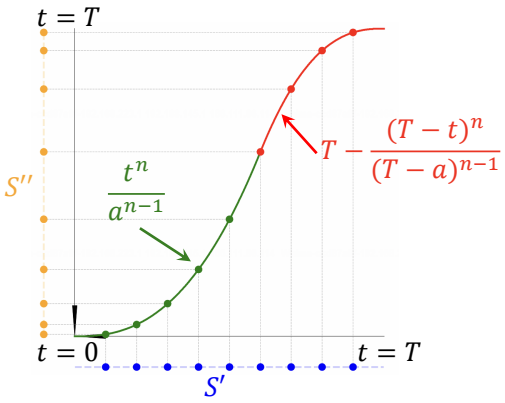


Figure 3: Illustration of the Time Dynamic Sampling strategy.

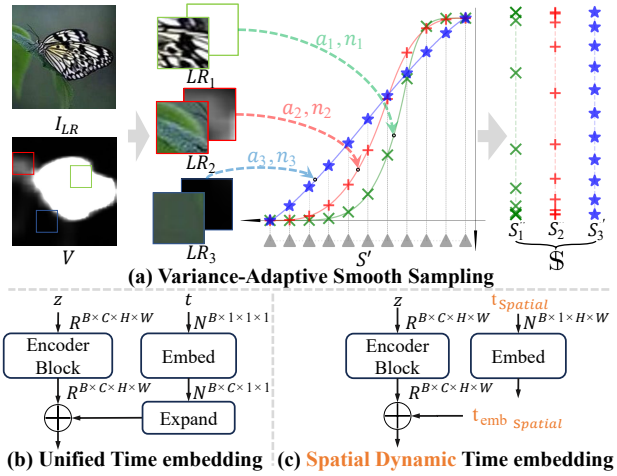


Figure 4: (a) The Variance Adaptive Smooth Sampling strategy. (b) The common unified embedding injection strategy. (c) The proposed Spatial Dynamic Time Embedding injection strategy.

2) Adjustable non-uniformity, including uniform sampling as a special case. To achieve this, we propose a non-uniform denoising approach focused on high-frequency information by implementing a non-uniform resampling strategy in the time-step schedule. Specifically, for a given number of training steps T and training scheduler $S = \{1, 2, 3, ..., T\}$, the scheduler S' for common uniform sampling of T' steps is

$$S'(T, T') = \{t_k \| t_k = \lfloor k \cdot \frac{T}{T'} \rfloor, k \in \{1, 2, ..., T'\}\}. \quad (1)$$

To incorporate adaptive non-uniformity, we resample the uniform sampling scheduler using a two-stage polynomial function to derive the modified scheduler S'', which is defined as:

$$S''(a, n, T, T') = \{f(t, a, n, T, T') \| t \in S'(T, T')\}, \quad (2)$$

$$f(t, a, n, T, T') = \begin{cases} \frac{t^n}{a^{n-1}}, t < a \\ T - \frac{(T-t)^n}{(T-a)^{n-1}}, t \geq a \end{cases}, \quad (3)$$

where a denotes the transition point between early and late stages, and n is the power factor. Fig. 3 illustrates an example of the resampling function. When $n > 1$, sampling density is concentrated at $t = 0$ and $t = T$, while n approaches 1, the function gradually converges to uniform sampling. Compared to uniform sampling, TDS enhances high-frequency signal recovery by prioritizing early and late-stage sampling, which is crucial for high-frequency restoration.

Spatial Dynamic Sampling (SDS)

TDS improves high-frequency signal restoration at the overall image level. However, as discussed in the previous section, different image regions exhibit distinct denoising dynamics, requiring region-specific sampling within a single image. To address this, we introduce Spatial Dynamic Sampling (SDS), which extends TDS with Variance-Adaptive Smooth Sampling and Spatial Dynamic Time Embedding.

Variance-Adaptive Smooth Sampling To achieve dynamic scheduling across different image regions, as illustrated in Fig. 4.a, we first compute the local variance $V_0 \in R^{H \times W}$ of the image $I_{LR} \in R^{H \times W \times 3}$ on grayscale using a 33×33 field around each pixel and post-process it with Gaussian blurring of the same size and min-max normalization for smoothness:

$$V_g = \text{norm}(\text{GaussianBlur}(V_0)). \quad (4)$$

Subsequently, we derive the non-uniform sampling strategy for each position based on the local smoothed variance V_g. Specifically, we assign an independent time scheduler to each position and adaptively adjust it based on local variations. The overall spatial-adaptive time scheduler $\mathbb{S}$ is defined as the set of schedulers for each position:

$$\mathbb{S} = \{S''(a(v_{g_{i,j}}), n(v_{v_{g_{i,j}}}), T, T') \| v_{g_{i,j}} \in V_g\}, \quad (5)$$

where $v_{g_{i,j}}$ denotes the local variance at position (i, j) which controls a and n. The spatial timestep $t_{spatial_k}$ for the k^{th} denoising iteration represents a set of timesteps for each pixel,

$$t_{spatial_k} = \{t_k \in \mathbb{S}_{i,j} | i \in [0, H], j \in [0, W]\}, \quad (6)$$

where $\mathbb{S}_{i,j}$ denotes the single scheduler in $\mathbb{S}$ for position (i, j). Considering computational efficiency and simplicity of implementation, we use linear functions for the projection $a()$ and $n()$, which are defined as follows:

$$n(v) = v(N_{\max} - N_{\min}) + N_{\min}, a(v) = v(A_{\max} - A_{\min}) + A_{\min}, \quad (7)$$

where $N_{\max}$, $N_{\min}$, $A_{\max}$, and $A_{\min}$ denote the range of n and a. As shown in Fig. 4.a, we assign different a and n values based on the local variance. In this way, we ensure that high-frequency texture regions with high variance (red block in Fig. 4.a) undergo greater sampling inhomogeneity, concentrating steps in the early and late phases to enhance high-frequency recovery. Conversely, smooth regions with lower variance (blue block in Fig. 4.a) are sampled more uniformly, facilitating the restoration of smooth areas.

Spatial Dynamics Time Embedding While Variance-Adaptive Smooth Sampling allows for location-specific timestep allocation, segmenting images into separate patches to accommodate spatially varying timesteps will increase processing time and cause boundary artifacts. To adapt pre-trained denoising networks for spatial timestep embeddings in one step, we introduce a simple yet effective spatiotemporal embedding injection strategy. Specifically, existing methods generate individual timestep embedding $t_{emb} \in R^C$ and add them to main

branch features $z \in R^{B\times C\times H\times W}$ at all spatial locations via automatic expansion (Fig. 4.b), which can be formulated as

$$t_{emb} = \mathrm{Emb}(t), z = z + \mathrm{expand}(t_{emb}), \tag{8}$$

where the expanded version $\mathrm{expand}(t_{emb})$ matches the spatial dimensions of the main branch features z. In contrast, our spatial timestep embedding is the set of embeddings for each spatial location in $t_{spatial}$, which has the same spatial dimensions as the main branch features z, as depicted in Fig. 4.c Thus, it can be directly added to the main branch features at all spatial locations, which can be expressed as:

$$t_{emb_{spatial}} = \{\mathrm{Emb}(t_{i,j})|t_{i,j} \in t_{spatial}\}, z = z+t_{emb_{spatial}}. \tag{9}$$

This allows the network to adapt to spatially varying timestep embeddings while integrating flexibly into existing architectures without extra cropping or multiple forward passes.

4 Experiment

4.1 Experiment Setup

Implementation Details Our approach modifies only the sampling strategy during inference, requiring no extra training. Experiments were conducted on an NVIDIA A800 GPU (80GB) using the official codebases and pre-trained weights. The PyTorch framework was used for implementation, and the detailed hyperparameters are provided in the full version.

Testing Datasets Performance was evaluated on both synthetic and real-world data. Synthetic data used DIV2K [2017] with LR images degraded by BSRGAN [2021]. Real-world evaluations used DRealSR [2020], RealPhoto60 [2024], and RealSR [2020], along with face SR benchmarks WebPhoto-Test [2021a] and LFW-test [2019]. RealPhoto60, WebPhoto, and LFW were tested at ×2 scale, and others at ×4.

Evaluation Metrics For the quantitative evaluation of super-resolution, we employed widely used perceptual quality metrics, including NIQE [2013], CLIPIQA [2023], MUSIQ [2021], and QAlign [2023], to compare the performance of the competing methods. Additionally, we provide the full-reference metrics including PSNR [2010], SSIM [2004], and LPIPS [2018] in the full version.

4.2 Comparison with State-of-the-Art Methods

For a comprehensive comparison, we integrate the TSS framework with three recent SOTA diffusion-based super-resolution (SR) methods: StableSR [2024a], SUPIR [2024], and PASD [2025]. Their performance is evaluated against state-of-the-art (SOTA) SR approaches, including Real-ESRGAN [2021b], BSRGAN [2021], SwinIR [2021], SinSR [2024b], and ResShift [2024], using both synthetic and real-world datasets. Real-ESRGAN, BSRGAN, and SwinIR are traditional deep learning-based SR frameworks using CNNs or transformers, while SinSR and ResShift are diffusion-accelerated SR methods leveraging distillation and fine-tuning, respectively. For fairness, the results are obtained from official codebases and pre-trained models. Quantitative evaluation and qualitative comparisons are presented in Tab. 1 and Fig. 5, respectively. As presented in Tab. 1, our SUPIR^{TSS} outperforms the state-of-the-art acceleration strategies ReShift and SinSR on most benchmarks, achieving a 1.14 - 5.05 improvement in NIQE [2013] and a 0.09 - 1.52

Datasets	Methods	NIQE↓	CLIPIQA↑	MUSIQ↑	Qalign↑
DIV2K SR (x4)	Real-ESRGAN	4.87	0.5963	56.55	3.53
	BSRGAN	3.78	0.5804	60.25	3.68
	SwinIR	3.51	0.5677	58.27	3.92
	ResShift (N=15)	7.68	0.5963	44.32	3.25
	SinSR (N=1)	6.53	0.6745	55.28	3.66
	StableSR(N=100)	7.94	0.3429	27.69	2.49
	StableSR^{TSS}(N=100)	**7.16**	**0.3604**	**28.06**	**2.50**
	SUPIR(N=7)	5.02	0.3758	60.76	4.11
	SUPIR^{TSS}(N=7)	**3.56**	**0.5244**	**62.35**	**4.28**
	PASD(N=7)	7.78	0.4071	39.79	3.03
	PASD^{TSS}(N=7)	**6.11**	**0.4192**	**41.82**	**3.15**
RealSR SR (x4)	Real-ESRGAN	4.70	0.4818	59.50	3.92
	BSRGAN	4.66	0.5399	63.37	3.86
	SwinIR	4.69	0.4636	59.40	3.85
	ResShift (N=15)	7.43	0.5427	53.52	3.84
	SinSR (N=1)	6.24	0.6631	59.23	3.87
	StableSR(N=100)	5.06	0.5530	60.99	3.90
	StableSR^{TSS}(N=100)	**4.98**	0.5691	**61.15**	**3.93**
	SUPIR(N=7)	6.44	0.4435	57.85	3.66
	SUPIR^{TSS}(N=7)	**4.76**	**0.5017**	**58.74**	**3.96**
	PASD(N=7)	4.99	0.5341	60.34	4.01
	PASD^{TSS}(N=7)	**4.31**	**0.5739**	**62.00**	**4.08**
DRealSR SR (x4)	Real-ESRGAN	4.35	0.5769	56.88	4.34
	BSRGAN	4.60	0.6125	58.55	4.31
	SwinIR	4.39	0.5668	56.93	4.33
	ResShift (N=15)	6.03	0.6490	56.23	4.30
	SinSR (N=1)	5.51	0.7134	56.72	4.30
	StableSR(N=100)	4.38	0.6472	55.88	4.26
	StableSR^{TSS}(N=100)	4.73	**0.6523**	**56.12**	**4.27**
	SUPIR(N=7)	5.79	0.4760	52.69	4.09
	SUPIR^{TSS}(N=7)	**4.37**	**0.5362**	**54.26**	**4.27**
	PASD(N=7)	4.58	0.6612	58.21	4.30
	PASD^{TSS}(N=7)	**3.95**	**0.6923**	57.58	**4.32**
RealPhoto60 SR (x2)	Real-ESRGAN	3.92	0.5709	59.25	3.64
	BSRGAN	5.38	0.3305	45.46	2.11
	ResShift (N=15)	6.59	0.6642	61.29	3.77
	SinSR (N=1)	5.91	0.7610	66.43	3.90
	StableSR(N=100)	4.39	0.5484	55.41	3.65
	StableSR^{TSS}(N=100)	**4.28**	**0.5575**	**56.40**	**3.68**
	SUPIR(N=7)	5.80	0.4597	65.42	3.62
	SUPIR^{TSS}(N=7)	**3.86**	**0.6277**	**67.86**	**4.38**
	PASD(N=7)	4.60	0.6244	63.85	3.99
	PASD^{TSS}(N=7)	**3.86**	**0.6427**	**66.38**	**4.22**
WebPhoto SR (x2)	Real-ESRGAN	5.92	0.4937	37.89	1.94
	BSRGAN	7.16	0.4316	38.86	1.78
	ResShift (N=15)	10.24	0.4590	29.63	1.84
	SinSR (N=1)	7.94	0.6610	51.10	2.36
	StableSR(N=100)	7.54	0.3412	28.03	1.72
	StableSR^{TSS}(N=100)	**7.00**	**0.3602**	**28.78**	**1.73**
	SUPIR(N=7)	8.56	0.3935	60.42	2.95
	SUPIR^{TSS}(N=7)	**5.19**	**0.4815**	58.72	**3.36**
	PASD(N=7)	11.18	0.3172	28.73	1.85
	PASD^{TSS}(N=7)	**8.57**	**0.3485**	**31.25**	**2.06**
LFW SR (x2)	Real-ESRGAN	5.06	0.5337	56.60	2.75
	BSRGAN	6.14	0.5566	58.48	2.60
	ResShift (N=15)	8.42	0.5570	52.48	2.64
	SinSR (N=1)	6.97	0.7553	65.40	2.98
	StableSR(N=100)	5.67	0.5246	55.51	3.18
	StableSR^{TSS}(N=100)	**5.42**	**0.5416**	**56.50**	**3.22**
	SUPIR(N=7)	6.21	0.4329	67.09	3.50
	SUPIR^{TSS}(N=7)	**4.36**	**0.5776**	**67.23**	**4.22**
	PASD(N=7)	6.51	0.5383	59.23	3.15
	PASD^{TSS}(N=7)	**5.10**	**0.5567**	**62.25**	**3.55**

Table 1: Quantitative comparison with SOTA real-world SR methods. '**Bold**' indicates improvement over the baseline, while 'Underline' denotes the best performance. SwinIR was evaluated only on x4 scale factors due to the lack of pretrained weights for x2.

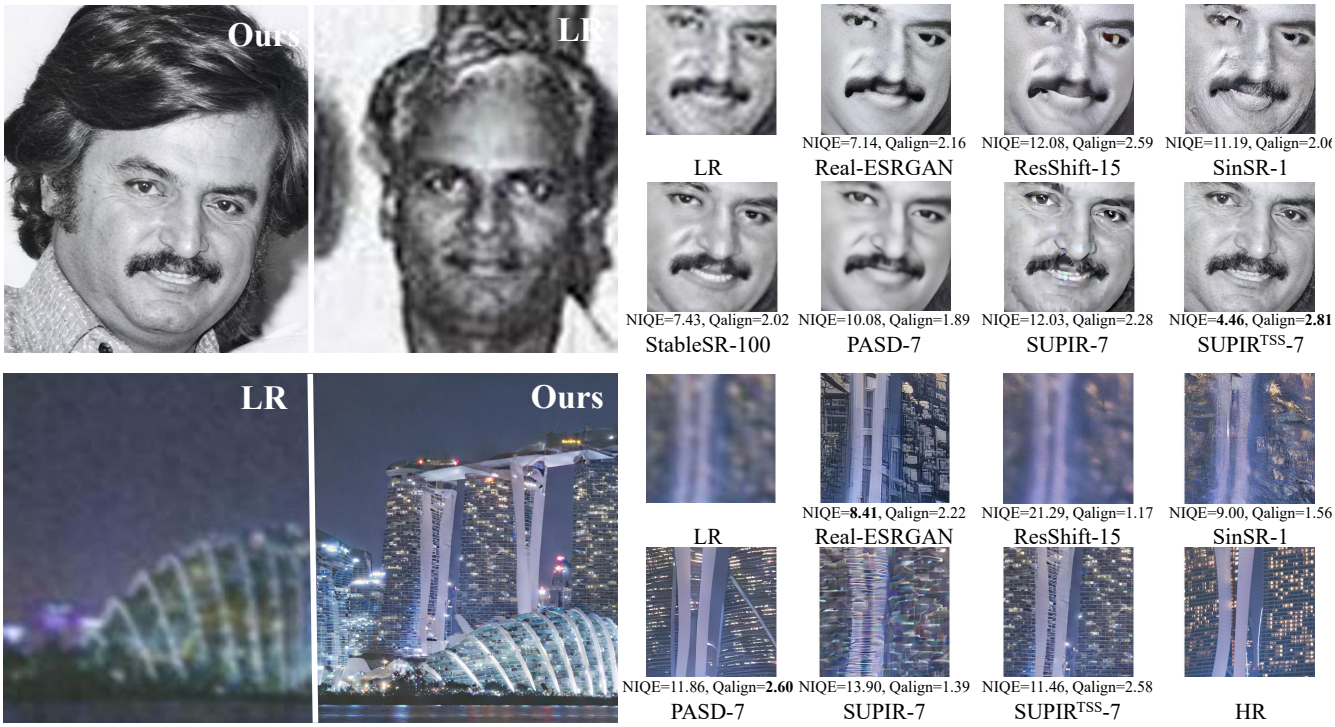


Figure 5: Qualitative comparison with state-of-the-art methods. Top: real-world sample from RealPhoto60 datasets. Bottom: synthetic sample from DIV2K valid datasets. More visual results are provided in the full version.

improvement in QAlign [2023] across both synthetic and real datasets. Notably, while baseline SUPIR's 7-step QAlign results are initially inferior to ReShift and SinSR on real datasets, the incorporation of the TSS framework enables a QAlign improvement of 0.09 - 1.24 over SinSR and 0.12 - 1.58 over ReShift across all datasets. Although SinSR requires only a single iteration, its distillation strategy involves additional training costs, whereas TSS achieves comparable results without any extra training. Furthermore, compared to ReShift, which requires training and 15 inference steps, SUPIR^{TSS} delivers superior performance across all metrics except CLIPIQA [2023], achieving these results in just 7 iterations, less than half the steps. Fig. 5 provides examples of synthetic and real degradation. Taking the first example in Fig. 5 as an example, most existing methods, including CNN- and Transformer-based approaches and diffusion acceleration strategies, fail to generate the region with high-frequency textures (wrinkles, beard, and eyebrow). In contrast, SUPIR^{TSS} produces more realistic and richer high-frequency textures, as evidenced by both qualitative visual quality and quantitative results (Fig. 5, Row. 2, Col. 4) In summary, both quantitative and qualitative analyses show that TSS significantly enhances the high-frequency restoration capabilities of recent diffusion-based SR methods, achieving state-of-the-art texture generation performance on both synthetic and real-world datasets without any additional training costs.

4.3 Ablation Study

Effectiveness of Time Dynamic Sampling To validate the effectiveness of Time Dynamic Sampling (TDS), we compare

Strategy	TDS / SDS	- / -	✓ / -	✓ / ✓
Step=7	NIQE ↓	5.80	3.99	**3.86**
	CLIPIQA ↑	0.4597	0.6212	**0.6277**
	MUSIQ ↑	65.42	67.53	**67.86**
	Qalign ↑	3.62	4.35	**4.38**
Step=10	NIQE ↓	5.23	3.89	**3.65**
	CLIPIQA ↑	0.5000	0.6391	**0.6571**
	MUSIQ ↑	65.93	67.42	**68.50**
	Qalign ↑	3.93	4.33	**4.38**
Step=14	NIQE ↓	4.41	3.59	**3.65**
	CLIPIQA ↑	0.5912	0.6438	**0.6439**
	MUSIQ ↑	68.24	68.51	**68.61**
	Qalign ↑	4.26	4.31	**4.33**

Table 2: Comparison of SUPIR with different sampling strategies.

SUPIR with its TDS-enhanced version on the RealPhoto60 dataset. As shown in Tab. 2, TDS improves quantitative metrics, achieving a 20% increase in QAlign[2023] and a 35% increase in CLIPIQA [2023] at 7 steps. Fig. 6 (Col. 3 vs. 1) visually compares SUPIR results with and without TDS on real samples. TDS enhances the butterfly texture reconstruction while uniform sampling introduces artifacts like color blocking. For further analysis, Fig. 7 illustrates the relationship between high-frequency SNR and the number of denoising steps. Compared to the original SUPIR (blue), SUPIR+TDS (green) consistently achieves higher SNR at the same step count, even with fewer total steps, indicating its effectiveness in high-frequency signal restoration.

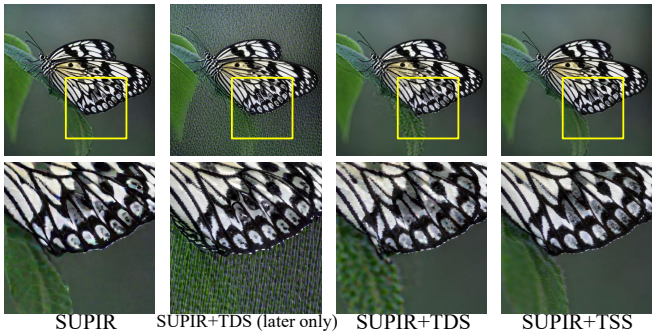


Figure 6: Visual comparison of SUPIR-7 with different sampling strategies on real samples in RealPhoto60.

Metrics	Uniform	Later Stage Only	TDS
NIQE ↓	5.80	5.80 (-0.00)	**3.99** (-1.81)
CLIPIQA ↑	0.4597	0.6032 (+0.1435)	**0.6212** (+0.1615)
MUSIQ ↑	65.42	66.58 (+1.16)	**67.53** (+2.11)
Qalign ↑	3.62	3.32 (-0.30)	**4.35** (+0.73)

Table 3: Ablation study on the necessity of early-stage sampling. Red for improvement, and blue for degradation.

Necessity of Early Stage Sampling To verify the need for early-stage sampling, we compare SUPIR with late-stage-only sampling ($a = T$ in Eq. 3) and Time Dynamic Sampling (TDS) on the RealPhoto60. As shown in Tab. 3, late-stage-only sampling fails to improve NIQE and results in a 0.3 drop in QAlign, while TDS nearly doubles the MUSIQ gain by balancing early and late sampling. Fig. 7 compares the high-frequency SNR of both methods. Due to insufficient early-stage denoising, late-stage-only sampling retains excessive HF noise at lower steps, sometimes underperforming the baseline. In contrast, TDS, with concentrated sampling in both early and late stages, effectively boosts HF SNR with fewer steps. Fig. 6 (Col. 2 vs. 3) further illustrates this effect. While late-stage-only sampling enhances local texture, fewer steps in the early stage leave residual noise, causing visible artifacts. TDS, by ensuring adequate sampling across both stages, enables realistic texture generation without artifacts.

Non-Uniform Function Selection We evaluated the impact of different non-uniform functions on sampling performance, including trigonometric, exponential, and polynomial functions. As shown in Tab. 4, all three functions outperform the baseline with a 0.6 - 0.75 QAlign improvement, showing that non-uniform sampling aligns with high-frequency signal recovery regardless of the function type. For better controllability, we select the polynomial function as the final setting.

Metrics	Uniform	Non-Uniform		
		Trigonometric	Exponential	Polynomial
NIQE ↓	5.80	3.96	4.36	**3.99**
CLIPIQA ↑	0.4597	**0.6224**	0.6033	0.6212
MUSIQ ↑	65.42	67.42	66.11	**67.53**
Qalign ↑	3.62	4.33	4.25	**4.35**

Table 4: Comparison of non-uniform functions used in TDS. The detailed function description is in the full version.

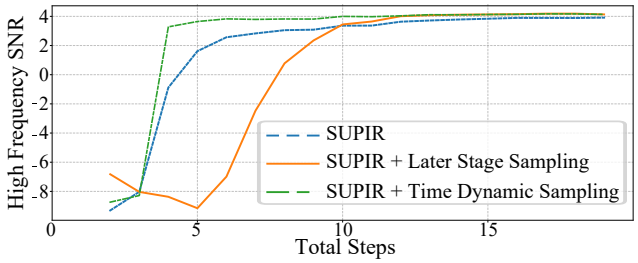


Figure 7: Comparison of SNR for high-frequency components at different steps counts when using different sampling strategies.

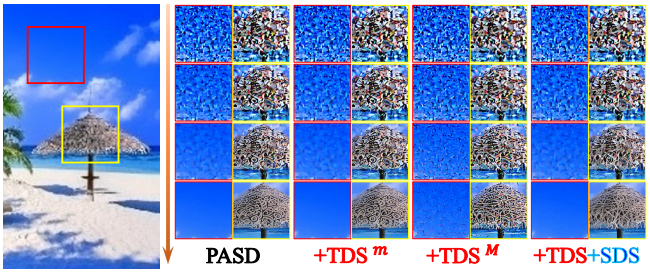


Figure 8: Last 4 steps in PASD's 7-step denoising process. "+TDS" strategies fixes param a and n to their **m**in and **M**ax value.

Effectiveness of Spatial Dynamic Sampling To assess the necessity of Spatial Dynamic Sampling (SDS), we compare SUPIR with TDS alone and the full TSS (TDS+SDS) on RealPhoto60. As shown in Tab. 2 (Col. 2 and 3), the inclusion of SDS brings quantitative metric improvements over TDS alone. Moreover, Fig. 6 further illustrates this effect. TDS alone (Col. 3) applies a spatial-unified scheduler, resulting in less sharp butterfly wings and an insufficiently smooth background. In contrast, SDS enables spatial-wise adaptive sampling, enhancing both texture-rich and smooth regions.

SDS in the Denoising Process For a more intuitive explanation, we recorded the denoising process of PASD with different sampling strategies in Fig. 8. Excessive inhomogeneity (Col. 3) adds noise to smooth regions, while insufficient inhomogeneity (Cols. 1, 2) fails to generate textures. In contrast, SDS (Col. 4) adaptively balances inhomogeneity, ensuring optimal reconstruction of both smooth and textured areas.

5 Conclusion

In this work, we explore key insights in the denoising process: high-frequency components require focused optimization in early and late iterations, while spatially varying content necessitates adaptive strategies for effective restoration. Based on these findings, we propose Time-Spatial-aware Sampling (TSS), a training-free, content-adaptive sampling strategy to accelerate diffusion-based image super-resolution. By leveraging temporal and spatial dependencies in high-frequency recovery, TSS enhances texture restoration while significantly reducing computational costs. Compatible with various diffusion SR frameworks, it achieves state-of-the-art results with half the steps of recent accelerated methods.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62072271.

Contribution Statement

† denotes authors who contributed equally as co-first authors.

References

[Agustsson and Timofte, 2017] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[Bao *et al.*, 2025] Jingwei Bao, Jinhua Hao, Pengcheng Xu, Ming Sun, Chao Zhou, and Shuyuan Zhu. Plug-and-play tri-branch invertible block for image rescaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1826–1834, 2025.

[Chen *et al.*, 2024] Haolan Chen, Jinhua Hao, Kai Zhao, Kun Yuan, Ming Sun, Chao Zhou, and Wei Hu. Cassr: Activating image power for real-world image super-resolution. *arXiv preprint arXiv:2403.11451*, 2024.

[Choi *et al.*, 2021] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

[Cochran *et al.*, 1967] W.T. Cochran, J.W. Cooley, D.L. Favin, H.D. Helms, R.A. Kaenel, W.W. Lang, G.C. Maling, D.E. Nelson, C.M. Rader, and P.D. Welch. What is the fast fourier transform? *Proceedings of the IEEE*, pages 1664–1674, 1967.

[He *et al.*, 2024] Xiao He, Huaao Tang, Zhijun Tu, Junchao Zhang, Kun Cheng, Hanting Chen, Yong Guo, Mingrui Zhu, Nannan Wang, Xinbo Gao, et al. One step diffusion-based super-resolution with time-aware distillation. *arXiv preprint arXiv:2408.07476*, 2024.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, pages 6840–6851, 2020.

[Horé and Ziou, 2010] Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369, 2010.

[Ji *et al.*, 2020] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 466–467, 2020.

[Kawar *et al.*, 2022] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, pages 23593–23606, 2022.

[Ke *et al.*, 2021] 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*, pages 5148–5157, 2021.

[Liang *et al.*, 2021] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[Liu *et al.*, 2023] Guandu Liu, Yukang Ding, Mading Li, Ming Sun, Xing Wen, and Bin Wang. Reconstructed convolution module based look-up tables for efficient image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12217–12226, 2023.

[Ma *et al.*, 2024] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15762–15772, 2024.

[Mittal *et al.*, 2013] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, pages 209–212, 2013.

[Podell *et al.*, 2023] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[Qin *et al.*, 2023] Rui Qin, Ming Sun, Fangyuan Zhang, Xing Wen, and Bin Wang. Blind image super-resolution with rich texture-aware codebook. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 676–687, 2023.

[Qin *et al.*, 2024] Rui Qin, Ming Sun, Chao Zhou, and Bin Wang. A new dataset and framework for real-world blurred images super-resolution. In *European Conference on Computer Vision*, pages 56–75, 2024.

[Qu *et al.*, 2025] Yunpeng Qu, Kun Yuan, Kai Zhao, Qizhi Xie, Jinhua Hao, Ming Sun, and Chao Zhou. Xpsr: Cross-modal priors for diffusion-based image super-resolution. In *Computer Vision – ECCV 2024*, pages 285–303, 2025.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Saharia *et al.*, 2022] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, pages 4713–4726, 2022.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Srivastava *et al.*, 2019] Y Srivastava, V Murali, and SR Dubey. A performance evaluation of loss functions for deep face recognition, computer vision. In *Pattern Recognition, Image Processing, and Graphics: 7th National Conference*, pages 22–24, 2019.

[Wang *et al.*, 2004] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, pages 600–612, 2004.

[Wang *et al.*, 2021a] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9168–9178, 2021.

[Wang *et al.*, 2021b] 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*, pages 1905–1914, 2021.

[Wang *et al.*, 2023] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023.

[Wang *et al.*, 2024a] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024.

[Wang *et al.*, 2024b] 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*, pages 25796–25805, 2024.

[Wei *et al.*, 2020] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 101–117, 2020.

[Wu *et al.*, 2023] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

[Wu *et al.*, 2024a] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *arXiv preprint arXiv:2406.08177*, 2024.

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

[Xie *et al.*, 2025] Rui Xie, Yinhong Liu, Penghao Zhou, Chen Zhao, Jun Zhou, Kai Zhang, Zhenyu Zhang, Jian Yang, Zhenheng Yang, and Ying Tai. Star: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. *arXiv preprint arXiv:2501.02976*, 2025.

[Yang *et al.*, 2025] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *European Conference on Computer Vision*, pages 74–91, 2025.

[Yin *et al.*, 2024] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6613–6623, 2024.

[Yu *et al.*, 2024] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25669–25680, 2024.

[Yue *et al.*, 2024] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 2024.

[Zhang *et al.*, 2018] 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*, pages 586–595, 2018.

[Zhang *et al.*, 2021] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[Zhao *et al.*, 2023] Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. Quality-aware pre-trained models for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22302–22313, 2023.

[Zhu *et al.*, 2024] Haowei Zhu, Dehua Tang, Ji Liu, Mingjie Lu, Jintu Zheng, Jinzhang Peng, Dong Li, Yu Wang, Fan Jiang, Lu Tian, et al. Dip-go: A diffusion pruner via few-step gradient optimization. *arXiv preprint arXiv:2410.16942*, 2024.