DiffSSR+: Optimized Progressive Diffusion With Stereo-Consistent Sampling for Stereo Super-Resolution

Jinyi Luo, *Student Member, IEEE*, Zejia Fan, Wenhan Yang, *Member, IEEE*, and Jiaying Liu, *Fellow, IEEE*

***Abstract*—Recent single-image super-resolution methods leverage the generative priors of diffusion models to achieve superior perceptual quality in the reconstructed images. However, these perceptual-driven methods suffer from severe fidelity distortion in stereo super-resolution tasks due to the stereo inconsistency introduced by the stochastic nature of the diffusion generation process. We address this challenge by leveraging stereo constraints to guide diffusion at both noise sampling and diffusion feature fusion, allowing the model to handle complex low-resolution degradation while maintaining cross-view consistency, achieving a favorable perceptual-fidelity balance. To be specific, we address the fundamental problem that the inherent stochasticity of diffusion disrupts stereo consistency by proposing a stereo-consistent noise sampling optimization strategy, where we obtain the left-view noise by sampling an approximate optimal initial noise and optimizing the variance noise via a perceptual-consistency loss, while the right-view noise is obtained by disparity-aware warping. This approach drives the diffusion process toward outputs that satisfy stereo high-frequency detail consistency while delivering improved perceptual quality. To enhance stereo consistency at the structural and semantic levels, we design a Disparity-Aware State-Space Module, which performs stereo feature fusion via a sequentially bi-directional cross-scanning approach that is more suitable for the diffusion feature space than traditional attention operation. Together, we build a diffusion-based stereo super-resolution framework, DiffSSR+, which progressively integrates stereo guidance into the diffusion process to achieve a better fidelity-perception trade-off, and is plug-and-play to a broad range of diffusion SR architectures.**

***Index Terms*—Diffusion model, disparity warping, feature fusion, noise optimization, plug-and-play, stereo super-resolution.**

Received 16 February 2026; revised 12 June 2026; accepted 9 July 2026. Date of publication 2 September 2026; date of current version 5 November 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62332010, in part by the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology), in part by the Phase I Open Research Projects, Yungang Academy under Grant 2025YGKFKT004, and in part by the Interdisciplinary Frontier Research Project of PCL under Grant PCL2025QYB013. The Associate Editor coordinating the review of this article and approving it for publication was Dr Jong-Seok Lee. *(Corresponding author: Jiaying Liu.)*

Jinyi Luo, Zejia Fan, and Jiaying Liu are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100080, China (e-mail: luojinyi@stu.pku.edu.cn; zejia@pku.edu.cn; liujiaying@pku.edu.cn).

Wenhan Yang is with Pengcheng Laboratory, Shenzhen, Guangdong 518066, China (e-mail: yangwh@pcl.ac.cn).

Our project is publicly available at https://pku-jamesluo.github.io/Project_DiffSSR/.

This article has supplementary downloadable material available at https://doi.org/10.1109/TMM.2026.3730265, provided by the authors.

Digital Object Identifier 10.1109/TMM.2026.3730265

I. Introduction

STEREO image super-resolution (Stereo SR) aims to recover high-resolution (HR) stereo image pairs from low-resolution (LR) inputs while preserving cross-view correspondence. This task is crucial for 3D-aware visual applications such as augmented reality, autonomous systems, and robotic perception. However, inconsistent reconstruction of fine-grained details across views can introduce stereo mismatches, which in turn severely degrade both visual realism and 3D application performance. Therefore, achieving both high perceptual quality and stereo consistency remains a central challenge in stereo image super-resolution.

Current methods in stereo image super-resolution are typically built upon the framework of single-image super-resolution (SISR), whose methods can be generally categorized into fidelity-driven and perception-driven approaches. Fidelity-driven methods [1], [2] typically focus on minimizing pixel-level discrepancies with respect to the ground truth. However, this optimization often results in over-smooth outputs, which degrades perceptual visual quality. In contrast, perception-driven methods [3], [4] focus on generating SR results that are better aligned to the perceptual quality of the ground truth. Nevertheless, relaxing fidelity constraints in these approaches frequently results in the synthesis of false textures and increased instability in modeling image manifolds.

In practice, most existing stereo image super-resolution methods still follow the fidelity-driven reconstruction paradigm. Representative approaches based on cross-view attention [5], [6], [7] achieve promising stereo-consistent SR results through stereo feature interaction. However, their fully deterministic inference pipelines cannot handle real world scenarios where complex degradation patterns occur, often leading to noticeably reduced perceptual quality and even unacceptable outputs under severe degradations. On the other hand, although perceptual-driven methods based on generative models [8], [9], [10] provide improved perceptual quality and strong robustness to diverse degradation types in real-world single-image SR by exploiting the generative prior from the model, when directly applied to stereo SR scenario, these methods introduce severe stereo inconsistency, as the two views frequently diverge in the generated high-frequency details. This issue is caused by the inherent stochasticity of generative models, which leads to substantial degradation in stereo fidelity. This trade-off between stereo

1520-9210 © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Peking University. Downloaded on November 12,2026 at 03:00:38 UTC from IEEE Xplore. Restrictions apply.

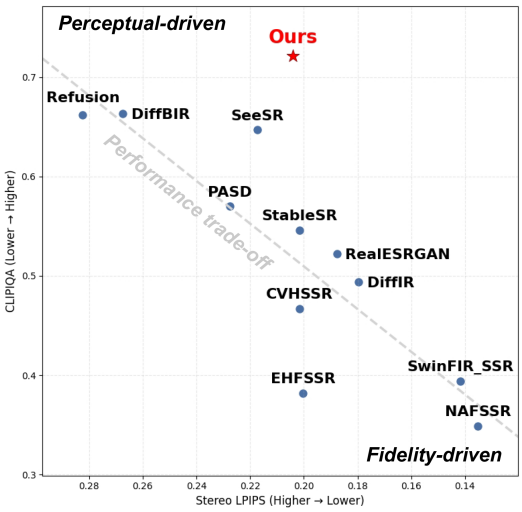


Fig. 1. Existing methods exhibit a performance trade-off between perceptual quality and stereo fidelity. In contrast, our method (marked in red) achieves superior perceptual performance while maintaining stereo fidelity comparable to fidelity-driven methods.

fidelity and perceptual quality forms a performance bottleneck for most existing SR methods, as shown in Fig. 1.

In this paper, we aim to propose a diffusion-based stereo super-resolution framework designed to preserve the strong perceptual advantages of generative models while simultaneously enhancing stereo consistency. We incorporate stereo constraints into both noise space and diffusion feature space, achieving a better fidelity-perception trade-off. To be specific, for noise enhancement, we propose a stereo-consistent noise sampling and optimization strategy that tackles high-frequency inconsistencies from diffusion stochasticity. We adopt an optimal initial noise under the Optimal Bound Conditions (OBC) [11] for a single view, then warp it to the other view using disparity information. The stochastic variance noise is progressively optimized at test-time via a combined perceptual and stereo-consistency loss. This strategy effectively improves high-frequency detail generation while preserving stereo consistency. For feature fusion enhancement, we propose a Disparity-Aware State-Space Model (DASSM). We apply disparity fusion to combine the diffusion features from the two views, and utilize a Stereo Bi-directional Cross-Scan Model (SBCSM), where the combined stereo features are scanned in a horizontal-vertical scheme. This fusion captures structural and semantic correlations between stereo features, enabling complementary enhancement on stereo consistency. Our proposed framework is plug-and-play, enabling robust stereo extensions for a wide range of existing and future diffusion-based SR methods.

Our contributions can be summarized as follows:

- We propose a plug-and-play **Diff**usion-based **S**tereo image **S**uper-**R**esolution framework, termed **DiffSSR+**, which integrates stereo-consistent constraints into both noise sampling and feature spaces, achieving robustness to complex degradations and strong cross-view consistency. Extensive experiments demonstrate that our framework is both effective and broadly applicable to diffusion-based SR models.
- We propose a stereo-consistent noise sampling and optimization strategy for diffusion-based stereo restoration. We sample an optimal initial noise and optimize the variance noise for the left view, while deriving the noise for the right view via disparity-aware warping. The approach enhances single-view perceptual quality while suppressing stereo inconsistency caused by diffusion stochasticity.
- We introduce a Disparity-Aware State-Space Model (DASSM) for effective stereo feature interaction and fusion, including a Stereo Bi-directional Cross-Scan Model (SBCSM) design for linear sequence feature scanning, achieving improved stereo-consistent reconstruction at structural and semantic levels.

This paper is built based on our previous work [12], where we have made several substantial contributions to method improvement and comprehensive content addition as follows: (1) We propose a **stereo-consistent noise optimization** method. We sample an approximate optimal initial noise and optimize the variance noise in the stereo diffusion process via a combined perceptual-consistency loss to enhance stereo-consistent high-frequency detail reconstruction. The details are illustrated in Section III-E. (2) We introduce a **Spatial Bi-directional Cross-Scan Model (SBCSM)** for stereo feature fusion. The model performs stereo feature fusion via vertical-horizontal bi-directional scanning, while refine the output sequencing to reduce the learning difficulty of spatial reordering, leading to further improvements in structural and semantic reconstruction performance. The relative content is presented in Section III-C. (3) We extend the original Diff-SSR into a **plug-and-play framework** applicable to a series of diffusion-based SR models. Compared to our previous version, we achieve notable performance gains of 16.3% in LPIPS, 19.2% in DISTS, and 14.0% in MANIQA on the Flickr1024RS dataset. We further validate the general applicability of our framework across multiple state-of-the-art diffusion-based SR backbones, achieving LPIPS improvements of 3.7% on SeeSR [10] and 17.7% on PASD [13]. Our approach offers a convenient and effective stereo extension for potential future SOTA SISR methods. The relative results are provided in Section IV-B.

The rest of our paper is organized as follows: In Section II, we review the related works on real-world image super-resolution, diffusion-based super resolution and stereo image super resolution. In Section III, we introduce the proposed diffusion-based stereo super-resolution method. We then provide experiment results and analysis in Section IV. Finally, the conclusion remarks are given in Section V.

II. Related Work

A. Real-World Image Super-Resolution

Real-world image super-resolution (Real-ISR) aims to recover images with high signal fidelity and perceptual clarity from inputs affected by low resolution and multiple complex degradations. vMany early deep learning–based methods for image restoration and super-resolution leverage convolutional neural networks (CNNs) [2], [14], [15], [16] and Transformers [17], [18], [19] to model correlations in low-resolution pixels

Authorized licensed use limited to: Peking University. Downloaded on November 12,2026 at 03:00:38 UTC from IEEE Xplore. Restrictions apply.

for high-resolution reconstruction and general restoration tasks. While these approaches improve spatial resolution with high fidelity to the original content, their capacity to reconstruct fine details is limited when information loss is severe. Moreover, they frequently encounter difficulties in dealing with complex degradations, such as ringing artifacts or extra noise, which are commonly present in real-world low-resolution images.

More recent Real-ISR methods focus on modeling complex degradation processes to better reflect real-world scenarios, while leveraging generative models to recover perceptually realistic details. RealSR [20] estimates blur and noise distributions from LR images with a degradation framework to model realistic degradations. In contrast, BSR-GAN [21] and Real-ESRGAN [3] employ randomized and complex degradation synthesis pipelines and adopt Generative Adversarial Networks (GANs) to better learn from real-world degradations. However, despite their relatively comprehensive degradation modeling, these methods remain limited by their generative capacity and struggle to reconstruct perceptually plausible details.

B. Diffusion-Based Image Super-Resolution

Recent perceptual-oriented image super-resolution methods [22], [23], [24] exploit the strong generative prior of diffusion models to enhance perceptual reconstruction quality. These diffusion-based super-resolution methods can generally be classified into three main groups, according to how LR information is integrated into the diffusion process.

The first group comprises condition-driven methods that directly inject LR information as diffusion conditions, such as noise initialization or feature aggregation. Representative approaches include ResShift [25], which constructs a diffusion Markov chain between LR and HR latent spaces; StableSR [9], which employs a time-aware encoder to fuse LR latents into denoising layers; and DiffBIR [8], which adopts a two-stage pipeline for progressive fidelity reconstruction and perceptual enhancement. The second group couples the SR process with diffusion mechanisms. For instance, UPSR [26] reweights diffusion noise using an uncertainty map estimated from the LR image, while InvSR [27] reformulates super-resolution via diffusion inversion.

Although these two groups improve the ability of diffusion models to synthesize fine-grained details, they often exhibit limited semantic fidelity. To address this issue, the third group incorporates semantic prompts to guide reconstruction. SeeSR [10] introduces a degradation-aware prompt extractor with hard and soft text embeddings, PASD [13] provides semantic guidance through multiple high-level networks with pixel-aware cross-attention, and DiT4SR [28] leverages a Diffusion Transformer to integrate semantic information.

Despite their success in SISR task, these diffusion-based methods are not well suited for stereo SR. The stochastic nature of diffusion often leads to texture inconsistencies across stereo views, resulting in perceptually implausible distortions.

C. Stereo-Image Super-Resolution

The initial study of stereo-image super-resolution focuses on incorporating the stereo matching [29] and attention-based solutions [30], [31]. Disparity maps derived from parallax attention maps help tackle occlusions, enhancing stereo consistency [32]. The iPASSRnet [33] leverages symmetry cues and a dual parallax attention mechanism (biPAM) in a siamese network structure to enhance both images in a stereo pair.

Several state-of-the-art methods utilize cross-view attention models to further facilitate stereo information interaction. NAFSSR [5] integrates cross-view attention module (SCAM) between successive NAFblocks [34]. Following works including Steformer [35], CVHSSR [7], EHFSSR [36], and CVCNet [37], explore cross-/intra-view or hybrid feature interaction via attention-based architectures, while StereoINR [38] leverages implicit neural representations to model non-local cross-view correspondence. These existing works mainly focus on signal fidelity and stereo consistency, while paying limited attention to high-quality perceptual reconstruction. In this work, we aim to improve both perceptual quality and stereo consistency in stereo image super-resolution.

III. Methodology

A. Preliminaries

1) Diffusion Models: DDPM [39] serves as a foundation of modern diffusion-based generation models. A typical DDPM forward process gradually applies Gaussian noise for T steps on the high-quality image to move from a natural high-quality image domain to Gaussian noise distribution, and the backward process reconstructs the high-quality image by progressively removing noise from the initial noisy image.

Practically, let β_t denote the pre-determined variance schedule, $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.Given a high quality image $\mathbf{x}_0 \sim q(x_0)$, the noisy image at timestep t is obtained by:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

while during the inference stage, the denoised image at the previous timestep is obtained from the current timestep t as:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where $\boldsymbol{\epsilon}_\theta$ denotes the trained noise prediction network parameterized by θ.

2) Selective State-space Model: State-space models (SSMs) provide a principled framework for modeling dynamic systems by capturing how internal states evolve over time in response to external inputs. An SSM represents the system behavior through a latent state that encodes historical information and enables the prediction of future system responses.

Formally, given an input signal $x(t) \in \mathbb{R}^L$, the system produces the corresponding output responses $y(t) \in \mathbb{R}^L$ by an intermediate hidden state $h(t) \in \mathbb{C}^N$, whose temporal evolution is controlled by the time derivative $\dot{h}(t) := \frac{d}{dt}h(t)$. The hidden state is maintained and continuously updated by the model to capture long-term dependencies in the sequence.

Specifically, the state evolution and output generation are defined as:

$$\dot{h}(t) = Ah(t) + Bx(t), \tag{3}$$

Authorized licensed use limited to: Peking University. Downloaded on November 12,2026 at 03:00:38 UTC from IEEE Xplore. Restrictions apply.

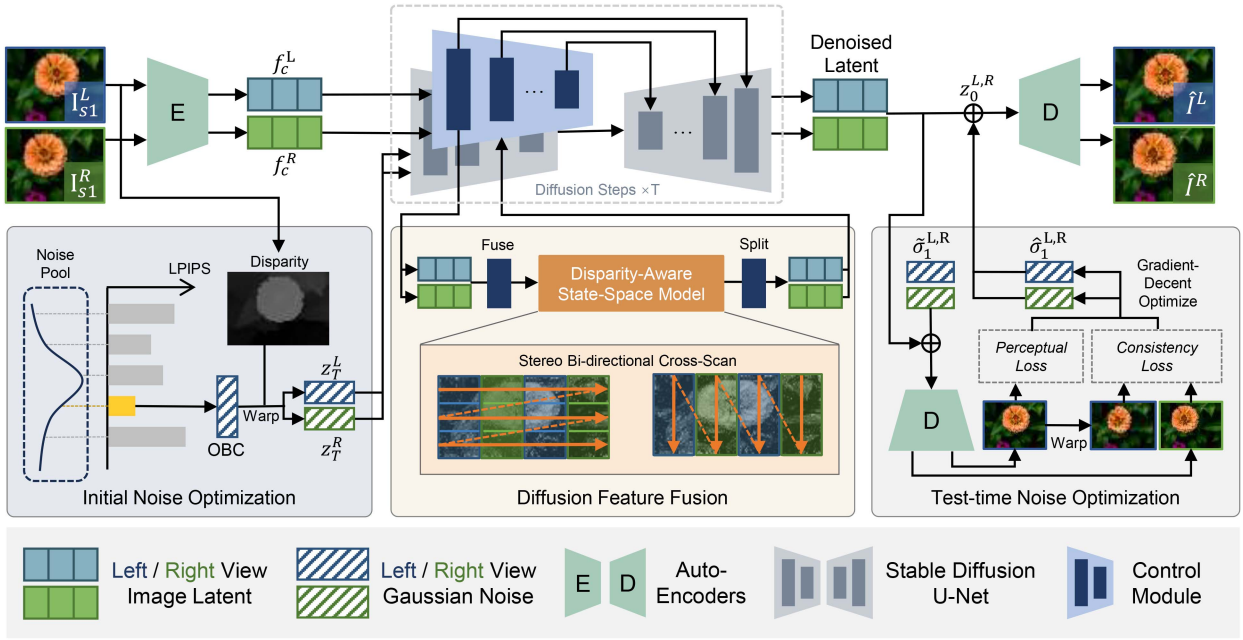


Fig. 2. The framework of our proposed DiffSSR+. The framework progressively optimizes the sampled noise at both initial noise sampling stage where the noise is obtained by the OBC sampling and disparity-aware warping, and the test-time optimization stage where the variance noise is optimized by the combined perceptual-consistency loss. The DASSM performs stereo feature fusion in each diffusion step in a Stereo Bi-directional Cross-Scanning scheme to jointly enhance consistency and perceptual quality.

$$y(t) = Ch(t) + Dx(t). \tag{4}$$

The hidden state is continuously updated by the current state through matrix A and the input signal B in (3), while the output state is obtained by mapping the hidden state with matrix C, and the input with matrix D in (4).

Recent studies have shown that selective state-space models achieve strong performance in efficient feature modeling across both natural language processing and computer vision tasks [40], [41], demonstrating their applicability to a wide range of scenarios.

B. Diffusion-Based Stereo SR Framework

As discussed in Section II-B, Stable Diffusion provides a strong generative prior that can be leveraged in super-resolution tasks to better reconstruct fine-details and improve perceptual performance. In this work, we aim to develop a stereo-friendly diffusion-based super-resolution framework that jointly achieves high perceptual quality from generative modeling while preserving stereo consistency with high fidelity.

Our approach begins by extending the unconditional diffusion noise estimation in (2) by incorporating image latents c as control information:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t)\right) + \boldsymbol{\sigma}_t, \tag{5}$$

we have re-written the variance noise as $\boldsymbol{\sigma}_t$ for brevity.

However, such a single-view information flow lacks explicit stereo interaction, where the complementary information are not analyzed for stereo feature fusion. Moreover, the overall diffusion process is affected by the stochasticity of both the initial noise $\mathbf{x}_T$ and the per-step sampling variance noise $\boldsymbol{\sigma}_t$, which leads to inconsistencies in fine textures across the reconstructed stereo images.

To address these challenges, we propose a progressive Diffusion Stereo-Super-Resolution (DiffSSR+) framework that jointly enhances stereo consistency and perceptual quality from the perspectives of network-level information interaction and diffusion stochasticity optimization. The overall architecture of the proposed framework is illustrated in Fig. 2.

In the preprocessing stage, a fidelity-driven stereo restoration module is first employed to perform coarse reconstruction, aiming to suppress noise and distortion artifacts in the low-resolution stereo inputs I^L and I^R. This module produces coarsely restored intermediate results I^L_{s1} and I^R_{s1}, which provide a more reliable structural foundation for subsequent processing. These intermediate results are then encoded by the Stable Diffusion auto-encoders to obtain the conditional latent representations f^L_c and f^R_c, which serve as conditioning signals for the diffusion process. An optimal boundary condition–based sampling strategy to generate an approximate optimal initial noise $\mathbf{z}^L_T$, and its disparity-warped counterpart $\mathbf{z}^R_T$ is constructed accordingly, enabling stereo-consistent initialization of the denoising process. During diffusion inference, the noisy latent representations are iteratively updated in a stereo-aware manner, allowing effective information exchange between the two views throughout the denoising process. At the final timestep, the stochastic variance noise injected into the latent space is further optimized using a perceptual-consistency loss to reduce residual inter-view discrepancies. The refined latent representations are finally decoded to generate the super-resolved stereo outputs $\hat{I}^L$ and $\hat{I}^R$.

Authorized licensed use limited to: Peking University. Downloaded on November 12,2026 at 03:00:38 UTC from IEEE Xplore. Restrictions apply.

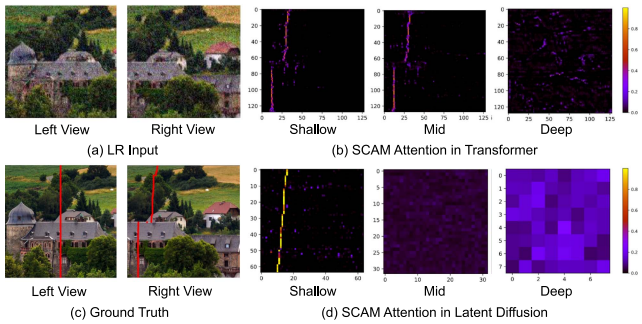


Fig. 3. Attention maps of SCAM applied to two types of SR frameworks. (a) Input LR stereo Images. (b) SCAM Attention in Transformer-based SR network [6] from shallow to deep, where red pixels indicate concentrated cross-view attention. (c) Ground truth. (d) SCAM Attention in Diffusion-based SR network [8].

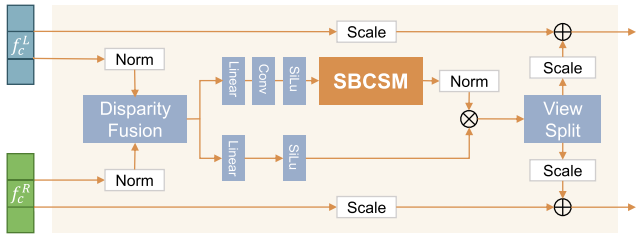


Fig. 4. The structure of the Disparity-Aware State-Space Model (DASSM), where the features from the two views are combined through the disparity-fusion operation and further fused by the Stereo Bi-directional Cross Scan Model (SBCSM).

Following DiffBIR [8], We replicate the encoder and middle-blocks in the SD network, constructing a parallel control branch. Intermediate features from the control branch are fused using the proposed DASSM. The fused features are added back to the denoise branch and passed to the next layers of the control branch. This design enables joint optimization of perceptual quality and stereo consistency via adaptive noise optimization and explicit stereo interaction. Moreover, these designs are backbone-agnostic. As discussed in Section II-B, state-of-the-art diffusion-based SR methods have introduced increasingly complex architectures, such as semantic-prompt-based designs in SeeSR [10] and additional intermediate modules in PASD [13]. Despite these architectural variations, the proposed DiffSSR+ framework can seamlessly adapt to different SoTA diffusion SR backbones and consistently deliver notable performance gains, as demonstrated in Section IV-D.

C. Disparity-Aware State-Space Model

The Stereo Cross-Attention Module (SCAM) [5] provides a possible solution for stereo feature fusion by computing cross-attention between the corresponding spatial pixels in the two views along the same horizontal scanline. However, we observe that SCAM exhibits inherent limitations when applied to diffusion-based super-resolution frameworks. Fig. 3 visualizes the behavior of SCAM when integrated into Transformer-based SR networks such as SwinFIR [6] and diffusion-based networks such as DiffBIR [8]. For a fair comparison, both SCAM models are trained under the same settings (Flickr1024, batch size $= 1$, 30 k iterations), corresponding to the quantitative results reported in Table X. As shown in Fig. 3(b), the SCAM attention maps in Transformer network follow the underlying cross-view disparity structure. Specifically, for a given position in the left view, the attention is concentrated around the disparity-shifted location in the right view, revealing consistent cross-view structural correspondence.

However, in diffusion-based networks, the latent space is inherently more complex, as it entangles both low- and high-level representations. This complexity makes it challenging for SCAM to provide effective stereo guidance, which leads to the inefficient modeling of cross-view correspondences in attention-based models, as shown in Fig. 3(d). Moreover, the limited attention scope of SCAM restricts its ability to capture broader contextual information, which is critical for synthesizing spatially continuous textures in super-resolution.

To address these issues, we propose the Disparity-Aware State-Space Module (DASSM), whose overall architecture is illustrated in Fig. 4. We first perform disparity-warping feature fusion based on stereo geometry to align features associated with the same physical object in the left and right views to neighboring spatial positions. The disparity map is estimated from the intermediate restored images I_{s1}^{L} and I_{s1}^{R}, which guides the feature fusion process. The fused representations are then processed by two parallel branches. One branch employs our proposed Stereo Bi-directional Cross Scan Model (SBCSM) to efficiently capture both horizontal cross-view consistency and vertical intra-view continuity in a linear sequential manner, while the other branch serves as a residual connection.

The detailed operations of the SBCSM are illustrated in Fig. 5. Leveraging the linear scanning property of state-space models, we perform feature modeling and fusion on the stereo cyclopean representation through a bi-directional spatial scanning strategy. Specifically, we design distinct scanning orders along both the horizontal and vertical directions. The horizontal scans focus on analyzing the distribution similarity of stereo features across the two views, enabling the extraction of complementary information that is consistent between views. In contrast, the vertical scans model the intra-view continuity within each image, which is crucial for enhancing the coherence of local texture reconstruction. The bidirectional arrows in Fig. 5(b) and (c) indicate that scanning is performed in both directions and the resulting features are aggregated by summation. In practice, we sequentially apply vertical scanning followed by horizontal scanning to decouple the modeling of intra-view and cross-view information. An ablation study on different scanning orders is presented in Table VIII. We refine the output sequence by reordering it to match the original image order to reduce the learning difficulty of the subsequent projection layers and leads to improved overall performance.

After scanning, the two branch outputs are multiplied and separated back into the left and right views. The final result is obtained by adaptively scaling and integrating the inputs with the processed outputs using learnable parameters.

Authorized licensed use limited to: Peking University. Downloaded on November 12,2026 at 03:00:38 UTC from IEEE Xplore. Restrictions apply.

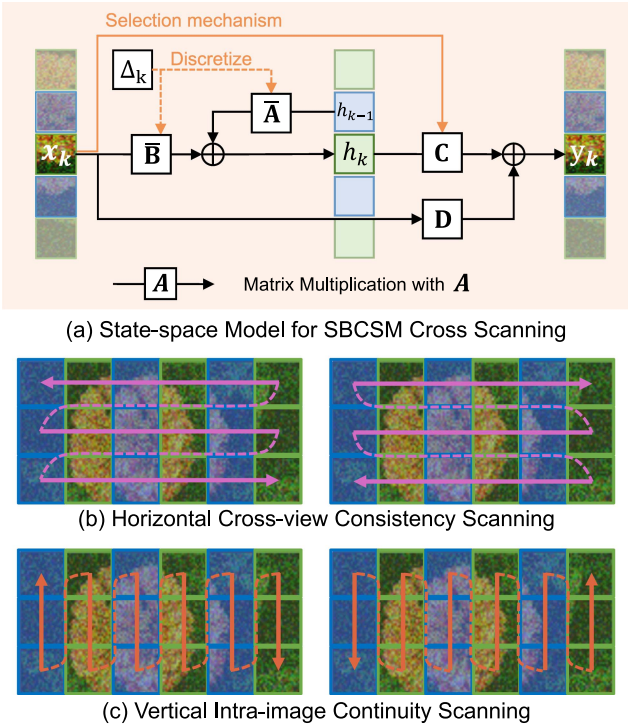


Fig. 5. The model design of Stereo Bi-directional Cross Scan Model and the horizontal / vertical scanning schemes in SBCSM.

DASSM is particularly well suited for stereo fusion in our diffusion-based framework, as it jointly models local pixel-level and global semantic correspondences. Specifically, disparity-aware fusion explicitly captures stereo relationships to enhance local pixel continuity across views, while the dynamic selective update of SBCSM preserves global semantic consistency during scanning.

D. Stereo-Consistent Noise Warping

The randomness of noise sampling in diffusion models introduces stochastic variations in the generated details, which can lead to severe cross-view inconsistencies when stereo super-resolution is performed with independent noise sampling strategy. To address this issue, we observe that the diffusion process becomes approximately deterministic once the sampled noise is fixed. Moreover, the generated content exhibits approximate translation equivariance with respect to synchronized translations of the input noise and low-resolution images. This property forms the theoretical foundation of our noise-warping strategy.

Let the noise input of the diffusion model be denoted as z, the low-resolution conditional input as c, and the resulting super-resolved output as y. We simplify the disparity-based warping operation between stereo views as a translation operation. Specifically, for a translation scale d, we define a translation matrix A_d. Under this formulation, the relationship between the right-view image c_R and the left-view image c_L can be expressed as: $c_R = A_d c_L$. We observe that under stereo conditional inputs, warping the left-view noise to the right view yields an image whose distribution differs minimally from that of the left view.

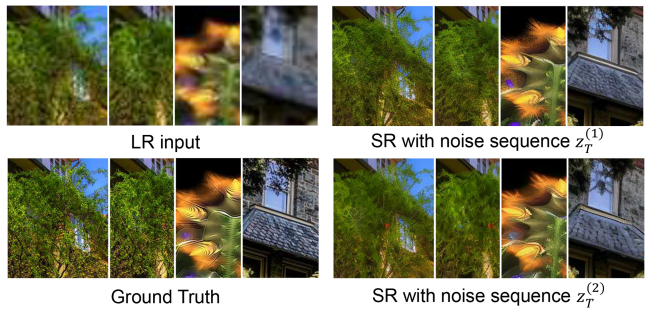


Fig. 6. Similarity of the SR results between different LR images under the same initial noise sequence.

Formally,

$$Az_L = \arg\min_{\epsilon\sim\mathcal{N}} \mathrm{KL}(P(Ay_L \mid Ac_L, \epsilon)) \,\|\, P(y_L \mid c_L, z_L))). \quad (6)$$

Further theoretical justification for (6) is provided in the supplementary materials.

Fig. 6 empirically validates our conclusion on noise translation equivariance. When we apply the same sampled noise sequence on different LR images, the generated SR results tend to exhibit similar high-frequency detail patterns. Therefore, sharing the noise samples between the two views strengthens stereo consistency while avoiding substantial inconsistencies in the reconstructed details.

Motivated by the above theoretical analysis and empirical observations, we adopt a noise-warping strategy to ensure stereo consistency in diffusion-based super-resolution. Specifically, the noise for the left view follows a standard Gaussian distribution, while the noise for the right view is obtained by warping the left-view noise based on the estimated disparity map. Any missing values introduced by occlusions during the warping process are filled with Gaussian noise.

E. Perceptual-Oriented Noise Sampling Optimization

The stereo noise warping strategy effectively alleviates stereo inconsistency caused by independently sampled noise across views. However, the randomness of the initial noise can still introduce visually implausible details, degrading perceptual quality in both views. To address this issue, we propose a perceptual-oriented noise sampling optimization that enhances perceptual quality under stereo-consistency constraints by jointly optimizing the initial noise and the stochastic variance noise during denoising.

Ma et al. [11] analyzed the influence of initial noise on the final reconstruction in diffusion-based super-resolution, and proposed an Optimal Boundary Condition (OBC) sampling strategy. By fixing the initial noise z_T to an optimal point $\tilde{z}_T$ whose approximation can be obtained via sampling based on a reference-set criterion, the overall perceptual quality of images generated based on this initial noise can be improved. Formally, the optimal noise $\tilde{z}_T$ can be selected from a randomly sampled noise pool $\mathcal{K}$ containing K noises. Given a small subset of the training set which contains R pairs of LR-HR pairs

Authorized licensed use limited to: Peking University. Downloaded on November 12,2026 at 03:00:38 UTC from IEEE Xplore. Restrictions apply.

$\mathcal{R} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{R}$, with an image distance measurement function $M(\cdot,\cdot)$, we approximate the optimal noise by:

$$\tilde{z}_T \approx \arg\min_{\epsilon \in \mathcal{K}} \sum_{i=i}^{R} M\left(h_\theta\left(\epsilon, c_i\right), \mathbf{y}_i\right), \tag{7}$$

where c_i is the encoded control signal of the LR input $\mathbf{x}_i$, and M is practically instantialized by the LPIPS [42] metric. The acquired noise $\tilde{z}_T$ is adopted as the initialization of the left view noise, while the right view noise is obtained by disparity-based warping operation to maintain stereo consistency, as shown in the *Initial Noise Optimization* in Fig. 2.

This strategy imposes empirical constraints on the sampling space that guide the diffusion process toward regions with higher-quality outcomes. Although this reduces some stochastic variation, the inherent randomness of the generative prior still allows the outputs to remain diverse and semantically plausible. Sampling under these constraints also produces more stable and visually natural results, while substantially improving cross-view texture consistency.

Despite approximating optimal perceptual performance at noise initialization, the stochastic denoising process of (5) introduces random variance noise $\boldsymbol{\sigma}_t$ at each step, which can disrupt this optimality and generate undesired textures. Observing that fine-texture details are primarily determined by the later diffusion steps, we further refine the noise through test-time optimization at the end of the denoising process, as shown in *Test-time Noise Optimization* in Fig. 2.

Only the variance noise of the final timestep is chosen as the optimization target, since, as shown in Section IV-D, optimizing this single timestep is sufficient to achieve good performance while keeping computational costs manageable. We first randomly sample a random variance noise and perform noise warping to obtain the two view noises $\tilde{\sigma}_1^{L,R}$. The variance noise is added to the last denoised latent to finish the last step latent sampling. The superscript ^{L,R} is omitted here for brevity:

$$\tilde{z}_0 = \frac{1}{\sqrt{\alpha_1}}\left(z_1 - \frac{1-\alpha_1}{\sqrt{1-\bar{\alpha}_1}} h_\theta(z_1, c, t=1) + \tilde{\boldsymbol{\sigma}}_1\right). \tag{8}$$

The final latent is then decoded to acquire a coarse result with decoder $\mathcal{D}$. We optimize the variance noise by leveraging the obtained coarse result, explicitly targeting improvements in both perceptual quality and stereo consistency, guided by a *perceptual-consistency* loss. To be specific, we adopt CLIP-IQA [43] as the base score to form a reference-free test-time adaptable loss:

$$\mathcal{L}_p = -\text{CLIPIQA}\left(\mathcal{D}\left(\tilde{z_0}^{L,R}\right)\right). \tag{9}$$

Since $\mathcal{L}_p$ alone does not explicitly enforce stereo correspondence and may weaken the consistency ensured by noise warping, we further introduce a stereo constraint loss based on perceptual similarity instantiated by LPIPS [42]:

$$\mathcal{L}_c = \text{LPIPS}\left(\mathcal{D}(\tilde{z_0}^{L}), \text{W}\left(\mathcal{D}(\tilde{z_0}^{R}), Dis(I_{s1}^{L}, I_{s1}^{R})\right)\right), \tag{10}$$

where W denotes disparity warping, and $Dis(I_{s1}^{L}, I_{s1}^{R})$ denotes the disparity map calculated from the first-stage coarse SR results. We perform simple first-order gradient decent on the combination of perceptual-consistent loss to optimize and acquire the final denoised latent:

$$\begin{aligned} &\mathcal{L} = \mathcal{L}_p + \mathcal{L}_c, \\ &\hat{\boldsymbol{\sigma}}_1 = \tilde{\boldsymbol{\sigma}}_1 - \gamma \nabla_\sigma \mathcal{L}, \\ &z_0 = \frac{1}{\sqrt{\alpha_1}}\left(z_1 - \frac{1-\alpha_1}{\sqrt{1-\bar{\alpha}_1}} h_\theta(z_1, c, t=1) + \hat{\boldsymbol{\sigma}}_1\right), \end{aligned} \tag{11}$$

where γ is the optimization step size and ∇ is the gradient operator. This noise optimization approach further achieves test-time, case-specific adaptive perceptual quality enhancement, while preventing stereo consistency from being perturbed by perceptual-oriented updates.

IV. Experimental Results

A. Experimental Settings

1) Benchmark: To validate the effectiveness of our proposed framework, we compare our method with several state-of-the-art real-world SISR methods including Real-ESRGAN [3], StableSR [9], DiffBIR [8], DiffIR [44], SeeSR [10] and PASD [13], and stereo-image super-resolution methods including NAFSSR [5], Refusion [45], SwinFIR-SSR [6], CVHSSR [7] and EHFSSR [36]. We conduct super-resolution experiments on two datasets with real-world scenarios sourced from Flickr1024 [46] with simulated degradation, including the Flickr1024RS [47] with simple random degradation and Flickr1024Blind with a more complex degradation process.

For quantitative performance evaluation, two kinds of metrics are adopted in total, including full-reference metrics like LPIPS [42], DISTS [48], WaDIQaM [49], and no-reference metrics containing BRISQUE [50], MUSIQ [51], MANIQA [52] and PI [53].

2) Implementation Details: For training data synthesis, we adopt the Real-ESRGAN two-stage degradation pipeline to generate our paired training data. For our progressive framework construction, we adopt the SwinFIR-SSR [6] as our fidelity-driven stereo restoration module for coarse reconstruction, and the DKT-Stereo module for disparity-map estimation. More details can be found in the supplementary materials.

B. Quantitative Results Analysis

We report quantitative results on Flickr1024RS and Flickr1024Blind in Table I and Table II. Our method achieves consistent improvements across both full-reference and no-reference metrics, indicating a well-balanced trade-off between fidelity and perceptual quality. For example, compared with DiffBIR, our method achieves a 26.4% improvement in LPIPS, reflecting the fidelity gains from stereo consistency optimization. Traditional NR metrics such as BRISQUE and PI tend to favor stochastic texture enhancement in monocular restoration, whereas our stereo consistency constraint suppresses cross-view inconsistent details. In contrast, the semantics-aware metric

Authorized licensed use limited to: Peking University. Downloaded on November 12,2026 at 03:00:38 UTC from IEEE Xplore. Restrictions apply.

TABLE I
QUANTITATIVE RESULTS ON THE FLICKR1024RS DATASET.THE BEST AND SECOND-BEST RESULTS ARE MARKED IN BOLD AND UNDERLINED, AND SYMBOLS ↑ AND ↓ INDICATE WHETHER HIGHER OR LOWER VALUES ARE PREFERRED, RESPECTIVELY.

Methods		LPIPS ↓	DISTS ↓	WaDIQaM ↑	BRISQUE ↓	CLIPIQA ↑	MUSIQ ↑	MANIQA ↑	PI ↓
SISR methods	Real-ESRGAN [3]	0.294	0.163	-0.084	21.138	0.522	67.224	0.385	3.034
	StableSR [9]	0.318	0.156	-0.084	26.584	0.546	64.633	0.365	3.465
	DiffBIR [8]	0.325	0.154	-0.088	10.405	0.663	69.030	0.457	**2.727**
	DiffIR [44]	0.284	0.161	**-0.079**	22.586	0.494	65.528	0.350	3.432
	SeeSR [10]	0.251	0.132	-0.084	17.456	0.647	70.357	0.645	3.062
	PASD [13]	0.278	0.133	-0.086	22.840	0.570	67.243	0.611	3.231
Stereo SR methods	NAFSSR [5]	0.468	0.250	-0.093	53.540	0.349	54.681	0.295	6.132
	Refusion [45]	0.374	0.197	-0.100	14.691	0.662	66.243	0.391	3.319
	SwinFIR-SSR [6]	0.404	0.220	-0.085	52.682	0.394	59.460	0.338	5.767
	CVHSSR [7]	0.338	0.170	-0.065	46.618	0.467	61.300	0.365	4.507
	EHFSSR [36]	0.398	0.156	-0.086	51.130	0.382	52.680	0.511	5.186
Ours	DiffSSR	0.276	0.146	-0.081	**10.363**	0.717	**72.481**	0.580	3.012
	DiffSSR+	**0.232**	**0.118**	-0.081	11.114	0.727	72.561	0.665	3.142
	DiffSSR+ (SeeSR)	0.250	0.126	-0.084	15.733	0.698	71.628	**0.663**	2.954
	DiffSSR+ (PASD)	0.246	0.127	-0.082	19.909	**0.786**	70.272	0.641	3.982

TABLE II
QUANTITATIVE RESULTS ON THE FLICKR1024BLIND DATASET

Methods		LPIPS ↓	DISTS ↓	WaDIQaM ↑	BRISQUE ↓	CLIPIQA ↑	MUSIQ ↑	MANIQA ↑	PI ↓
SISR methods	Real-ESRGAN [3]	0.401	0.224	-0.103	15.304	0.557	65.260	0.397	3.207
	StableSR [9]	0.447	0.218	-0.102	23.310	0.426	55.541	0.285	3.678
	DiffBIR [8]	0.435	0.195	-0.105	**8.903**	0.653	66.540	0.441	**2.586**
	DiffIR [44]	0.388	0.233	-0.100	18.255	0.499	62.493	0.344	3.500
	SeeSR [10]	0.349	0.182	-0.101	20.343	0.605	65.450	0.599	3.600
	PASD [13]	0.412	0.198	-0.108	26.282	0.536	63.482	0.544	3.841
Stereo SR methods	NAFSSR [5]	0.652	0.321	-0.108	66.462	0.262	35.820	0.199	7.656
	Refusion [45]	0.477	0.239	-0.111	14.748	0.671	62.830	0.356	3.330
	SwinFIR-SSR [6]	0.569	0.299	-0.106	62.147	0.282	47.943	0.257	7.267
	CVHSSR [7]	0.739	0.310	-0.114	63.914	0.227	24.491	0.297	7.972
	EHFSSR [36]	0.737	0.309	-0.114	64.406	0.229	23.253	0.298	7.929
Ours	DiffSSR	0.383	0.213	-0.097	14.954	0.687	69.536	0.540	3.203
	DiffSSR+	**0.318**	0.172	-0.097	10.620	0.702	69.893	0.626	3.169
	DiffSSR+ (SeeSR)	0.333	**0.166**	-0.104	16.828	0.685	**70.682**	**0.650**	3.094
	DiffSSR+ (PASD)	0.339	0.186	**-0.095**	20.533	**0.761**	66.318	0.611	4.124

CLIPIQA shows a 39.1% improvement, indicating that the generated results are more perceptually coherent and better aligned with human visual preference.

Compared with our previous work DiffSSR [12], the proposed DiffSSR+ yields substantial gains on Flickr1024, including a 16.3% reduction in LPIPS, a 19.2% reduction in DISTS, and a 14.0% improvement in MANIQA, demonstrating simultaneous optimization of fidelity and perceptual performance. Moreover, we apply DiffSSR+ to two representative diffusion-based SISR backbones, SeeSR and PASD. As shown in Table II, DiffSSR+ improves LPIPS by 3.7% for SeeSR and 17.7% for PASD on Flickr1024Blind, validating its general applicability and effectiveness in extending existing diffusion-based SR backbones to stereo super-resolution.

TABLE III
ABLATION STUDY ON OBC HYPER PARAMETERS

K	R	LPIPS↓	DISTS↓	BRISQUE↓	CLIPIQA↑
None	None	0.328	0.178	14.675	0.682
40	8	0.324	0.180	15.388	0.675
40	16	0.328	0.177	15.252	0.682
80	8	**0.322**	**0.176**	**11.566**	**0.683**
100	8	**0.322**	**0.176**	**11.566**	**0.683**

C. Qualitative Results Analysis

We provide the visualization result comparison in Fig. 7. As the figures have shown, perceptual-driven methods such as DiffBIR, SeeSR, and PASD produce perceptually clearer results for each single view, but introduce noticeable stereo inconsistencies across views, *e.g.*, misaligned vertical textures on building structures. Fidelity-driven methods achieve better stereo consistency but exhibit inferior perceptual quality, particularly in fine details such as text reconstruction. In contrast, our method effectively reduces stereo detail inconsistencies while maintaining strong perceptual clarity in super-resolution, leading to a more favorable perception–consistency trade-off. Additional results are provided in the supplementary material.

D. Ablation Studies

1) OBC Settings: We conduct an ablation study on the OBC hyper-parameters, including the noise pool size K and the noise training set size R, with results summarized in Table III. Increasing R does not consistently improve performance, indicating that the OBC distribution is largely insensitive to the input LR images. In contrast, performance improves with larger K, but the marginal gain reduces as $K \geq 80$, where the same optimal noise is selected. Considering both performance and

Authorized licensed use limited to: Peking University. Downloaded on November 12,2026 at 03:00:38 UTC from IEEE Xplore. Restrictions apply.

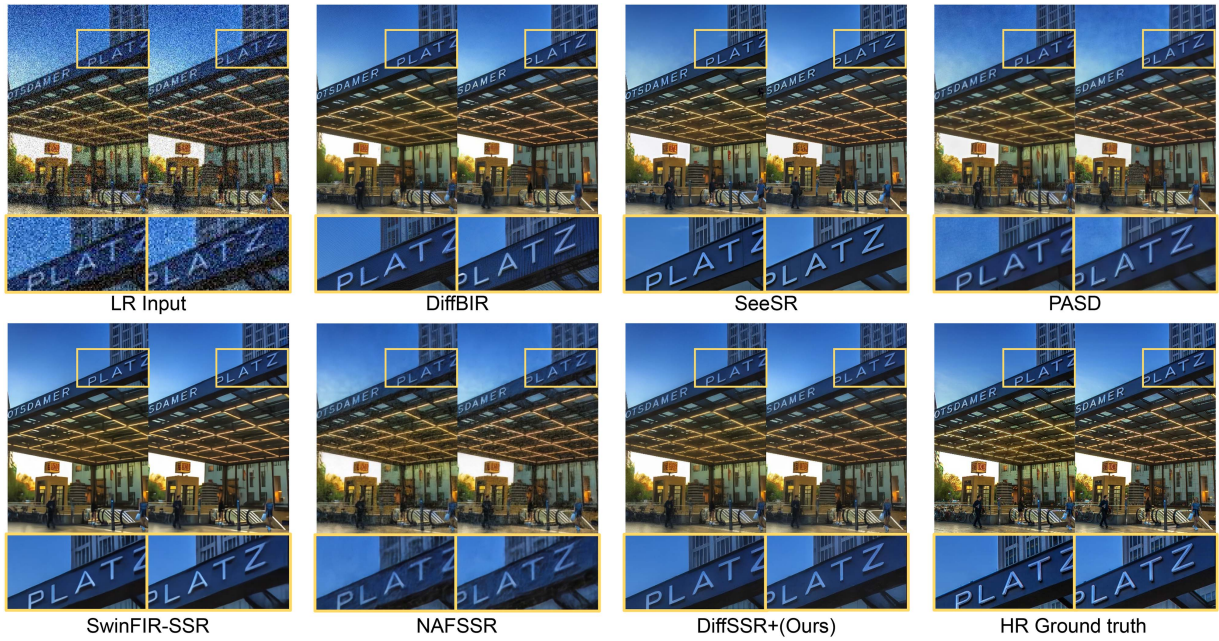


Fig. 7. Qualitative comparison of our proposed DiffSSR+ with other SoTA perceptual-driven SISR methods and fidelity-driven Stereo SR methods.

TABLE IV
ABLATION STUDY ON OBC TRAINING SET SELECTION

Setting	LPIPS↓	DISTS↓	BRISQUE↓	CLIPIQA↑
None	0.328	0.178	14.675	0.682
First	**0.320**	**0.174**	11.560	**0.694**
Random	0.323	0.176	**11.259**	0.692
Representative	0.325	0.176	12.634	0.689

TABLE V
ABLATION STUDY ON NOISE OPTIMIZATION HYPER PARAMETERS

γ	Opt. Step	LPIPS↓	DISTS↓	BRISQUE↓	CLIPIQA↑
None		0.328	0.178	14.675	0.682
1e2	1	0.322	0.176	11.566	0.680
1e3	1	**0.320**	**0.174**	11.560	0.694
1e4	1	0.349	0.178	12.630	0.628
1e3	2	0.321	0.176	12.805	0.697
1e3	3	0.321	**0.174**	**11.558**	**0.701**

TABLE VI
ABLATION STUDY ON NOISE OPTIMIZATION TARGET TIMESTEPS

Target Noise	LPIPS↓	DISTS↓	CLIPIQA↑	Mem. (MB) ↓
None	0.328	0.178	0.682	17556
(L1)	**0.320**	**0.174**	**0.694**	26472
(L1,L2)	0.329	0.177	0.675	71050

computational cost, we set $K = 80$ and $R = 8$ as the final experimental setting.

Moreover, we investigate the impact of the OBC training set selection on SR performance. We consider three settings: using the first R images, randomly sampling R images, and using the first R images under R different representative degradation parameter settings of the RealESRGAN degradation pipeline from the Flickr1024 dataset as the OBC training set. As shown in Table IV, the proposed initial noise optimization consistently improves performance under all training-set selection configurations, demonstrating the robustness of the method to the selection of training set.

2) Backbone Analysis: We further perform qualitative comparisons by applying our method to various diffusion-based SISR backbones. As shown in Fig. 8, the original backbones struggle in stereo SR, exhibiting stereo inconsistencies and degraded perceptual quality. For instance, despite incorporating semantic guidance, SeeSR [10] misinterprets vehicles as stones, resulting in implausible reconstructions. In contrast, our method consistently improves all evaluated backbones by correcting stereo inconsistencies and semantic errors while enhancing perceptual quality, demonstrating strong generality and robustness across different diffusion-based architectures.

3) Noise Optimization Hyperparameters: We analyze the impact of the key hyperparameters in test-time noise optimization in Table V, including the optimization step size γ and the number of optimization steps per sample, which corresponds to the repetitive optimization steps performed on the last sampling noise. The results show that our method is robust to the choice of γ within a reasonable range. Moreover, a single optimization step already yields a substantial improvement in LPIPS, while additional optimization steps primarily contribute to additional improvements on no-reference metrics.

We further analyze the effect of the target timestep for noise optimization in Table VI. Extending the optimization target from the last step (L1) to the last two steps (L1,L2) yields no performance gain while substantially increasing memory consumption, indicating that optimizing only the final timestep is already sufficient and practical.

4) Fusion Scanning Schemes: We conduct an ablation study on different scanning schemes within the SBCSM models. In addition to validating the effectiveness of our bi-directional scanning, we also investigate alternative scanning strategies adopted

Authorized licensed use limited to: Peking University. Downloaded on November 12,2026 at 03:00:38 UTC from IEEE Xplore. Restrictions apply.



Fig. 8. Visualization of the DiffSSR+ framework applied to different diffusion-based SISR backbones, demonstrating consistent improvements in stereo detail and structural consistency across backbones.

TABLE VII
ABLATION STUDY ON FUSION SCANNING SCHEMES

Scanning scheme	LPIPS↓	DISTS↓	BRISQUE↓	CLIPIQA↑
VisMamba	0.469	0.265	**10.769**	0.643
GatedCNN	0.459	0.268	14.965	0.626
Ours-SingleDirection	0.331	0.181	14.535	**0.691**
Ours-BiDirection	**0.328**	**0.178**	14.675	0.682

TABLE VIII
ABLATION STUDY ON FUSION SCANNING ORDERS

Scanning Order	LPIPS↓	DISTS↓	CLIPIQA↓	Runtime(s/img)↓
H-Only	0.331	0.181	0.691	156.27
V-Only	0.342	0.185	0.689	156.60
HV-Simultaneous	0.328	0.178	0.682	158.05
H-then-V	0.324	0.175	**0.696**	163.56
V-then-H	**0.321**	**0.174**	**0.696**	162.57

TABLE IX
MODULE COMPARISON BETWEEN DIFFSSR AND DIFFSSR+

Model	Stereo Restoration	Noise Sampling	Noise Optimization	Stereo Fusion
DiffSSR	SwinFIR	Warp	-	SCSM
DiffSSR+	SwinFIR	Warp+OBC	✓	SBCSM

TABLE X
ABLATION STUDY ON BASIC DIFFSSR COMPONENTS

Stereo Restoration	Noise Sampling	Stereo Fusion	Restoration Guidance	LPIPS↓	DISTS↓	MUSIQ↑
SwinIR	Random	-	-	0.435	0.195	66.540
SwinFIR	Random	-	-	0.413	0.194	66.890
SwinFIR	Same noise	SCAM	-	0.407	0.195	70.419
SwinFIR	Warp noise	SCAM	-	0.399	0.186	70.620
SwinFIR	Warp noise	DASSM	-	0.395	**0.184**	**71.648**
SwinFIR	Warp noise	DASSM	✓	**0.383**	0.212	69.536

in vision tasks, including VisionMamba [54] and GatedCNN-style [55]. The results in Table VII demonstrate the superiority of our bi-directional design over other single-directional scanning schemes.

Furthermore, we investigate the effect of the scanning order in bi-directional scanning on model performance. Specifically, we evaluate five settings: horizontal-only scanning, vertical-only scanning, simultaneous horizontal and vertical scanning, horizontal-to-vertical sequential scanning, and vertical-to-horizontal sequential scanning. Table VIII compares different scanning strategies. Bidirectional scanning consistently surpasses its unidirectional counterparts on FR metrics, while the vertical-to-horizontal scanning order delivers the best overall performance with a small but consistent margin.

5) Basic DiffSSR Components: We further analyze the effectiveness of the basic components in DiffSSR. The full DiffSSR model consists of the coarse restoration module, feature fusion module, classifier-based restoration guidance used in conditional diffusion models, and the noise warping strategy. It excludes SBCSM, initial noise optimization, and test-time variance noise optimization. The specific configuration comparison is shown in Table IX. Accordingly, we remove these components for ablation studies. The results in Table X validate the contribution of each component in our design.

6) Computational Costs: To evaluate the inference overhead of the proposed method, we compare the average per-image inference time and peak GPU memory consumption after individually disabling DASSM and test-time noise optimization on

Authorized licensed use limited to: Peking University. Downloaded on November 12,2026 at 03:00:38 UTC from IEEE Xplore. Restrictions apply.

TABLE XI
COMPARISON OF COMPUTATIONAL COSTS ON DIFFSSR+ MODULES

Version	Runtime (s/img)	Peak GPU Mem. (M)
Baseline	145.53	17548
DiffSSR+ w/o DASSM	146.29	26140
DiffSSR+ w/o NoiseOpt	156.82	17556
DiffSSR+	158.05	26472

1024×1024 crops from the Flickr1024Blind dataset. As shown in Table XI, both designs introduce only marginal extra inference time, while DASSM causes only slight additional memory consumption. The Noise Optimization strategy results in a moderate yet acceptable memory increase due to backpropagation through the decoder, which can be alleviated by reducing the size of the noise patches used for tiling.

V. CONCLUSION

In this work, we introduce a novel diffusion-based stereo super-resolution framework that effectively addresses the fidelity–perception trade-off, particularly the inherent stereo inconsistency in diffusion SR models. We achieve joint improvements by incorporating constraints into the noise sampling process through a stereo-consistent noise sampling optimization strategy, which enhances high-frequency detail consistency and quality. In addition, we propose a DASSM model for improved stereo feature fusion, improving structural and semantic reconstruction between views. Extensive experiments demonstrate the effectiveness of our framework, which can also be readily applied to a broad range of diffusion-based SISR backbones, enabling potential stereo extensions for future state-of-the-art SR methods.

Nevertheless, our work has some limitations left for future exploration. For example, Although our Optimal Initial Noise Sampling does not increase inference overhead and is faster in training than mainstream diffusion training, its computational cost scales linearly with the noise pool and training set size during noise finding, which may constrain further performance improvements. Future work could address these issues by investigating diffusion acceleration techniques, such as few-step inference with model distillation, to improve computational efficiency while achieving even better performance.

REFERENCES

[1] Y. Zhang et al., "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 294–310.

[2] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[3] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *Proc. Int. Conf. Comput. Vis. Workshop*, 2021, pp. 1905–1914.

[4] C. Saharia et al., "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2024.

[5] X. Chu, L. Chen, and W. Yu, "NAFSSR: Stereo image super-resolution using NAFNet," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1238–1247.

[6] D. Zhang et al., "SwinFIR:Revisiting the SwinIR with fast Fourier convolution and improved training for image super-resolution," 2022, *arXiv:2208.11247*.

[7] W. Zou et al., "Cross-view hierarchy network for stereo image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1396–1405.

[8] X. Lin et al., "DiffBIR: Towards blind image restoration with generative diffusion prior," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 430–448.

[9] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy, "Exploiting diffusion prior for real-world image super-resolution," *Int. J. Comput. Vis.*, vol. 132, pp. 5929–5949, 2024.

[10] R. Wu et al., "SeeSR: Towards semantics-aware real-world image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 25456–25467.

[11] Y. Ma, H. Yang, W. Yang, J. Fu, and J. Liu, "Solving diffusion ODEs with optimal boundary conditions for better image super-resolution," in *Proc. Int. Conf. Learn. Representations*, 2024, pp. 33601–33619.

[12] Z. Fan, J. Luo, W. Yang, Z. Guo, and J. Liu, "Progressive interactive diffusion model for real-world stereo super-resolution," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2025, pp. 1–6.

[13] T. Yang, R. Wu, P. Ren, X. Xie, and L. Zhang, "Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 74–91.

[14] C. Tian et al., "Coarse-to-fine CNN for image super-resolution," *IEEE Trans. Multimedia*, vol. 23, pp. 1489–1502, 2021.

[15] A. Shocher, N. Cohen, and M. Irani, "Zero-shot" super-resolution using deep internal learning," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3118–3126.

[16] X. Yang et al., "DRFN: Deep recurrent fusion network for single-image super-resolution with large factors," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 328–337, Feb. 2019.

[17] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5790–5799.

[18] Z. Lu et al., "Transformer for single image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2022, pp. 456–465.

[19] Z. Chen et al., "Dual aggregation transformer for image super-resolution," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 12278–12287.

[20] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 3086–3095.

[21] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 4771–4780.

[22] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "ILVR: Conditioning method for denoising diffusion probabilistic models," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 14347–14356.

[23] H. Chung, B. Sim, and J. C. Ye, "Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp. 12403–12412, 2022.

[24] H. Li et al., "SRDiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47–59, 2022.

[25] Z. Yue, J. Wang, and C. C. Loy, "ResShift: Efficient diffusion model for image super-resolution by residual shifting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 13294–13307.

[26] L. Zhang, W. You, K. Shi, and S. Gu, "Uncertainty-guided perturbation for image super-resolution diffusion model," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 17980–17989.

[27] Z. Yue, K. Liao, and C. C. Loy, "Arbitrary-steps image super-resolution via diffusion inversion," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 23153–23163.

[28] Z.-P. Duan et al., "DiT4SR: Taming diffusion transformer for real-world image super-resolution," in *Proc. Int. Conf. Comput. Vis.*, 2025, pp. 18948–18958.

[29] Y. Luo et al., "Single view stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 155–163.

[30] L. Wang et al., "Learning parallax attention for stereo image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12242–12251.

[31] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10073–10082.

[32] W. Song, S. Choi, S. Jeong, and K. Sohn, "Stereoscopic image super-resolution with stereo consistent feature," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12031–12038.

[33] Y. Wang et al., "Symmetric parallax attention for stereo image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 766–775.

Authorized licensed use limited to: Peking University. Downloaded on November 12,2026 at 03:00:38 UTC from IEEE Xplore. Restrictions apply.

[34] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 17–33.
[35] J. Lin, L. Yin, and Y. Wang, "SteFormer: Efficient stereo image super-resolution with transformer," *IEEE Trans. Multimedia*, vol. 25, pp. 8396–8407, 2023.
[36] J. Song, S. Arcot, and S. Changming, "Efficient hybrid feature interaction network for stereo image super-resolution," *IEEE Trans. Multimedia*, vol. 26, pp. 10094–10105, 2024.
[37] X. Zhu et al., "Cross view capture for stereo image super-resolution," *IEEE Trans. Multimedia*, vol. 24, pp. 3074–3086, 2022.
[38] Y. Liu et al.,, "StereoINR: Cross-view geometry consistent stereo super resolution with implicit neural representation," in *Proc. 33rd ACM Int. Conf. Multimedia*, 2025, pp. 1003–1012.
[39] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.
[40] Y. Liu et al., "VMamba: Visual state space model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 103031–103063.
[41] H. Guo et al., "MambaIR: A simple baseline for image restoration with state-space model," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 222–241.
[42] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
[43] J. Wang, K. C. Chan, and C. C. Loy, "Exploring CLIP for assessing the look and feel of images," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 2555–2563.
[44] B. Xia et al., "DiffIR: Efficient diffusion model for image restoration," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 13049–13059.
[45] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Refusion: Enabling large-size realistic image restoration with latent-space diffusion models," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2023, pp. 1680–1691.
[46] Y. Wang, L. Wang, J. Yang, W. An, and Y. Guo, "Flickr1024: A large-scale dataset for stereo image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2019, pp. 3852–3857.
[47] Y. Zhou et al., "Towards real world stereo image super-resolution via hybrid degradation model and discriminator for implied stereo image information," *Expert Syst. Appl.*, vol. 255, 2024, Art. no. 124457.
[48] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2022.
[49] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.
[50] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
[51] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 5128–5137.
[52] S. Yang et al., "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1190–1199.
[53] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6228–6237.
[54] L. Zhu et al., "Vision Mamba: Efficient visual representation learning with bidirectional state space model," in *Proc. IEEE Int. Conf. Mach. Learn.*, 2024.
[55] W. Yu and X. Wang, "MambaOut: Do we really need mamba for vision?," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 4484–4496.



Jinyi Luo (Student Member, IEEE) received the B.S. degree in artificial intelligence from Peking University, Beijing, China, in 2026, where he is currently working toward the Ph.D. degree with the Wangxuan Institute of Computer Technology. His research interests include deep learning based-image enhancement and low-level computer vision.



Zejia Fan received the B.S. degree in computer science and the Ph.D. degree in computer applied technology from Peking University, Beijing, China, in 2021 and 2026, respectively. Her research interests include image enhanement and generative models.



Wenhan Yang (Member, IEEE) received the B.S degree and Ph.D. degree (with Hons.) in computer science from Peking University, Beijing, China, in 2012 and 2018, respectively. He is currently an Associate Researcher with Pengcheng Laboratory, Shenzhen, Guangdong, China. He has authored more than 50 technical articles in refereed journals and proceedings, and holds nine granted patents. His research interests include image/video processing/restoration, bad weather restoration, human-machine collaborative coding. He was the recipient of the 2023 IEEE Multimedia Rising Star Runner-Up Award, IEEE ICME-2020 Best Paper Award, IFTC 2017 Best Paper Award, IEEE CVPR-2018 UG2 Challenge First Runner-up Award, and MSA-TC Best Paper Award of ISCAS 2022, and was the Candidate of CSIG Best Doctoral Dissertation Award in 2019. He was the Area Chair of IEEE ICME-2021/2022/2023/2024, Session Chair of IEEE ICME-2021, and Organizer of IEEE CVPR-2019/2020/2021 UG2+ Challenge and Workshop.



Jiaying Liu (Fellow, IEEE) received the Ph.D. degree (with Hons.) in computer science from Peking University, Beijing China, 2010. She is currently a Boya Distinguished Professor with the Wangxuan Institute of Computer Technology, Peking University. From 2007 to 2008, she was a Visiting Scholar with the University of Southern California, Los Angeles, California. In 2015, she was a visiting Researcher with Microsoft Research Asia, supported by the Star Track Young Faculties Award. She has authored more than 100 technical articles in refereed journals and proceedings, and holds 70 granted patents. Her research interests include multimedia signal processing, compression, and computer vision. She has been a member of Multimedia Systems and Applications Technical Committee (MSA TC), and Visual Signal Processing and Communications Technical Committee (VSPC TC) in IEEE Circuits and Systems Society. She was ACM ICMR Steering Committee Member and CAS Representative with IEEE ICME Steering Committee. She was also the Deputy Editor in Chief of *APSIPA Transactions on Signal and Information Processing*, an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *IEEE Multimedia Magazine*, and *Journal of Visual Communication and Image Representation*, General Chair of ACM MM Asia 2024, Technical Program Chair of ACM MM Asia 2025/2023/IEEE ICME 2021/ACM ICMR 2021/IEEE VCIP 2019. She was the APSIPA Distinguished Lecturer from 2016 to 2017. She is also a Distinguished Member of CCF and senior member of CSIG.

Authorized licensed use limited to: Peking University. Downloaded on November 12,2026 at 03:00:38 UTC from IEEE Xplore. Restrictions apply.