HiCoSR: High-Fidelity Single-Step Real-World Super-Resolution via Hierarchical Latent Diffusion

Anonymous authors
Paper under double-blind review

Abstract

High-fidelity real-world image super-resolution (RealSR) remains challenging for Latent Diffusion Models (LDMs) due to the fixed $8\times$ spatial compression in pre-trained variational autoencoders (VAEs), which causes fine structural information loss during encoding and often introduces artifacts in generated images. To address this limitation, we propose HiCoSR, a hierarchical and distribution-consistent framework designed to preserve spatial details in the latent space. Specifically, we design a HiCo-Encoder that introduces a parallel $4\times$ latent branch to maintain richer spatial and structural information beyond the capacity of the $8\times$ representation. Based on this dual-scale representation, we adopt a coarse-to-fine cascaded generation strategy to overcome the artifacts and semantic inaccuracies inherent in single-scale denoisers. The coarse $8\times$ stage first establishes a robust global layout, while the fine $4\times$ stage refines local structural details with a single denoising step at each stage. Extensive experiments on three RealSR benchmarks show that HiCoSR achieves strong reconstruction fidelity and comparable perceptual quality to other diffusion-based models, using only a pretrained SD-Turbo backbone.

1 Introduction

Real-world image super-resolution (RealSR) Cai et al. (2019); Ren et al. (2020) aims to restore high-quality images from low-quality inputs with complex, unknown degradations. Unlike ideal SR, which often assumes deterministic downsampling Anwar et al. (2020); Liang et al. (2021); Chen et al. (2025c), RealSR must handle diverse noise, blur, compression artifacts, and domain shifts. Earlier GAN-based methods improve perceptual quality, but they often suffer from unstable training and produce artifacts under severe degradations. Recent methods increasingly exploit the generative priors of pre-trained Text-to-Image (T2I) Latent Diffusion Models (LDMs), which provide stronger open-world texture synthesis than conventional GANs. These priors have substantially improved visual realism, and recent efforts further pursue better structural faithfulness, faster sampling, or stronger foundation backbones. Nevertheless, improving perceptual quality without sacrificing fidelity and efficiency remains challenging.

Despite strong visual capabilities, many LDM-based RealSR methods still inherit a key bottleneck from the spatial compression of pre-trained Variational Autoencoders (VAEs). In the commonly used SD latent space, the VAE compresses images by $8\times$, removing fine structures and high-frequency details during encoding and leaving an information gap that cannot be fully recovered by the decoder. Denoising only in this compressed space lacks deterministic structural anchors and must infer missing textures mainly from learned priors. Recent efforts may turn to stronger foundation backbones or richer latent representations to improve perceptual realism, but the synthesized details are not always grounded in the input image and may introduce semantic deviations, reduce structural fidelity, or increase computational cost. To address these limitations, we propose HiCoSR, a hierarchical and distribution-consistent framework for high-fidelity RealSR. Instead of relying on larger backbones to hallucinate missing details, HiCoSR focuses on preserving more source information before generation. At a high level, it constructs a multi-scale latent representation that keeps fine structures in a higher-resolution latent space while remaining compatible with the distribution of pre-trained T2I priors. One possible way to enrich the latent representation is to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

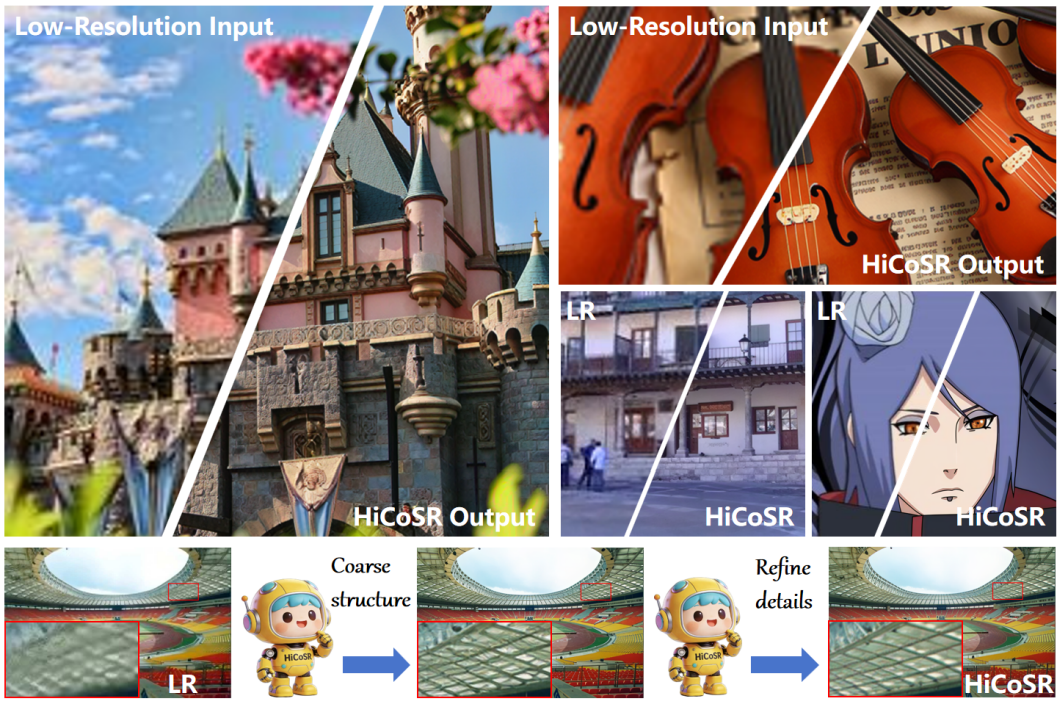

Figure 1: Top: RealSR results with faithful textures and fewer hallucinations. Bottom: the coarse-to-fine pipeline decouples global layout establishment from local detail refinement.

increase the channel dimension, as explored in recent large-backbone or high-capacity autoencoder designs like DiT/FLUX. However, studies suggest that large-channel latents can make diffusion training harder Chen et al. (2025b); Skorokhodov et al. (2025). We instead increase latent spatial resolution, which gives the model more explicit spatial anchors for fine structures and allows the higher-resolution branch to inherit and adapt pretrained lower-resolution diffusion weights.

Built on this representation, HiCoSR adopts a coarse-to-fine generation strategy that separates global structure establishment from local detail refinement. This staged design reduces the burden of single-scale denoising, keeps synthesized textures more tightly grounded in the input image, and enables efficient restoration with a single denoising step at each stage. TVT Yi et al. (2025) is a closely related baseline that also addresses VAE-induced detail loss through transfer VAE training. HiCoSR differs by coupling a parallel $4\times$ latent branch, aligned with the pre-trained latent space through shared-decoder supervision, with coarse-to-fine denoising across the $8\times$ and $4\times$ branches. This design jointly addresses detail preservation, compatibility with pre-trained generative priors, and progressive structural refinement. Extensive experiments on three RealSR benchmarks demonstrate that HiCoSR delivers strong reconstruction fidelity under single-step inference, surpassing the runner-up by +0.22dB PSNR on RealSR while maintaining competitive perceptual quality. Our main contributions are as follows:

- We propose a novel Hierarchical Consistent Encoder to reduce irreversible high-frequency loss from standard VAE compression. It outputs distribution-aligned latents at multiple resolutions, preserving source details and feature integrity.
- We design an efficient SD-Turbo-based multi-scale cascaded generator. Its coarse-to-fine denoising first restores global geometry in low-dimensional space and then refines fine textures in high-resolution latent space.
- Extensive experiments demonstrate that HiCoSR delivers strong reconstruction fidelity on multiple RealSR benchmarks while maintaining competitive perceptual quality and single-step inference efficiency.

2 RELATED WORKS

GAN-Based RealSR. Early RealSR methods Yuan et al. (2018); Bulat et al. (2018); Fritsche et al. (2019) bypassed deterministic assumptions by leveraging Generative Adversarial Networks (GANs) Goodfellow et al. (2014); Adler & Lunz (2018); Karras et al. (2020). Methods such as BSRGAN Zhang et al. (2021b) and Real-ESRGAN Wang et al. (2021) improve texture realism through stochastic degradation simulation and high-order modeling. However, GAN-based approaches remain limited by unstable training and restricted capacity Gulrajani et al. (2017); Mescheder et al. (2018), often producing artifacts under severe degradations.

Diffusion Models for RealSR. Diffusion Models (DMs) Batzolis et al. (2021); Li et al. (2021); Zhao et al. (2022); Saharia et al. (2021b); Ho et al. (2021); Yang et al. (2024a); Saharia et al. (2021a); Zhu et al. (2024); Yang et al. (2024b) have become powerful priors for ill-posed restoration tasks. Early methods such as SR3 Saharia et al. (2022b) perform iterative denoising in pixel space, while Latent Diffusion Models (LDMs) Rombach et al. (2022) reduce computation by shifting denoising into compressed feature spaces. Recent works further improve efficiency and robustness, including ResShift Yue et al. (2023) with residual-space Markov chains and DiffBIR Lin et al. (2024) with pre-trained LDM priors for complex degradations.

Text-to-Image Priors for RealSR. A growing paradigm leverages pre-trained text-to-image (T2I) models Nichol et al. (2021); Saharia et al. (2022a); Xia et al. (2021a;b); Tao et al. (2022); Li et al. (2019); Zhang et al. (2021a); Ding et al. (2021) as generative priors for high-fidelity RealSR. StableSR Wang et al. (2024a), SeeSR Wu et al. (2024b), and OSEDiff Wu et al. (2024a) inject low-resolution conditions through time-aware guidance, semantic awareness, or trainable layers. To reduce computational burden, AdcSR Chen et al. (2025a) adopts adversarial distillation, while TVT Yi et al. (2025) uses transfer VAE training to preserve fine structures. Recent large-backbone methods such as SUPIR Yu et al. (2024), DiT4SR Duan et al. (2025), and FluxSR Li et al. (2025) further improve visual fidelity but introduce heavy memory and computation costs. In contrast, our multi-scale cascaded framework is built upon SD-Turbo and decouples coarse structural synthesis from fine-grained detail refinement for efficient restoration.

3 METHOD

To address the high-frequency detail loss caused by the fixed $8\times$ spatial compression in standard VAEs, we propose HiCoSR, a hierarchical and distribution-consistent framework for high-fidelity RealSR. HiCoSR first introduces a Hierarchical Consistent Encoder, which adds a parallel $4\times$ latent branch to preserve richer spatial structures beyond the capacity of the standard $8\times$ representation. Built upon this dual-scale latent representation, HiCoSR further adopts a coarse-to-fine cascaded denoising strategy: the $8\times$ branch establishes a reliable global layout, while the $4\times$ branch refines local textures and fine-grained structures. This design reduces the burden of single-scale denoising and enables faithful restoration with a single denoising step at each stage.

3.1 HIERARCHICAL CONSISTENT ENCODER

Standard Latent Diffusion Models rely on a pre-trained VAE encoder that compresses an input image of size $H \times W$ into a highly compact latent space of size $\frac{H}{8} \times \frac{W}{8}$. While computationally efficient, this aggressive spatial reduction inevitably eradicates fragile high-frequency textures, establishing a strict upper bound on the restorable details in RealSR tasks. To articulate our proposed solution clearly, we denote the standard highly compressed latent representation with $\frac{1}{8}$ spatial resolution as the $f8$-latent. Conversely, we define the latent representation with $\frac{1}{4}$ spatial resolution as the $f4$-latent, which inherently possesses a larger spatial capacity to encapsulate fine-grained structural information.

Multi-Scale Latent Extraction. To break the detail-preservation bottleneck of standard VAEs while maintaining the efficiency of latent-space processing, we propose the HiCo-Encoder, as illustrated in Figure 2 (a). Featuring a progressive feature fusion architecture, the HiCo-Encoder is designed to extract multi-scale representations without discarding critical source details. During the forward pass, the input image undergoes the standard downsampling blocks of the VAE, yielding the base $f8$-latent that captures the global semantic context. Instead of relying solely on this heavily com-

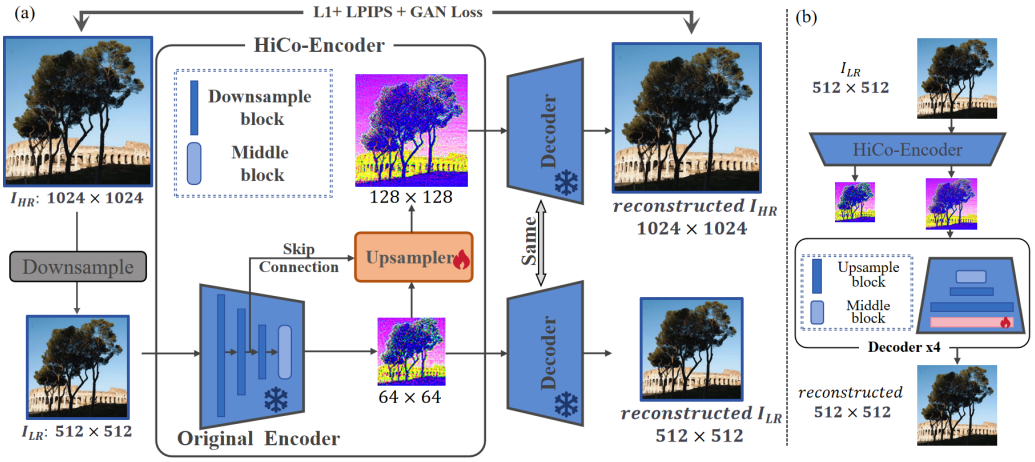

Figure 2: HiCo-Encoder and Decoder $\times 4$. (a) HiCo-Encoder extracts multi-scale latents via a frozen encoder and trainable upsampler. (b) Decoder $\times 4$ reconstructs images from $f4$-latents with matched output resolution.

Table 1: Reconstruction quality comparison of different encoder-decoder configurations.

Configuration	Pixel Fidelity		Perceptual Quality		
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓
Original VAE	33.83	0.9101	0.0601	0.0669	20.25
HiCo-Encoder + Decoder ($8\times$)	39.46	0.9770	0.0197	0.0262	6.23
HiCo-Encoder + Decoder ($4\times$)	**40.95**	**0.9796**	**0.0140**	**0.0171**	**4.23**

pressed representation, we introduce a lightweight upsampling module to systematically recover the spatial dimensions. Specifically, the generated $f8$-latent is spatially expanded and integrated with the intermediate hierarchical features extracted during the encoding phase via explicit skip connections. By fusing the deep semantic base with the shallow structural features at the corresponding resolution, the network effectively constructs the expanded $f4$-latent. This dual-output mechanism ensures that both the coarse global geometry and the fine local textures are explicitly preserved within the latent domain.

Shared-Decoder Supervision. A critical challenge in manipulating multi-scale latents is ensuring that the newly constructed $f4$-latent remains strictly aligned with the identical distribution domain of the standard $f8$-latent. Naive upsampling would cause the latent features to deviate from the original continuous latent space, rendering the subsequent denoising process incorrect. To achieve identical distribution alignment, we propose a shared-decoder supervision strategy. During the independent training of HiCo-Encoder, we utilize a high-resolution ground truth image $I_{HR} \in \mathbb{R}^{3\times1024\times1024}$ and its synthetically downsampled low-resolution counterpart $I_{LR} \in \mathbb{R}^{3\times512\times512}$. The input I_{LR} is fed into our encoder to simultaneously produce a 64×64 $f8$-latent and a 128×128 $f4$-latent. Crucially, both latents are subsequently passed through the exact same frozen pre-trained VAE decoder, which inherently performs an $8\times$ spatial upsampling. The decoded $f4$-latent projects to a 1024×1024 image strictly supervised by I_{HR}. This unified decoding paradigm compels the $f4$-latent to leverage its expanded spatial capacity to encode high-frequency details for the 1024×1024 target, while implicitly aligning it with the same latent distribution space as the $f8$-latent.

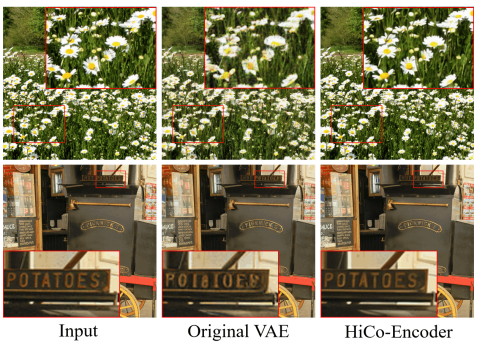

Figure 3: Visualization of HiCo-Encoder.

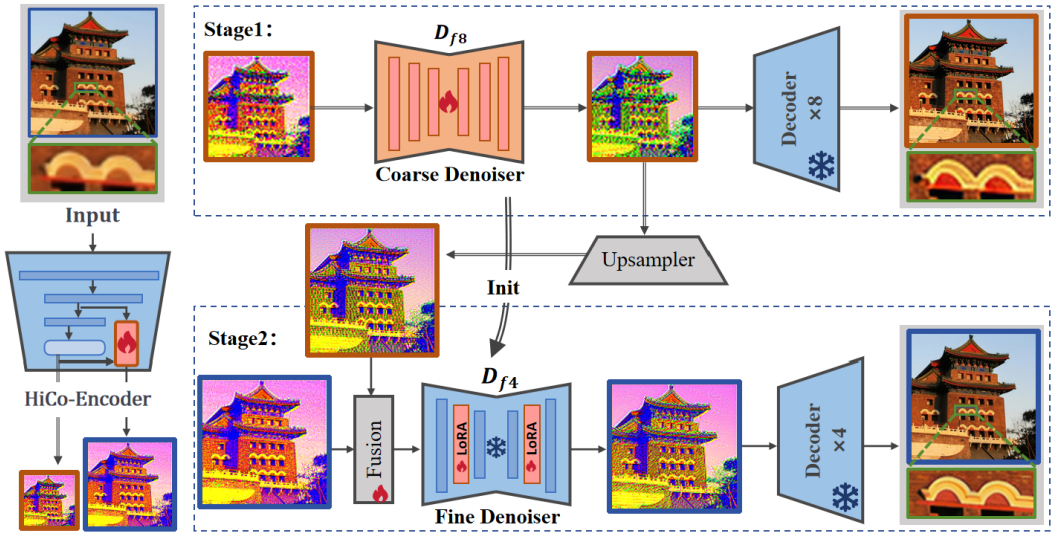

Figure 4: Overall HiCoSR pipeline. The coarse denoiser $\mathcal{D}_{f8}$ establishes global structure, while the fine denoiser $\mathcal{D}_{f4}$ refines local details after spatial upsampling and feature fusion.

Decoder Roles in Training and Inference. The shared Decoder $\times 8$ serves as a frozen supervision network during independent HiCo-Encoder training: it maps the 64×64 coarse latent to 512×512 RGB and the 128×128 fine latent to 1024×1024 RGB. This resolution difference allows the fine branch to learn a richer representation under the same decoding prior. For the subsequent restoration pipeline, the coarse and fine predictions must instead be decoded at the same target resolution. We therefore retain the original Decoder $\times 8$ for the coarse branch and construct a dedicated Decoder $\times 4$ for the fine branch by removing the final upsampling operation and fine-tuning the remaining network separately, as illustrated in Figure 2(b). These decoders are fixed during cascaded denoiser training. At inference, Decoder $\times 4$ converts the final denoised $f4$-latent directly to the target RGB image; for example, a 128×128 fine latent yields a 512×512 output. Decoder adaptation is completed before inference, and no additional pixel-domain resizing is needed to correct the decoding scale.

As summarized in Table 1 and Figure 3, the proposed HiCo-Encoder achieves the best reconstruction metrics and superior visual quality. Compared to the original VAE, it successfully preserves significantly more fine-grained details from the input image.

3.2 Coarse-to-Fine Cascaded Denoising

Existing LDM-based RealSR approaches typically perform denoising within a single-scale latent space, compelling the network to simultaneously reconstruct global semantic structures and fine-grained local details. This dual requirement imposes a severe representation burden on the denoiser. Consequently, the network often falls into a suboptimal compromise: while it might recover sharp local textures, it frequently suffers from semantic inaccuracies and structural artifacts due to the lack of dedicated global guidance. To alleviate this burden and overcome the limitations inherent in single-scale denoisers, we propose a decoupled, coarse-to-fine cascaded generation strategy, as illustrated in Figure 4. The core intuition is to first establish correct semantic information at a lower resolution and subsequently refine local details at a higher resolution. Specifically, the $8\times$ latent space, naturally aligned with pre-trained T2I priors, is exceptionally suited to secure a robust global layout. This reliable semantic prior is then propagated to the $4\times$ latent branch, which provides the essential spatial capacity to further synthesize intricate local details while maintaining strict semantic correctness. This cascaded design elegantly disentangles global semantic layout from local detail refinement, enabling the network to fully exploit the dual-scale features to achieve both exceptional realism and high structural fidelity with a single denoising step at each stage.

Global Semantic Establishment. The cascaded process begins in the $8\times$ latent space to establish a semantically correct and structurally accurate foundation. Given the $f8$-latent z_{f8} extracted by the

Table 2: Quantitative comparison on three benchmarks. **Bold** and underline denote the best and second-best single-step results.

Method	Steps	Params	Fidelity						Perceptual Quality			
			PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FSIM↑	VIF↑	MANIQA↑	MUSIQ↑	FID↓	Q-Align↑
RealSR Benchmark												
StableSR	200	1.41B	24.43	0.7380	0.2636	0.2034	0.8328	0.1671	0.5910	61.80	133.59	3.123
SeeSR	50	2.51B	25.15	0.7210	0.3007	0.2224	0.8411	0.1607	0.6451	69.82	125.39	3.720
DiT4SR	40	2.72B	23.40	0.6662	0.3192	0.2235	0.8183	0.1411	0.6570	67.71	120.92	3.380
SinSR	1	119M	26.26	0.7357	0.3191	0.2358	0.8564	0.1649	0.5420	60.73	134.99	3.166
OSEDiff	1	1.78B	25.15	0.7341	0.2920	0.2128	0.8358	0.1608	0.6335	69.08	123.52	3.696
TSD-SR	1	2.21B	24.82	0.7123	0.2716	0.2091	0.8454	0.1648	0.6344	**70.71**	114.08	3.659
TVT	1	1.72B	25.81	0.7597	0.2594	0.2062	0.8581	**0.1882**	0.6225	69.86	**110.10**	**3.769**
HiCoSR	1	1.03B	**26.48**	**0.7695**	**0.2476**	**0.1817**	**0.8608**	0.1862	**0.6538**	65.36	115.37	3.443
DRealSR Benchmark												
StableSR	200	1.41B	28.41	0.7950	0.2726	0.2069	0.8443	0.1531	0.4914	51.41	151.55	2.750
SeeSR	50	2.51B	28.07	0.7684	0.3174	0.2315	0.8410	0.1469	0.6052	65.09	147.50	3.589
DiT4SR	40	2.72B	25.80	0.6821	0.3656	0.2437	0.7965	0.1259	0.6234	64.51	153.47	3.345
SinSR	1	119M	28.35	0.7488	0.3704	0.2488	0.8476	0.1432	0.4919	55.78	173.10	3.107
OSEDiff	1	1.78B	27.92	0.7836	0.2966	0.2163	0.8445	0.1565	0.5898	64.69	135.42	3.544
TSD-SR	1	2.21B	27.77	0.7490	0.2951	0.2124	0.8419	0.1570	0.5881	**65.92**	**130.38**	3.573
TVT	1	1.72B	28.25	0.7898	0.2900	0.2206	**0.8535**	**0.1684**	0.5784	65.59	134.27	**3.642**
HiCoSR	1	1.03B	**28.57**	**0.7983**	**0.2788**	**0.1968**	0.8526	0.1640	**0.5986**	61.54	138.40	3.373
DIV2K-Val Benchmark												
StableSR	200	1.41B	24.24	0.6265	0.3129	0.1977	0.8293	0.1023	0.5506	58.41	24.26	3.133
SeeSR	50	2.51B	23.68	0.6043	0.3194	0.1968	0.8232	0.0979	0.6222	68.67	25.90	3.977
DiT4SR	40	2.72B	21.79	0.5491	0.3448	0.2108	0.7915	0.0846	0.6415	68.01	30.68	3.721
SinSR	1	119M	24.41	0.6018	0.3238	0.2067	0.8347	0.0984	0.5393	62.78	35.49	3.508
OSEDiff	1	1.78B	23.72	0.6109	0.2942	0.1975	0.8246	0.1011	0.6131	67.96	26.34	3.835
TSD-SR	1	2.21B	22.89	0.5786	**0.2691**	**0.1833**	0.8246	0.1002	0.6168	**70.73**	**23.59**	3.866
TVT	1	1.72B	24.23	0.6292	0.2773	0.1860	**0.8411**	**0.1161**	0.6038	68.67	24.78	**3.905**
HiCoSR	1	1.03B	**24.63**	**0.6374**	0.3004	0.1902	0.8340	0.1100	**0.6120**	66.21	26.93	3.744

HiCo-Encoder, we process it through a coarse denoiser $\mathcal{D}_{f8}$, which is initialized with pre-trained SD-Turbo weights. This stage leverages the pre-trained T2I prior to transform z_{f8} into a clean latent representation at the $8\times$ scale, denoted as $\hat{z}_{f8} = \mathcal{D}_{f8}(z_{f8})$. By isolating the semantic generation to the $8\times$ latent space, we ensure a robust global structure for the subsequent refinement.

Spatial Alignment and Feature Fusion. To propagate the global structural information to the finer $4\times$ latent branch, spatial alignment is required. A direct upsampling operation might shift the established latent distribution, causing a mismatch with the original $f4$-latent. Therefore, we introduce a separately trained spatial upsampler $\mathcal{U}$ that maps $\hat{z}_{f8}$ to the target spatial resolution while preserving distribution consistency, yielding $\tilde{z}_{f8} = \mathcal{U}(\hat{z}_{f8})$. This upsampled latent is then integrated with the original $f4$-latent z_{f4} using a lightweight fusion module $\mathcal{F}$ to obtain the unified representation $z_{fused} = \mathcal{F}(\tilde{z}_{f8}, z_{f4})$. This tailored fusion mechanism effectively bridges the multi-scale representations without introducing domain shifts.

Fine Detail Refinement. Finally, the unified representation z_{fused} is fed into the fine denoiser $\mathcal{D}_{f4}$. To maximize efficiency and preserve the strong generative priors, $\mathcal{D}_{f4}$ shares the core architecture and pre-trained weights with the coarse denoiser $\mathcal{D}_{f8}$. To adapt these shared weights to the higher-resolution $4\times$ latent space, we integrate Low-Rank Adaptation (LoRA) Hu et al. (2022) modules into $\mathcal{D}_{f4}$ for parameter-efficient fine-tuning. Specifically, for a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$ within the shared layers, the adaptation is formulated as:

$$W = W_0 + \Delta W = W_0 + \frac{\alpha}{r} BA, \tag{1}$$

where $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$ are trainable low-rank matrices with the rank $r \ll \min(d, k)$, and α is a constant scaling factor. Guided by the reliable semantic prior established in the previous stage, $\mathcal{D}_{f4}$ focuses entirely on refining local structural details and synthesizing high-fidelity textures. This allows the network to predict the final clean latent representation at the $4\times$ scale via $\hat{z}_{f4} = \mathcal{D}_{f4}(z_{fused})$. By decoupling the generation process and employing efficient weight adaptation, our framework successfully synthesizes high-resolution details while strictly adhering to the global semantics, with a single denoising step at each stage.

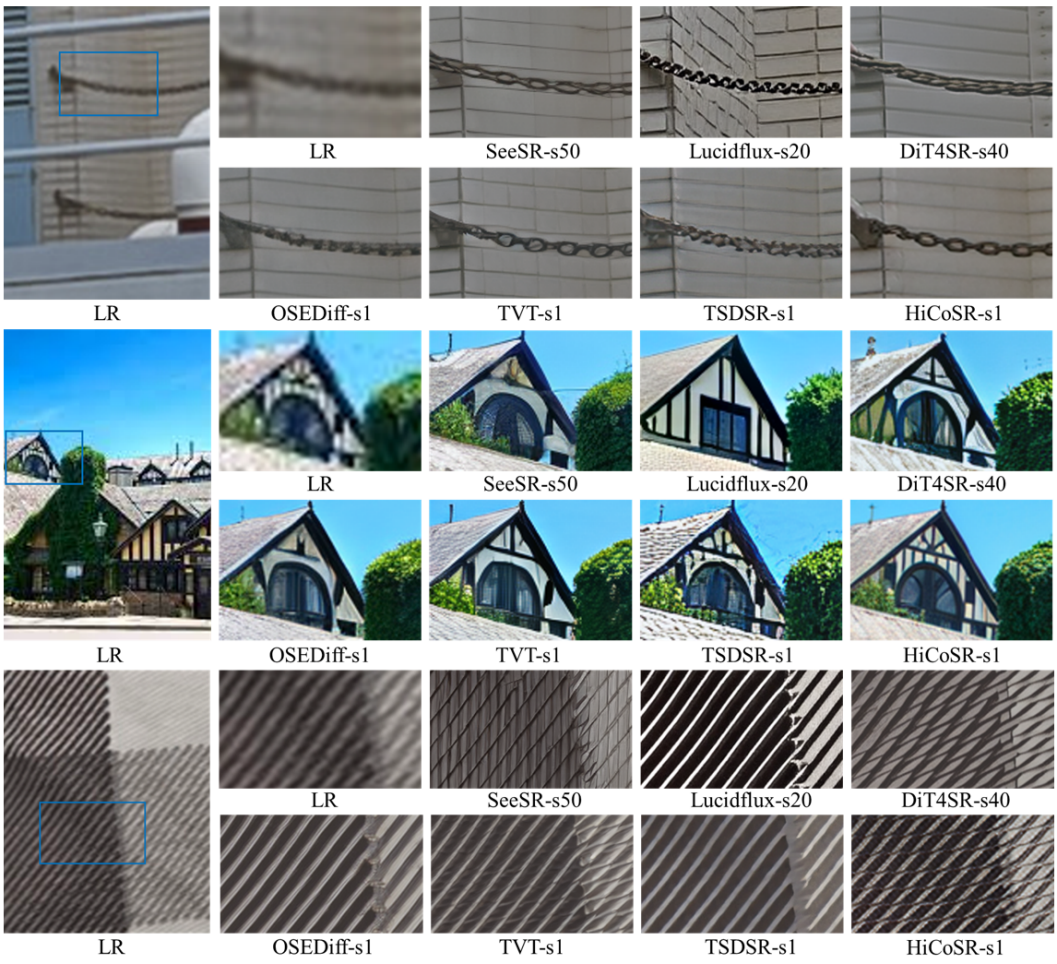

Figure 5: Visual comparison of our HiCoSR against state-of-the-art RealSR methods on real-world images.

Table 3: Ablation study on RealSR benchmark.

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MANIQA↑	MUSIQ↑	FID↓	Q-Align↑
w/o $\mathcal{D}_{f4}$	25.36	0.7331	0.2673	0.1978	0.6398	65.49	**107.31**	**3.527**
w/o HiCo	24.61	0.7223	0.2912	0.2142	0.6316	**66.05**	125.87	3.512
w/o $\mathcal{D}_{f8}$	26.07	0.7664	**0.2445**	0.1875	0.6407	65.16	112.98	3.356
HiCoSR	**26.48**	**0.7695**	0.2476	**0.1817**	**0.6538**	65.36	115.37	3.443

4 Experiment and Analysis

4.1 Experimental Setting

Implementation and Training Details. Our coarse-to-fine cascaded denoising framework is trained on a hybrid dataset consisting of 500K samples from LSDIR and 10K high-quality facial images from FFHQ. During training, we crop random 512×512 patches from these references and synthesize their low-quality counterparts using the stochastic high-order degradation pipeline Wang et al. (2021). For model initialization, the coarse-stage denoiser $\mathcal{D}_{f8}$ leverages pre-trained weights from SD-Turbo, while the fine-stage denoiser $\mathcal{D}_{f4}$ inherits these weights with additional zero-initialized LoRA adapters to ensure optimization stability. For adversarial training, we use a frozen DINOv3 Siméoni et al. (2025) backbone to extract multi-level features, followed by a lightweight decoder head that distinguishes generated images from ground-truth images. The total optimization objective $\mathcal{L}_{total}$ combines three key terms for reconstruction fidelity, semantic realism, and latent

stability:

$$\mathcal{L}_{total} = \lambda_{mse}\mathcal{L}_{mse} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{out}\mathcal{L}_{out}. \tag{2}$$

Specifically, we employ MSE loss ($\mathcal{L}_{mse}$) for pixel-wise reconstruction, adversarial loss $\mathcal{L}_{adv}$ based on the DINOv3-backboned discriminator for structure-aware realism, and outlier regularization loss $\mathcal{L}_{out}$ to penalize extreme latent deviations. Here, $\mathcal{L}_{adv}$ denotes the generator-side hinge adversarial objective:

$$\mathcal{L}_{adv} = -\mathbb{E}_{I_{gen}}\left[D(I_{gen})\right], \tag{3}$$

where the discriminator is trained to separate generated images from ground-truth images. The outlier regularization is defined as:

$$\mathcal{L}_{out} = \left\|\max(0, |z_{gen} - z_{gt}| - \tau)\right\|_2^2. \tag{4}$$

where z_{gen} and z_{gt} denote the latent representations of the generated and ground-truth images, respectively. For the fine stage, we empirically set $\lambda_{mse} = 1.0$, $\lambda_{adv} = 0.25$, and $\lambda_{out} = 10.0$, with the threshold $\tau = 1.5$. All models are optimized using the Adam optimizer with a batch size of 8 and a constant learning rate of 5×10^{-5}.

Two-Stage Training of the Cascaded Denoisers. After independently training HiCo-Encoder, the intermediate upsampler, and Decoder $\times 4$, we optimize the denoisers sequentially. In the coarse stage, we initialize $\mathcal{D}_{f8}$ from SD-Turbo and update all of its parameters while keeping the encoder and Decoder $\times 8$ frozen. We use $\lambda_{adv} = 1.0$ to emphasize structure-aware supervision at this stage. In the fine stage, $\mathcal{D}_{f4}$ inherits the trained coarse-stage weights, which are then frozen, and only its newly inserted LoRA adapters are optimized within the denoiser. The encoder and separately trained Decoder $\times 4$ remain fixed. We reduce λ_{adv} to 0.25 as the fine branch refines local details under the structural guidance established by the coarse branch. Both stages use the same three-term objective, with all other loss weights and training settings unchanged. At inference, the two trained denoisers are applied sequentially, with one denoising step per stage, followed by the final Decoder $\times 4$.

Evaluation Protocol. We evaluate HiCoSR on three benchmarks: RealSR Cai et al. (2019), DRealSR Wei et al. (2020), and DIV2K-Val Agustsson & Timofte (2017), where DIV2K-Val LR inputs are synthesized using the Real-ESRGAN degradation pipeline Wang et al. (2021). We compare against representative multi-step diffusion methods, including StableSR Wang et al. (2024a), SeeSR Wu et al. (2024b), and DiT4SR Duan et al. (2025), as well as single-step methods including OSEDiff Wu et al. (2024a), SinSR Wang et al. (2024b), TSD-SR Dong et al. (2024), and TVT Yi et al. (2025). All baselines are evaluated with their official models and recommended configurations. Following common RealSR protocols, we report fidelity metrics (PSNR/SSIM on the Y channel, LPIPS, DISTS, FSIM, and VIF) and no-reference perceptual metrics (MANIQA, MUSIQ, FID, and Q-Align); metric directions are indicated in Table 2.

4.2 Comparison with State-of-the-Arts

Quantitative comparisons. Tab. 2 presents comprehensive quantitative comparisons on three benchmarks. HiCoSR shows **strong reconstruction fidelity**, achieving the best or second-best scores across all six fidelity metrics on RealSR and DRealSR. On RealSR, it obtains the best PSNR (26.48dB), SSIM (0.7695), LPIPS (0.2476), DISTS (0.1817), and FSIM (0.8608), surpassing the runner-up by **+0.22dB PSNR** and **11.9%** relative DISTS improvement. On DRealSR and DIV2K-Val, HiCoSR also achieves the best PSNR/SSIM, confirming that the HiCo-Encoder preserves structural information typically lost in standard VAE compression. For no-reference perceptual metrics, HiCoSR remains competitive with diffusion-based methods while prioritizing faithful reconstruction. Notably, all results are achieved with **a single denoising step at each stage**, whereas DiT4SR requires 40 steps and StableSR requires 200 steps.

Efficiency Comparison. Table 4 compares inference latency. Although cascaded denoising adds computation, omitting the CLIP text encoder and RAM captioning model keeps HiCoSR's latency (0.13 s) comparable to single-step baselines. HiCoSR also has the second-smallest parameter count (1.03B) among the compared methods (Table 2).

Qualitative Comparison. Figure 5 compares HiCoSR with state-of-the-art RealSR methods. Existing methods often struggle to balance detail restoration and artifact suppression, whereas ours generates natural textures with higher fidelity. Specifically, in the first example (top row), methods like TSDSR introduce severe structural distortions and unnatural artifacts when reconstructing

Table 4: Inference latency on an RTX 5090 at 512×512 resolution.

	StableSR	SeeSR	DiT4SR	SinSR	OSEDiff	TSD-SR	TVT	**HiCoSR**
Time (s)	10.09	4.08	10.00	0.11	0.12	0.08	0.23	**0.13**

the chain and wall lines. Meanwhile, SeeSR and DiT4SR tend to over-smooth the image, failing to recover the distinct links of the chain. In the second example (bottom row) containing intricate structures, baselines such as TVT hallucinate incorrect, rigid window grilles, and SeeSR completely loses the window details. TSDSR also suffers from noticeable noise and messy artifacts around the roof area. In contrast, HiCoSR restores the chain links and curved window grilles more faithfully, producing fewer hallucinations and cleaner local structures under real-world degradations.

4.3 Ablation Study

To validate the effectiveness of each core component, we conduct ablation studies on the RealSR benchmark (Tab. 3, Fig. 6). **(1) w/o** $\mathcal{D}_{f4}$: Removing the fine denoiser restricts the model to perform denoising solely within the $8\times$ compressed latent space, resulting in significant fidelity degradation (-1.12dB PSNR, -0.0364 SSIM). This confirms that the pre-trained VAE discards fine-grained structural information, necessitating the $4\times$ branch to restore local details. **(2) w/o HiCo-Encoder**: Severing the feature fusion with the authentic f4-latent and relying only on upsampled $f8$ features causes the most severe performance drop (-1.87dB PSNR, -0.0472 SSIM, +0.0325 DISTS). Without high-frequency constraints from the source image, the network is forced to hallucinate details, proving the decisive value of HiCo-Encoder for high-fidelity restoration. **(3) w/o** $\mathcal{D}_{f8}$: Ablating the coarse denoiser forces the network to handle both global semantics and local details in the $4\times$ space simultaneously. While fidelity remains competitive, perceptual quality degrades (MANIQA -0.0131, Q-Align -0.087) with structural distortions, demonstrating that global prior guidance is indispensable for coherent detail refinement. In summary, these three components are highly complementary and indispensable for achieving optimal fidelity-perception balance.

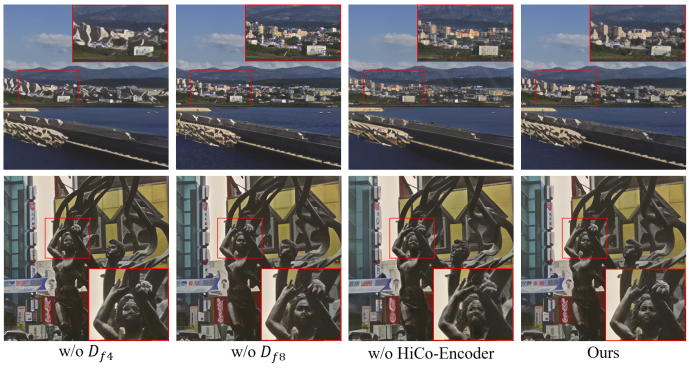

Figure 6: Ablation visual results.

5 Conclusion

In this paper, we presented HiCoSR, which tackles the fundamental information loss bottleneck in LDM-based RealSR caused by aggressive VAE compression. Our HiCo-Encoder introduces a parallel $4\times$ latent branch to preserve fine-grained structural details, while a shared-decoder supervision strategy ensures strict distribution alignment with the original latent space. Building on this hierarchical representation, our coarse-to-fine cascaded denoising explicitly decouples global semantic establishment from local detail refinement, enabling faithful texture recovery with efficient single-step inference. Extensive experiments validate that HiCoSR achieves strong reconstruction fidelity across multiple benchmarks, with consistent gains in key fidelity-oriented metrics over existing single-step methods. Remarkably, these gains are obtained using only a lightweight SD-Turbo backbone with single-step inference, demonstrating that addressing the root cause of detail loss is more effective than brute-force model scaling. We believe this work provides new insights into leveraging multi-scale latent spaces for faithful image super-resolution.

REFERENCES

Jonas Adler and Sebastian Lunz. Banach wasserstein gan. *Advances in neural information processing systems*, 31, 2018.

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

Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *ACM computing surveys (CSUR)*, 53(3):1–34, 2020.

Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *CoRR*, abs/2111.13606, 2021. URL https://arxiv.org/abs/2111.13606.

Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 185–200, 2018.

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

Bin Chen, Gehui Li, Rongyuan Wu, Xindong Zhang, Jie Chen, Jian Zhang, and Lei Zhang. Adversarial diffusion compression for real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28208–28220, 2025a.

Junyu Chen, Chongjian Ge, Enze Xing, Hao He, Jing Wang, Tianxin Wu, Lechao Lin, Shuyang Lin, Zhaoyu Guo, Jun Yang, Ping Luo, Huchuan Lu, and Song Han. DC-AE 1.5: Accelerating diffusion model convergence with structured latent space. *arXiv preprint arXiv:2508.00413*, 2025b.

Xiangyu Chen, Xintao Wang, Wenlong Zhang, Xiangtao Kong, Yu Qiao, Jiantao Zhou, and Chao Dong. Hat: Hybrid attention transformer for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025c.

Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers, 2021. URL https://arxiv.org/abs/2105.13290.

Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. *arXiv preprint arXiv:2411.18263*, 2024.

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

Song Fei, Tian Ye, Lujia Wang, and Lei Zhu. Lucidflux: Caption-free universal image restoration via a large-scale diffusion transformer. *arXiv preprint arXiv:2509.22414*, 2025.

Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3599–3608. IEEE, 2019.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *CoRR*, abs/2106.15282, 2021. URL https://arxiv.org/abs/2106.15282.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. Controllable text-to-image generation. *arXiv preprint arXiv:1909.07083*, 2019.

Haoying Li, Yifan Yang, Meng Chang, Huajun Feng, Zhihai Xu, Qi Li, and Yue-ting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *CoRR*, abs/2104.14951, 2021. URL https://arxiv.org/abs/2104.14951.

Jianze Li, Jiezhang Cao, Yong Guo, Wenbo Li, and Yulun Zhang. One diffusion step to real-world super-resolution via flow trajectory distillation. *arXiv preprint arXiv:2502.01993*, 2025.

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*, pp. 1833–1844, 2021.

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

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pp. 3481–3490. PMLR, 2018.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. *CoRR*, abs/2112.10741, 2021. URL https://arxiv.org/abs/2112.10741.

Haoyu Ren, Amin Kheradmand, Mostafa El-Khamy, Shuangquan Wang, Dongwoon Bai, and Jungwon Lee. Real-world super-resolution using generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 436–437, 2020.

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*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *CoRR*, abs/2111.05826, 2021a. URL https://arxiv.org/abs/2111.05826.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement, 2021b. URL https://arxiv.org/abs/2104.07636.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022a. URL https://arxiv.org/abs/2205.11487.

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*, 45(4):4713–4726, 2022b.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Improving the diffusability of autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54915–54943. PMLR, 2025.

Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16515–16525, 2022.

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*, 132(12):5929–5949, 2024a.

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

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

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 *Proceedings of the European Conference on Computer Vision*, 2020.

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

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

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a.

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Towards open-world text-guided face image generation and manipulation. *arxiv preprint arxiv: 2104.08910*, 2021b.

Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, and Bin Cui. Improving diffusion-based image synthesis with context prediction, 2024a. URL https://arxiv.org/abs/2401.02015.

Ling Yang, Bohan Zeng, Jiaming Liu, Hong Li, Minghao Xu, Wentao Zhang, and Shuicheng Yan. Editworld: Simulating world dynamics for instruction-following image editing, 2024b. URL https://arxiv.org/abs/2405.14785.

Qiaosi Yi, Shuai Li, Rongyuan Wu, Lingchen Sun, Yuhui Wu, and Lei Zhang. Fine-structure preserved real-world image super-resolution via transfer vae training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12415–12426, 2025.

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*, pp. 25669–25680, 2024.

Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 701–710, 2018.

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*, 36: 13294–13307, 2023.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a.

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*, pp. 4791–4800, 2021b.

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

Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations, 2022. URL https://arxiv.org/abs/2207.06635.

Haowei Zhu, Ling Yang, Jun-Hai Yong, Hongzhi Yin, Jiawei Jiang, Meng Xiao, Wentao Zhang, and Bin Wang. Distribution-aware data expansion with diffusion models, 2024. URL https://arxiv.org/abs/2403.06741.

Table 5: Quantitative comparison of text conditioning strategies on RealSR benchmark.

Setting	Fidelity				Perceptual Quality			
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MANIQA↑	MUSIQ↑	FID↓	Q-Align↑
Fixed caption	26.47	0.7656	0.2517	0.1892	0.6191	61.6	109.9	3.188
Learnable embedding	26.48	0.7695	0.2476	0.1817	0.6538	65.36	115.37	3.443

A Implementation Details

In this section, we provide additional experimental settings for individually trained components and DINOv3-based adversarial supervision.

Training of HiCo-Encoder and Related Components. Several components in our framework are trained separately. For the HiCo-Encoder, we employ a combination of L1 loss, LPIPS loss Zhang et al. (2018), and PatchGAN loss Demir & Unal (2018) in the pixel domain. The weight ratio for these three losses is set to $1.0 : 1.0 : 0.1$, respectively. The same loss functions and weights are adopted when independently training the intermediate upsampler and the decoder $\times 4$.

DINOv3-based Adversarial Supervision. We use a DINOv3-backboned discriminator as an auxiliary training signal to encourage structure-aware realism. Specifically, the DINOv3 Siméoni et al. (2025) backbone is kept frozen, and we extract feature maps from multiple intermediate layers. Shallow features provide local boundary and texture cues, while deeper features provide higher-level semantic context. These multi-level features are processed by a lightweight decoder head to distinguish generated images from ground-truth images. We train this discriminator with the standard hinge adversarial objective:

$$\mathcal{L}_D = \mathbb{E}_x \left[\max(0, 1 - D(x))\right] + \mathbb{E}_{\hat{x}} \left[\max(0, 1 + D(\hat{x}))\right], \tag{5}$$

where x and $\hat{x}$ denote ground-truth and generated images, respectively. The generator-side adversarial term is $\mathcal{L}_{adv} = -\mathbb{E}_{\hat{x}}[D(\hat{x})]$, which is used together with the reconstruction and latent regularization losses described in the main paper. This discriminator is used only during training and introduces no additional inference cost.

B Caption-Free Adaptation Strategy

Our framework employs the pre-trained SD-Turbo as the base generative model. Since SD-Turbo is originally designed for Text-to-Image (T2I) generation, its native architecture includes a CLIP text encoder to extract conditional features from input prompts. However, adapting T2I models to real-world image super-resolution (RealSR) using explicit text prompts introduces two major limitations. First, extracting or generating accurate captions for degraded images adds unnecessary computational overhead, which limits inference efficiency. Second, semantic text guidance may induce hallucinations, leading to uncontrollable outputs that compromise image fidelity. Therefore, unlike previous methods such as SeeSR and OSEDiff, HiCoSR is designed as a caption-free framework.

To adapt the pre-trained T2I model to this caption-free setting, we replace explicit text prompts with a globally learnable embedding. During training, this embedding is optimized end-to-end together with the restoration network, enabling the model to learn an implicit conditional distribution tailored to RealSR. To validate this design, we compare it with a fixed-caption baseline using "*a high-resolution image full of vivid details*." As shown in Table 5, the learnable embedding consistently improves fidelity and perceptual metrics, demonstrating the effectiveness of task-specific implicit conditioning.

C Additional Ablation Study

In this section, we provide additional ablation studies to further validate the rationale and effectiveness of key component designs in our framework.

Ablation on Discriminator Architectures. We use a DINOv3-based discriminator to provide structure-aware supervision during adversarial training. To investigate the impact of different fea-

Table 6: Ablation study on RealSR benchmark using different feature extractors as the discriminator backbone.

Setting	Fidelity				Perceptual Quality			
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MANIQA↑	MUSIQ↑	FID↓	Q-Align↑
DINOv1	25.34	0.7581	0.2545	0.2055	0.5964	62.53	119.5	2.993
XCIT	25.28	0.7487	0.2597	0.2051	0.6357	**65.44**	128.89	3.315
DINOv3	**26.48**	**0.7695**	**0.2476**	**0.1817**	**0.6538**	65.36	**115.37**	**3.443**

Table 7: Ablation study on RealSR benchmark using different decoder designs.

Setting	Fidelity				Perceptual Quality			
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MANIQA↑	MUSIQ↑	FID↓	Q-Align↑
Decoder $\times 8$	25.47	0.7396	0.2494	0.1842	0.6864	67.84	108.69	3.355
Decoder $\times 4$	26.48	0.7695	0.2476	0.1817	0.6538	65.36	115.37	3.443

ture spaces, we compare DINOv3 with other feature extractors, including XCIT and DINOv1. As shown in Table 6, the DINOv3-based discriminator achieves the best overall balance across fidelity and perceptual metrics.

Ablation on Decoder Design. We compare the dedicated Decoder $\times 4$ described in Section 3.1 with the original Decoder $\times 8$ followed by pixel-domain downsampling to the same target resolution. As shown in Table 7, Decoder $\times 4$ achieves comparable quality while directly matching the target resolution and reducing unnecessary upsampling overhead. Therefore, we adopt the fine-tuned Decoder $\times 4$ in HiCoSR.

D Additional Visual Comparison Results

We provide additional qualitative results to further evaluate the visual performance of HiCoSR. Figure 7 and Figure 8 compare our model against recent RealSR baselines, including SeeSR Wu et al. (2024b), DiT4SR Duan et al. (2025), LucidFlux Fei et al. (2025), OSEDiff Wu et al. (2024a), TSD-SR Dong et al. (2024), and TVT Yi et al. (2025). Despite using SD-Turbo Sauer et al. (2024) in a caption-free single-step setting, HiCoSR recovers realistic textures while preserving faithful structures.

To illustrate the effectiveness of the cascaded generative structure, Figure 9 visualizes stage-wise outputs. The coarse denoiser $\mathcal{D}_{f8}$ establishes the global structure, while the fine denoiser $\mathcal{D}_{f4}$ leverages the preserved $4\times$ latent information to correct local distortions and refine textures.

E Use of AI Assistance

AI assistance was used only for grammar checking and text polishing. All research ideas, methods, and analyses are solely by the authors.

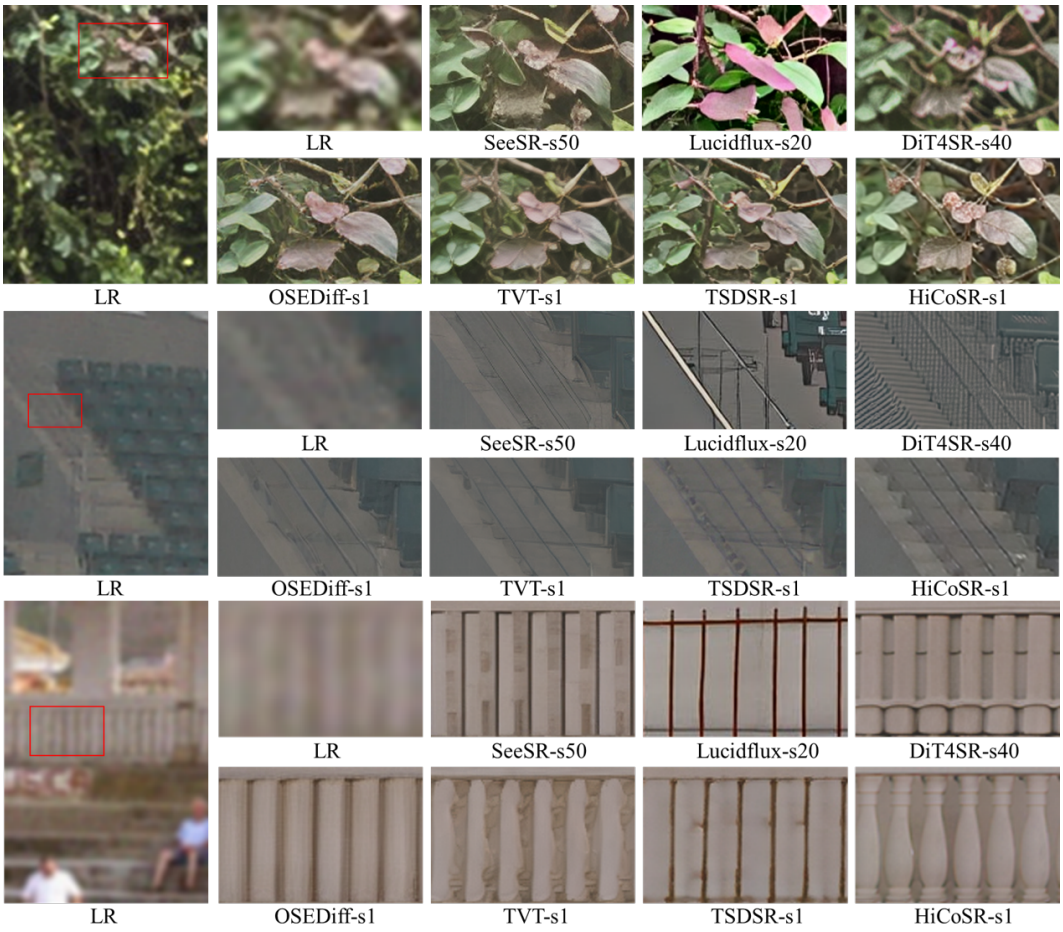

Figure 7: Additional visual comparisons with state-of-the-art RealSR methods.

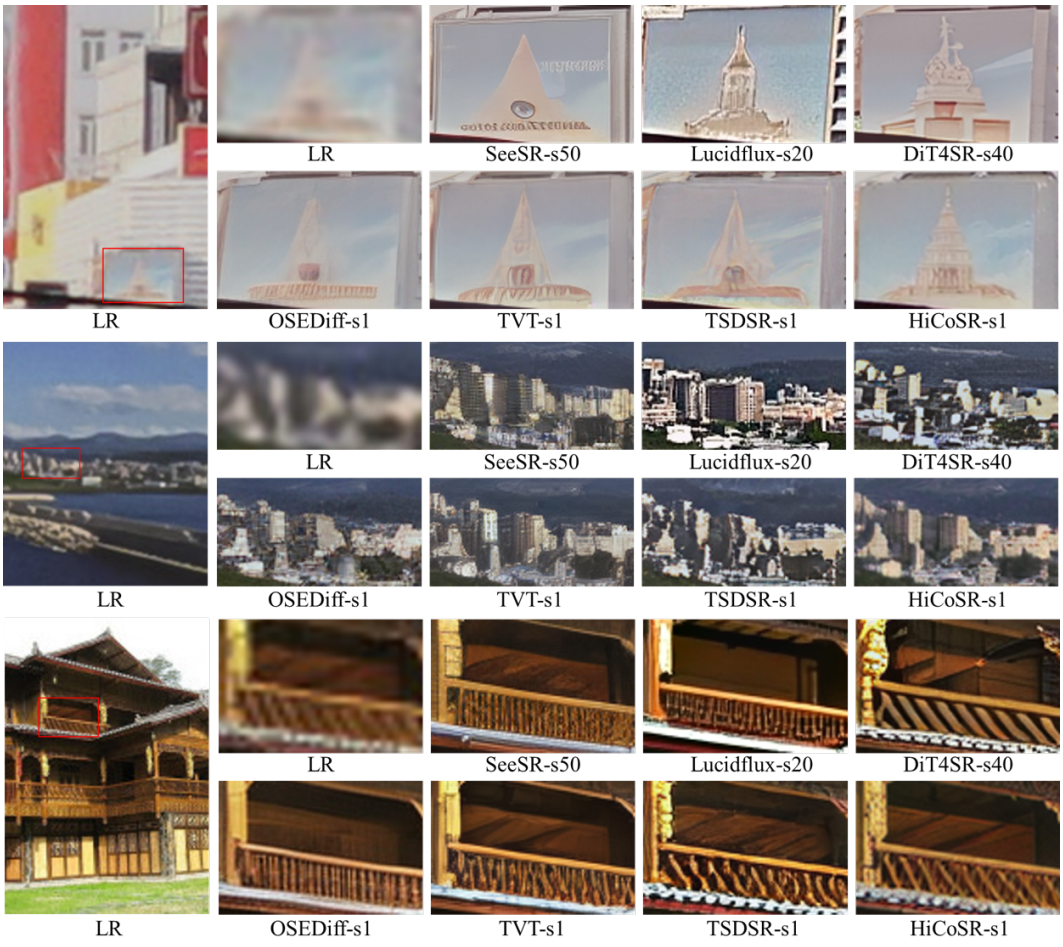

Figure 8: Additional visual comparisons with state-of-the-art RealSR methods.

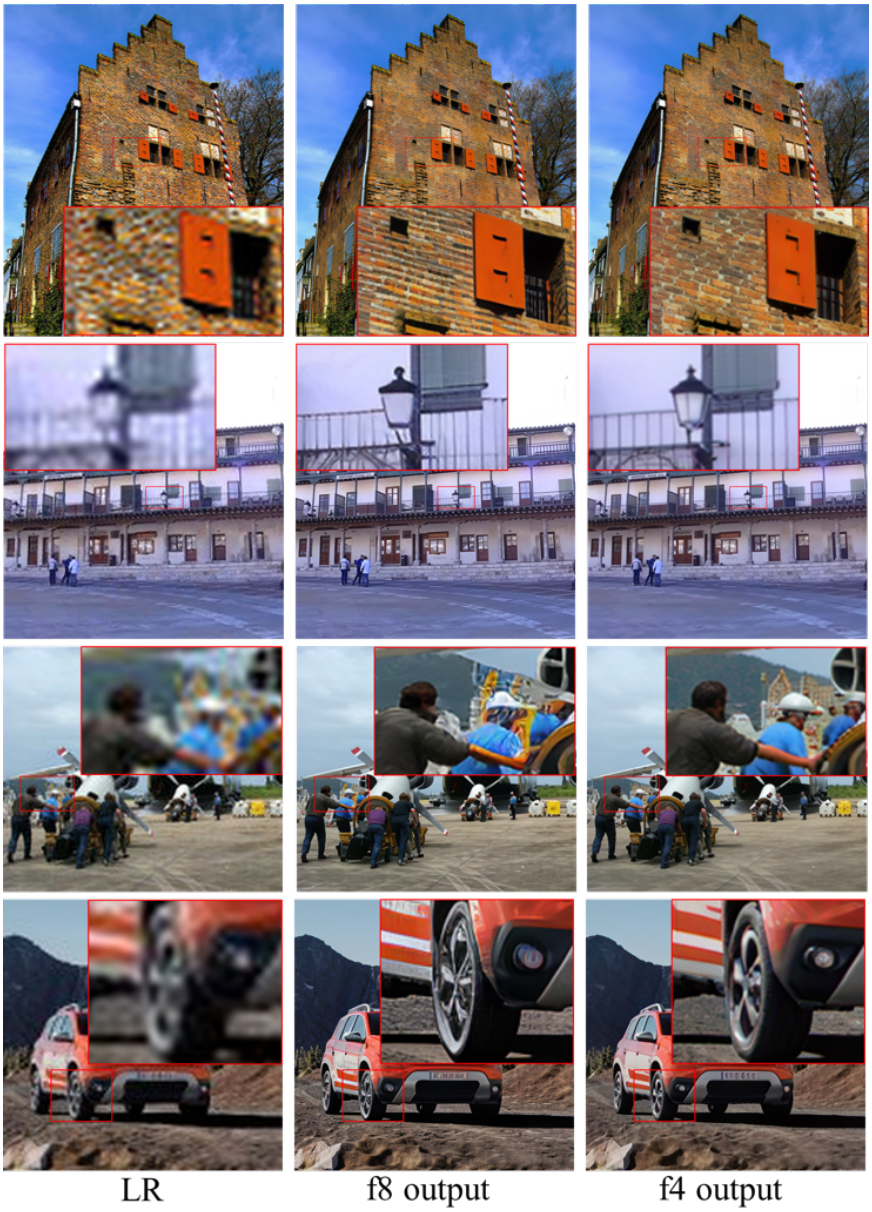

Figure 9: Visual demonstration of our coarse-to-fine cascaded generation process.