Realism Control One-step Diffusion for Real-World Image Super-Resolution

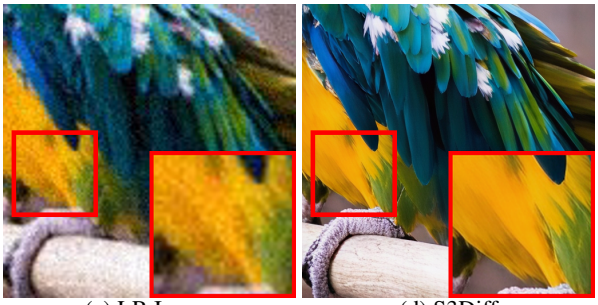
Zongliang Wu[1, 2, 3, *], **Siming Zheng** [3, *], **Peng-Tao Jiang**[3, †], **Xin Yuan**[2, †]

[1] Zhejiang University, Hangzhou, China
[2] School of Engineering, Westlake University, Hangzhou, China
[3] vivo Mobile Communication Co., Ltd
{wuzongliang, xyuan}@westlake.edu.cn, {zhengsiming, pt.jiang}@vivo.com

Abstract

Pre-trained diffusion models have shown great potential in real-world image super-resolution (Real-ISR) tasks by enabling high-resolution reconstructions. While one-step diffusion (OSD) methods significantly improve efficiency compared to traditional multi-step approaches, they still have limitations in balancing fidelity and realism across diverse scenarios. Since the OSDs for SR are usually trained or distilled by a single timestep, they lack flexible control mechanisms to adaptively prioritize these competing objectives, which are inherently manageable in multi-step methods through adjusting sampling steps. To address this challenge, we propose a Realism Controlled One-step Diffusion (RCOD) framework for Real-ISR. RCOD provides a latent domain grouping strategy that enables explicit control over fidelity-realism trade-offs during the noise prediction phase with minimal training paradigm modifications and original training data. A degradation-aware sampling strategy is also introduced to align distillation regularization with the grouping strategy and enhance the controlling of trade-offs. Moreover, a visual prompt injection module is used to replace conventional text prompts with degradation-aware visual tokens, enhancing both restoration accuracy and semantic consistency. Our method achieves superior fidelity and perceptual quality while maintaining computational efficiency. Extensive experiments demonstrate that RCOD outperforms state-of-the-art OSD methods quantitatively and visually, with flexible realism control capabilities in the inference stage.

Code — https://zongliang-wu.github.io/RCOD-SR
Supplementary Materials (SM) —
https://arxiv.org/abs/2509.10122



(a) LR Image ↑PSNR: 19.16, ↓NIQE:10.021
(d) S3Diff ↑PSNR: 25.47, ↓NIQE: 4.418
(e) Ours-Fid. ↑PSNR: **26.94**, ↓NIQE: 4.049
(f) Ours-Real. ↑PSNR: 25.66, ↓NIQE: **3.913**

Figure 1: While previous one-step diffusion methods, such as S3Diff (Zhang et al. 2024) only yield one optimal result (b), our approach offers the flexibility to control images (c-d) with different fidelity-realism trade-offs during inference, enhancing practical applicability across different scenarios.

Introduction

Image super-resolution (SR) (Dong et al. 2015; Zhang et al. 2018b, 2021; Ledig et al. 2017; Liang et al. 2021) aims to recover a high-resolution (HR) image from its low-resolution (LR) counterpart.

Traditional image super-resolution simplifies the degradation process as known noise, blur, or downsampling. In recent years, real-world image super-resolution (Real-ISR) (Zhang et al. 2021; Wang et al. 2021) has attracted more attention due to the increasing demand for reconstructing high-resolution images under real-world unknown degradations, which is more challenging and practical in real applications.

While recent advances in Stable Diffusion (SD) models (Ho, Jain, and Abbeel 2020; Song et al. 2020), especially the large-scale pretrained text-to-image (T2I) models have demonstrated unprecedented capabilities in various downstream vision tasks (Zhang, Rao, and Agrawala 2023; Rombach et al. 2022). Some works leveraging pre-trained SD models for multi-step SR, such as DiffBIR (Lin et al. 2023),

[*]These authors contributed equally.
[†]Corresponding author.

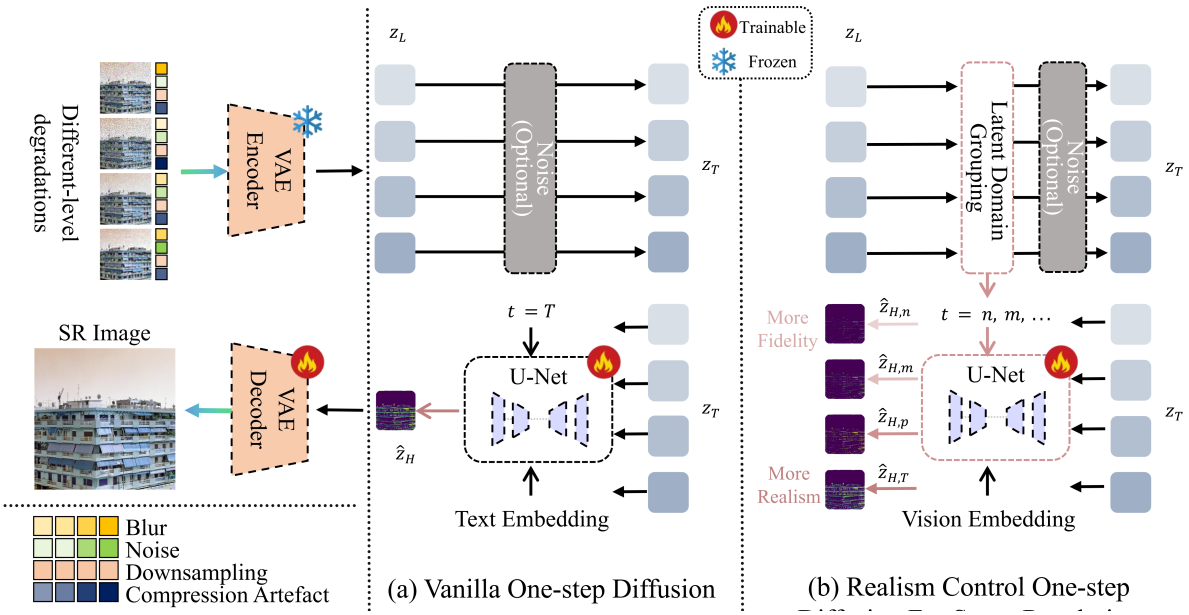


Figure 2: Realism control one-step diffusion (RCOD) training process. The left part illustrates several synthesized real-world LR images by applying diverse degradations with varying types and intensities on an HR image. (a) Existing vanilla one-step diffusion (OSD) methods for super-resolution (SR): These LR images are directly sent into the diffusion forward and reverse process; the denoising U-Net tends to learn to recover the 'average' degradation, leading to a monotonous generation ability within the latent domain. (b) Our proposed Realism Control One-Step Diffusion employs a latent domain grouping strategy. This allows for adaptive control of timesteps (denoising degrees) during the forward process according to the degradation degree in the latent domain. As a result, the denoising U-Net can acquire a more diverse generation capability based on the timestep.

StableSR (Wang et al. 2024a), and SeeSR (Wu et al. 2024b), have achieved remarkable SR quality through iterative latent space optimization. Though these methods achieve impressive perceptual quality, they suffer from computational inefficiency. The caused latency by multi-step sampling makes real-time applications impractical.

To address the efficiency concerns, recent attempts focus on one-step diffusion frameworks (Wu et al. 2024a; Zhang et al. 2024), which distill multi-step diffusion priors into single-step inference through knowledge distillation and achieve 10×to 100× speedup over previous multi-step diffusion-based SR methods. However, existing one-step diffusion approaches face a fundamental dilemma: the deterministic single-step generation inherently lacks the controllable fidelity-realism balance that multi-step methods achieve via step-wise noise scheduling. As illustrated in Fig. 1, previous one-step diffusion super-resolution (SR) methods, such as S3Diff (Zhang et al. 2024), can only generate a single optimal result but can not meet the need for dynamic adjustment between fidelity and realism. The root cause lies in current OSD training paradigms. Most methods align all the LR inputs under different unknown degradations with a single convergence space through single timestep conditioned training, which results in a balanced static preference for fidelity or realism and prevents adaptive adjustments for scenario-specific requirements.

Bearing the above concerns in mind, we propose a novel framework that provides one-step diffusion Real-ISR methods with the capability to monotonically control the level of realism. This framework, which we denote as **Realism Controlled One-step Diffusion (RCOD)**, can be easily integrated into existing one-step diffusion methods for Real-ISR. Specifically, during the training phase, we incorporate a Latent Domain Grouping (LDG) strategy into the latent diffusion process, grouping training data according to a latent domain metric. Through this strategy, the diffusion denoising network learns to perceive variations in degradation across training samples, thereby gaining adaptive restoration capabilities. Furthermore, to address the inherent limitations caused by text prompts, we introduce a Visual Prompt Injection Module (VPIM) to enhance prompt quality. Our contributions are summarized as follows:

- We propose a simple but effective latent domain grouping (LDG) strategy that reformulates the noise prediction process by partitioning the latent space into fidelity and realism oriented domains. This allows explicit control over detail preservation versus generative enhancement through simple training paradigm modifications, without requiring additional trainable parameters.
- During distillation, we introduce a degradation-aware sampling (DAS) strategy that reformulates timestep sampling in the pretrained model by adaptively aligning it with our LDG framework, enhancing controlling with regularization strength.
- To reduce the computational burden of VLM and dependencies on manual text prompts, we propose a visual

prompt injection module (VPIM) to replace text prompts with degradation-aware visual tokens, enhancing both restoration accuracy and semantic consistency.

- We empirically evaluate our approach on widely used stable diffusion-based and their distillation version Real-ISR methods, demonstrating quality improvement and the effectiveness of proposed approach.

Related Work

Real-World Image Super-Resolution

To address the challenge of recovering real-world low-resolution (LR) observations with unknown degradations, the Real-ISR task has been introduced. Due to the complex degradations involved, Real-ISR has been a challenging problem for some time (Ignatov et al. 2017; Liu et al. 2022; Ji et al. 2020). Initial methods trained their models with simple downsampling techniques (Kim, Kwon Lee, and Mu Lee 2016; Zhang et al. 2018b), which led to poor performance on real-world datasets. BSRGAN (Zhang et al. 2021) and Real-ESRGAN (Wang et al. 2021) were among the first to introduce a more realistic synthesis pipeline for LR images, enabling deep learning methods to be applied to real-world scenarios. However, these GAN-based methods often suffer from artifacts and training instability. With the emergence of the powerful pre-trained text-to-image generation model Stable Diffusion (SD) (Rombach et al. 2022), many efforts have been made to leverage its strong generative capabilities for solving the Real-ISR problem, such as DiffBIR (Lin et al. 2023), StableSR (Wang et al. 2024a) and SeeSR (Wu et al. 2024b), and FaithDiff (Chen, Pan, and Dong 2025). These SD-based methods improve fidelity and perceptual quality, but their application is limited due to the significant computational resources and time required. This is primarily due to the dozens to hundreds of timesteps involved in the diffusion denoising process.

One-step Diffusion Models for Real-ISR

To reduce the computational cost during inference, one-step diffusion methods have been proposed. These methods utilize model distillation techniques specifically designed for diffusion models, including progressive distillation (Meng et al. 2023; Salimans and Ho 2022), consistency models (Song et al. 2023), distribution matching distillation (Yin et al. 2024b,a), and variational score distillation (VSD) (Wang et al. 2024c; Nguyen and Tran 2024). In the context of Real-ISR, OSEDiff (Wu et al. 2024a) employs the VSD strategy to achieve one-step diffusion based on a pre-trained SD model. Similarly, S3Diff (Zhang et al. 2024) directly employs a distilled Stable Diffusion Turbo (SD-T) model for Real-ISR. Recent works including ADCSR (Chen et al. 2025) and TSD-SR (Dong et al. 2025) also improve OSD performance and efficiency. InvSR(Yue, Liao, and Loy 2025) offers a flexible sampling mechanism with arbitrary-steps (1-5) diffusion.

However, while time efficiency is improved, another dilemma arises: one-step diffusion methods struggle to generate diverse outputs balancing fidelity and realism—a critical requirement for real-world applications—unlike multi-step approaches that achieve this through progressive noise scheduling. In other words, current one-step diffusion methods for Real-ISR cannot control fidelity and realism as effectively as their multi-step counterparts. Some techniques, such as ControlNet models (Zhang, Rao, and Agrawala 2023), can enhance HR image generation by controlling semantic content but cannot achieve pixel-level control and require additional conditions, such as extra images or depth maps. Therefore, these techniques cannot be directly employed in one-step diffusion methods for Real-ISR. OFTSR (Zhu et al. 2024) achieves fidelity trade-offs through trajectory alignment distillation, but lacks degradation-aware mechanisms for real-world scenarios. PiSA-SR (Sun et al. 2025) controls fidelity and realism by training and inference with two different LoRA (Hu et al. 2022) modules and two-step diffusion, which obviously increases the computational cost compared to using a single step. Thus, a simple, efficient, and generalizable method for fidelity-realism control in OSD for Real-ISR remains essential.

Methodology

In this section, we first reveal the characteristic of denoising network in one-step diffusion for image super-resolution, and then propose our algorithm.

Preliminary: The Character of Denoising Network in One-step Diffusion for Real-ISR

In SD, the latent diffusion process starts with encoding an image into a latent representation z_0 using a VAE encoder. The forward diffusion process then adds Gaussian noise to z_0 over T steps via a Markov chain defined as:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{1-\beta_t}\boldsymbol{z}_{t-1}, \beta_t\mathcal{I}) \tag{1}$$

following a variance schedule $\beta_1, \ldots, \beta_T$. Here $\boldsymbol{z}_0 \sim q(\boldsymbol{z}_0)$. The forward process is: $\boldsymbol{z}_t = \alpha_t\boldsymbol{z}_0 + \sigma_t\boldsymbol{\epsilon}$, where $\alpha_t = \prod_{s=1}^{t}\sqrt{1-\beta_s}$, $\sigma_t = \sqrt{1-\alpha_t^2}$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathcal{I})$. Here, the mean of $\boldsymbol{z}_t$ becomes $\alpha_t\boldsymbol{z}_0$, and the variance is $\sigma_t^2\mathcal{I}$. With larger t, more noise is added, resulting in a greater deviation of the mean (scaled by $\alpha_t < 1$) and an increased variance (σ_t^2), distancing $\boldsymbol{z}_t$ from the original distribution $\boldsymbol{z}_0$.

The reverse process denoises $\boldsymbol{z}_t$ to recover $\boldsymbol{z}_{t-1}$:

$$p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{z}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{z}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{z}_t, t)), \tag{2}$$

where a time-conditional U-Net ϵ_θ predicts the noise $\boldsymbol{\epsilon}$ to estimate $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$. Multi-step diffusion models iteratively refine $\boldsymbol{z}_T$ back to $\boldsymbol{z}_0$ over T steps, while one-step diffusion methods, often distilled from multi-step models, predict the clean data directly from $\boldsymbol{z}_T$ in a single step, significantly reducing computation.

For Real-ISR tasks using SD-based OSD methods like (Wu et al. 2024a), the process from LR latent features $\boldsymbol{z}_L$ to HR latent features $\hat{\boldsymbol{z}}_H$ is formulated as a one-step denoising process:

$$\hat{\boldsymbol{z}}_H = F_\theta(\boldsymbol{z}_L; T, c_y) \triangleq \frac{\boldsymbol{z}_L - \beta_T\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_L; T, c_y)}{\alpha_T}, \tag{3}$$

where $\boldsymbol{z}_L$ is the LR latent representation, c_y is the text embedding, and ϵ_θ is the denoising network predicting noise

at timestep T. As shown in Fig. 2(a), vanilla OSD methods learn a direct latent feature mapping from LR to HR images with or without adding additional noise.

In Real-ISR, images exhibit diverse degradation types and levels. SD-based methods use powerful generative image priors to recover LR images. However, OSD typically trains on all data with a fixed timestep T with a constant noise level. This results in a model that generates a uniform amount of detail and converges to a confined domain, which may not suit images with varying degradation severity. Tab. 1 (a) demonstrates the results of training OSEDiff on two different degradation pipelines (DP). 'Orig.+ New Deg.' denotes a DP applying more degradations than the standard DP ('Orig.'). It indicates that higher degradation in training results in a greater emphasis on realism. This exposes OSD's flaw: by locking training to a single fixed T, the model optimizes for an **"average" degradation**, yielding a limited generation flexibility that struggles to adapt its output to meet the specific scenario requirements.

In contrast, multi-step diffusion models provide greater flexibility in SR tasks. By selecting different timesteps t during the inference stage, these models control the degree of noise adding and removing in diffusion, balancing fidelity and realism effectively.

Therefore, to overcome the inherent limitation of OSD, which is incapable of adjusting its generation levels to adapt to varying scenarios, we propose a novel training strategy for real-world SR applications that can flexibly control the generation realism during the inference stage.

Metrics	Orig.	Orig.+New Deg.
PSNR ↑	**25.15**	24.59
MANIQA ↑	0.6326	**0.6462**

(a)

M_L	L1	MSE	**Cosine Sim.**
SSIM ↑	0.60	0.54	**0.78**
DISTS ↑	0.15	0.11	**0.42**
CLIPIQA ↑	0.06	0.05	**0.27**

(b)

Table 1: (a) Influence of degradation degree in training data. (b) |Spearman coefficient| ↑ comparison of different M_L distances and image quality metrics.

Latent Domain Grouping

To achieve dynamic fidelity-realism trade-offs control in OSD generated SR results, we focus on the most basic condition in denoising network, *i.e*, timestep condition. Unlike prompts from a text encoder or a vision encoder, the timestep condition is an unremovable component in the diffusion process. At the same time, it controls the mean and variance of noisy latent feature z_T. With the larger mean and variance difference between z_T and z_0, more contents will be generated during the denoising process. Fig. 3 shows an example

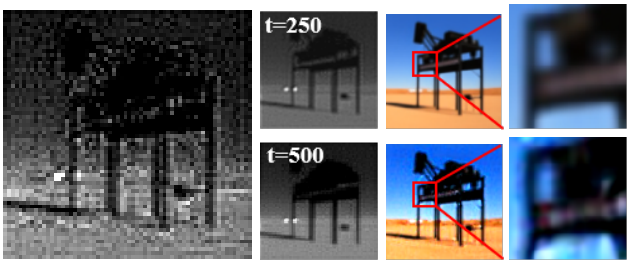


Figure 3: Influence of different timesteps t using SD-turbo.

of influence of different t. In the foundational model SD-turbo, the higher timestep value during the diffusion process usually reflects the the higher capability generating.

Therefore, to easily control the generation degree, we propose a latent domain grouping (LDG) strategy. Recall Eq. (3), we do not use a single fixed timestep T, but choose a timestep t according to a metrics:

$$\hat{\boldsymbol{z}}_H = F_\theta(\boldsymbol{z}_L; t, c_y), \tag{4}$$

$$t = k \cdot (n - \left\lfloor \frac{n \cdot (M_L - M_{L\text{-min}})}{(M_{L\text{-max}} - M_{L\text{-min}})} \right\rfloor),\ k \in \mathbb{Z}^+, \tag{5}$$

where M_L denotes a latent metric that can perceive the "level of degradation" of features in latent domain, $M_{L\text{-min}}$ is the minimum value of M_L in training data, k is interval of timestep, $\lfloor . \rfloor$ denotes the maximum integer no larger than the entry inside, n is number of groups for timestep.

To employ this strategy both on SD and distillation version of SD, *i.e*, SD-Turbo (SDT), which distilled a four specific steps from the original 1000-step diffusion process, we set n to be ≤ 4 and $k = 250$.

By the grouping strategy, denoising network can learn different degrees of generation according to timestep. In the training stage, grouping is based on the M_L described in the next subsection. In the inference stage, we can easily choose a timestep to control the level of realism for SR in different scenarios. Furthermore, due to our grouping strategy, the realism level increases monotonically with the timestep.

Latent Metric for Denoising Network

The latent metric M_L is designed to enable the denoising network in the latent domain to perceive the "level of degradation" of the low-resolution features $\mathbf{z}_L$. Here, we define the "level of degradation" as the extent to which $\mathbf{z}_L$ deviates from its HR counterpart $\mathbf{z}_H$. Thus, the definition of the latent metric M_L should reflect the characteristics of $\mathbf{z}_L$ that indicate this degradation.

A simple choice might be to use the difference between the mean values of $\mathbf{z}_L$ and $\mathbf{z}_H$, as the forward diffusion process scales the mean of $\mathbf{z}_L$ over time. However, as shown in Fig. 4(a), the mean values of $\mathbf{z}_L$ and $\mathbf{z}_H$ are similar across the training data. This suggests that mean difference fails to effectively capture the degradation level.

Instead, we use cosine similarity (CS) as M_L:

$$M_L = \frac{\mathbf{z}_L \cdot \mathbf{z}_H}{\|\mathbf{z}_L\| \|\mathbf{z}_H\|}. \tag{6}$$

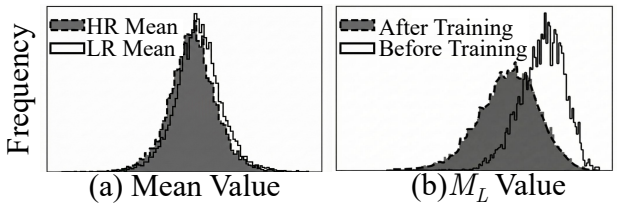


Figure 4: Distribution of (a) mean value of LR and HR training images in the latent domain, (b) the M_L metric in the latent domain of VAE before and after training.

Cosine similarity is a widely adopted measure in representation learning (e.g., contrastive learning (Chen et al. 2020)). It can quantify the divergence between high-dimensional latent features $\mathbf{z}_L$ and $\mathbf{z}_H$, which can reflect high-level changes caused by degradation. As shown in Fig. 4(b), most cosine similarity values lie between 0 and 1, providing a clear range to distinguish varying degrees of degradation for the denoising network.

Different distances inherently introduce a certain preference or bias, to evaluate the bias of different M_L, we calculate correlation coefficient (|Spearman Corr.| $\uparrow$) of CS, L1, and MSE with different metrics in Table 1 (b). The metrics are calculated between LR and HR images. The correlation reflects the degree of association between various M_L and different metrics. CS exhibits a higher correlation coefficient with objective (SSIM), perceptual (DISTS), and semantic (CLIPIQA) metrics compared to other distances in latent space. **More visualization results can be seen in Supplementary Materials (SM).**

Metric Estimation

Since $\mathbf{z}_H$ is introduced as a latent metric, we can only calculate it during training. In the inference stage, beyond manually choosing a timestep, we also consider an adaptive timestep selection, i.e., estimating an M_L for each LR image. Benefiting from the powerful representation ability of the pre-trained model, we use features from intermediate layers as input and a simple MLP as a metric estimation module (MEM), operating independently of OSD training.

Degradation-aware Sampling Distillation

Previous VSD method for SR (Wu et al. 2024a), utilizes the regularization network (a pre-trained SD) that sampled timesteps across a wide range (20-980). This aims to generate regularization latent features and, in turn, optimize the distribution of the OSD network. However, to better integrate the distillation process with the concept of degradation in the latent space, we propose a Degradation-Aware Sampling (DAS) strategy. DAS redefines how timesteps are sampled in the pre-trained model, adaptively aligning this process with our LDG framework to provide explicit control over regularization strength. The DAS can be written as:

$$t_r = S(\max(20, t-k), \min(980, t+k)) \tag{7}$$

where t_r is the sample timestep for regularization network, t is the chosen timestep in OSD network by LDG, $S(t_{min}, t_{max})$ denotes uniformly random sampling of an integer from the range $[t_{min}, t_{max}]$.

By applying DAS, the degradation grouping information is delivered from LDG, thereby aligning this process with LDG and control over regularization strength.

Visual Prompt Injection Module

In previous SD-based super-resolution (SR) methods like OSEDiff (Wu et al. 2024a) and SeeSR (Wu et al. 2024b), text encoders, sometimes paired with a vision-language model (VLM) as a text prompt extractor, improve non-reference (NR) metrics, which assess realism, but often constrain full-reference (FR) metrics, which assess fidelity to the ground truth. This trade-off arises because text prompts provide high-level semantic guidance to enhance realism, yet they may compromise structural accuracy.

LR feature $\mathbf{z}_L$ solely from the VAE encoder offers limited semantic information. Without additional context, the conditioned U-Net struggles to generate high-quality outputs. Text prompts attempt to bridge this gap by injecting external semantic cues, but they come with drawbacks: VLMs increase computational costs, and the prompts may not fully align with the image's content. Some methods like S3Diff (Zhang et al. 2024) use fixed text to reduce complexity, yet still struggle to balance NR and FR metrics.

To address these issues, we propose the Visual Prompt Injection Module (VPIM), which replaces conventional text prompts with degradation-aware visual tokens. VPIM substitutes the text encoder (typically a CLIP text model) with a CLIP vision model and an MLP layer for dimension alignment. The LR image serves as its input, *i.e*, visual prompt, and the output is fed into the cross-attention of the U-Net. By adopting VPIM, we eliminate the need for VLMs, reduce computational overhead, and provide the U-Net with image-specific semantic information directly from the LR input. The visual prompt is tied to the image's pixel characteristics, leading to improvement in both fidelity and realism.

With the combination of LDG, latent metric, DAS, and VPIM, we proposed a Realism control one-step diffusion (RCOD) framework, a realism-flexible one-step diffusion model with enhanced performance that can be applied to various recent mainstream one-step diffusion Real-ISR methods, thereby improving their capabilities. **We provide a pseudo-code example of our RCOD in SM.**

Experiments

Experiments Setups

Datasets: Following the training and testing settings of prior works(Wu et al. 2024b,a; Zhang et al. 2024), we employ LSDIR (Li et al. 2023) and the first 10K face images from FFHQ (Karras, Laine, and Aila 2019) for training and degradation pipeline of Real-ESRGAN (Wang et al. 2021) for LR image synthesizing. The synthetic test data use cropped 512×512 synthetic data from DIV2K-Val (Agustsson and Timofte 2017) and degraded using the Real-ESRGAN pipeline (Wang et al. 2021). The real-world data include LR-HR pairs from RealSR (Cai et al. 2019) and

Datasets	Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	FID↓
DrealSR	SinSR	28.36	0.7515	0.3665	0.2485	6.991	55.33	0.4884	0.6383	170.57
	PiSA-SR	28.31	0.7804	0.2960	0.2169	6.200	66.11	0.6156	0.6970	130.61
	TSD-SR	25.67	0.7132	0.3538	0.2459	5.991	65.99	0.6327	0.7137	171.36
	InvSR	28.33	0.7502	0.3678	0.2481	6.941	55.27	0.4900	0.6385	170.57
	OSEDiff	27.92	0.7835	0.2968	0.2165	6.490	64.65	0.5899	0.6963	135.30
	$RCOD_O$-Fid.	**28.90**	**0.7906**	**0.2919**	0.2186	6.817	66.72	0.6275	0.7023	131.52
	$RCOD_O$-Neu.	28.30	0.7775	0.3080	0.2306	6.469	68.03	**0.6385**	0.7179	139.92
	$RCOD_O$-Real.	27.59	0.7600	0.3389	0.2499	6.172	**68.19**	0.6295	0.7325	158.16
	S3Diff	27.54	0.7491	0.3109	**0.2099**	6.212	63.94	0.6134	0.7130	**118.64**
	$RCOD_S$-Fid.	28.09	0.7800	0.3002	0.2149	5.871	65.74	0.6174	0.6963	136.62
	$RCOD_S$-Neu.	27.48	0.7526	0.3256	0.2251	5.636	67.40	0.6339	0.7278	138.87
	$RCOD_S$-Real.	26.95	0.7190	0.3607	0.2414	**5.448**	67.58	0.6317	**0.7478**	145.36
	$RCOD_S$-Adap.	27.83	0.7661	0.3098	0.2181	5.768	66.32	0.6223	0.7110	135.61
RealSR	SinSR	**26.28**	0.7347	0.3188	0.2353	6.287	60.80	0.5385	0.6122	135.93
	PiSA-SR	25.50	0.7417	0.2672	0.2044	5.500	70.15	0.6560	0.6702	124.09
	TSD-SR	23.41	0.6886	0.2805	0.2183	5.093	**70.77**	0.6311	**0.7193**	114.56
	InvSR	24.13	0.7125	0.2871	0.2123	5.626	68.54	0.6619	0.6790	138.88
	OSEDiff	25.15	0.7341	0.2921	0.2128	5.648	69.09	0.6326	0.6693	123.49
	$RCOD_O$-Fid.	26.01	**0.7427**	0.2796	0.2103	5.911	70.25	0.6647	0.6866	121.40
	$RCOD_O$-Neu.	25.39	0.7264	0.2939	0.2190	5.497	70.34	0.6750	0.7022	127.74
	$RCOD_O$-Real.	24.62	0.7011	0.3296	0.2375	5.341	70.76	0.6650	0.7084	143.74
	S3Diff	25.18	0.7269	0.2721	0.2005	5.269	67.82	0.6437	0.6727	**105.14**
	$RCOD_S$-Fid.	25.42	0.7392	**0.2647**	**0.1976**	5.095	69.46	0.6605	0.6509	113.93
	$RCOD_S$-Neu.	24.78	0.7130	0.2855	0.2073	5.024	70.55	**0.6757**	0.6886	114.80
	$RCOD_S$-Real.	24.08	0.6759	0.3228	0.2236	**4.900**	70.65	0.6719	0.7086	120.00
	$RCOD_S$-Adap.	25.23	0.7313	0.2714	0.2010	5.033	69.72	0.6646	0.6622	112.93

Table 2: Quantitative comparison with state-of-the-art methods on both synthetic and real-world benchmarks. The best and second best results are highlighted in **bold** and underline, respectively.

DRealSR (Wei et al. 2020), both with sizes of 128×128-512×512 for LR-HR pairs.

Evaluation Metrics: We employ widely used FR and NR metrics. FR metrics include PSNR, SSIM (Wang et al. 2004), LPIPS (Zhang et al. 2018a), and DISTS (Ding et al. 2020). NR metrics include NIQE (Zhang, Zhang, and Bovik 2015), MANIQA-pipal (Yang et al. 2022), MUSIQ (Ke et al. 2021), and CLIPIQA (Wang, Chan, and Loy 2023).

Method Comparison: We compare our method with state-of-the-art methods (SOTA) in three categories: multi-step diffusion Real-ISR methods, including StableSR (Wang et al. 2024a), ResShift (Yue, Wang, and Loy 2023), DiffBIR (Lin et al. 2023), and SeeSR (Wu et al. 2024b); one-step diffusion Real-ISR methods, such as SinSR (Wang et al. 2024b), OSEDiff (Wu et al. 2024a), S3Diff (Zhang et al. 2024), TSD-SR (Dong et al. 2025), PiSA-SR (default version) (Sun et al. 2025), and InvSR (Yue, Liao, and Loy 2025); and GAN-based methods, including BSRGAN (Zhang et al. 2021) and Real-ESRGAN (Wang et al. 2021). We quantitatively compare recent SOTA OSD methods in Table 2 on real-world data. **More qualitative comparisons and the full table including synthetic data, multi-step diffusion, and GAN-based methods are in SM.**

Implement Details

Model Setting: To verify our framework, we select two types of recent SOTA OSD Real-ISR methods: OSEDiff (Wu et al. 2024a), which is distilled from a pre-trained multi-step Stable Diffusion (SD) model, and S3Diff (Zhang et al. 2024), which directly uses SD-T (a distilled version of SD), as our base models. When RCOD is applied to them, they are named $RCOD_O$ and $RCOD_S$, respectively. The choice of $n = 3$ corresponds to three distinct generation levels in the inference stage: Fidelity, Neutral, and Realism. *i.e*, $t = 250,\ 500,$ and 750 during inference. Since the S3Diff is not directly distilled from a multi-step diffusion, we do not apply DAS on it. $RCOD_S$-Adap. employs MEM during inference. The input to the MEM in this case is $\boldsymbol{z}_L$ and the features in the last layer degradation estimation model used in (Zhang et al. 2024). The MEM is trained after $RCOD_S$ training, utilizing the same training data. **More details please refer to SM.**

Comparison with State-of-the-Arts

Quantitative Comparisons: We can observe in Table 2: i) By applying RCOD, on each dataset, the "-Fid." versions ($t = 250$) have better full-reference (FR) metrics such as PSNR, SSIM, LPIPS, and FID, while keeping the no-reference (NR) metrics relatively ordinary. In contrast, the "-Real." versions ($t = 750$) achieve obviously higher NR metrics like MANIQA, MUSIQ, and CLIPIQA. Most "-Neu." versions ($t = 500$) fall within the middle range of the previous two versions. This illustrates that we can effectively and simply control the realism level (usually measured by perceptual NR metrics) during the inference stage, and that the realism level increases monotonically with the timestep.

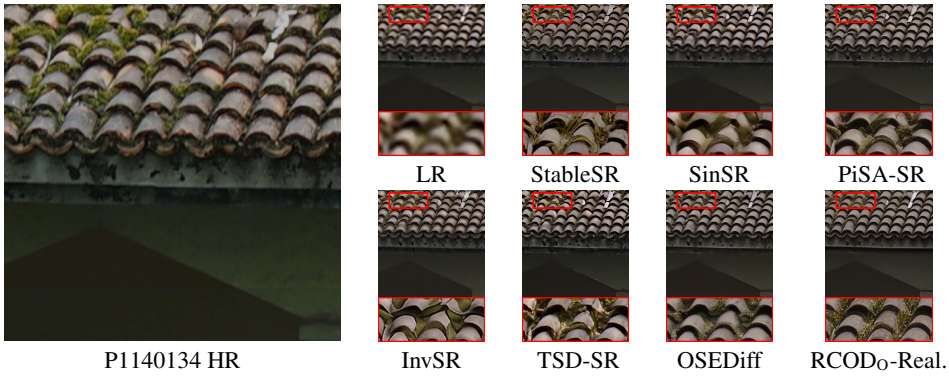


Figure 5: Visual comparison ($\times$4) of RCOD$_O$-Real. with other methods on DRealSR data.

ii) RCOD$_S$-Adap. has relatively balanced metrics between RCOD$_S$-Fid. and RCOD$_S$-Neu.. This indicates that most estimated M_L values are closer to 1 than to 0. This roughly matches the cosine similarity value distributions in Fig. 4 (b). iii) When applying RCOD, RCOD$_O$-Fid. and RCOD$_S$-Adap. perform better than their original methods (OSEDiff and S3Diff, respectively) on most metrics, including PSNR, SSIM, LPIPS, MANIQA, and MUSIQ. Even with a preference for fidelity in FR metrics, RCOD$_O$-Fid. shows superior performance on some NR metrics, such as MANIQA and MUSIQ, compared to the original OSEDiff method on real-world data. Additionally, RCOD$_S$-Real. usually achieves the best NR metrics (NIQE and CLIPIQA). iv) S3Diff shows better performance on the perceptual quality metric DISTS. This may arise from the negative online prompting (NOP) used in training, which provides a more accurate concept of high quality. However, since the text encoder and text prompt are replaced by VPIM in our RCOD$_S$, the NOP is also removed. Despite this, our RCOD$_S$-Adap. performs better on other NR metrics.

Time Efficiency: In Table 3, we compare inference time and trainable parameters. All methods are tested on an A100 GPU with a 512×512 input image. RCOD keeps similar time efficiency and trainable parameters as the original methods while have higher PSNR and MANIQA. RCOD$_O$ even inferences faster as the text extractor is removed.

Qualitative Comparisons: Fig. 5 compares the visual qualities of different Real-ISR methods. RCOD$_O$-Real. demonstrates its ability to recover more detailed and natural textures. Fig. 6 illustrates the changes in visual effects as t increases, *i.e*, from RCOD$_S$-Fid. ($t = 250$) to RCOD$_S$-Real. ($t = 750$). As t increases during the inference stage, more skin texture and wrinkles are recovered. RCOD$_S$-Adap. chooses a proper $t = 500$ in this case, where the M_L range is $[0.5, 0.75)$ according to Eq. 5. The M_L of the LR image (0.621) falls within this range.

Ablation Study We performed a series of ablation studies of the RCOD framework, the details can be found in SM.

Conclusion

We propose RCOD, a framework that enhances one-step diffusion methods for Real-ISR through flexible realism control. RCOD employs latent grouping with degradation-aware sampling during distillation and introduces a robust latent metric enabling denoising networks to assess degradation levels. Applied to two distinct one-step diffusion methods, RCOD achieves superior super-resolution performance across most FR and NR metrics while maintaining computational efficiency. We believe RCOD holds promise for diverse Real-ISR scenarios with varying requirements.

Figure 6: Visual comparison ($\times$4) of realism controls during RCOD$_S$ inference stage. The best results are in **bold**.

Metrics	PiSA-SR-adj.	OSEDiff	**RCOD$_O$-Fid.**	S3Diff	**RCOD$_S$-Adap.**
PSNR ↑	28.31	27.92	**28.90**	27.54	27.83
MANIQA ↑	0.6156	0.5899	**0.6275**	0.6134	0.6223
Infer. Time (s)	0.13	0.11	**0.09**	0.28	0.28
Trainable Param. (M)	**8.1**	8.5	9.5	34.5	35.5

Table 3: Efficiency comparison on an NVIDIA A100 GPU. The best results are highlighted in bold.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2024YFF0505603, National Natural Science Foundation of China under Grant 62271414, Zhejiang Provincial Science Fund for Distinguished Young Scholar Project under Grant LR23F010001, Zhejiang "Pioneer" and "Leading Goose" R&D Program under Grant 2024SDXHDX0006 and 2024C03182, the Key Project of Westlake Institute for Optoelectronics under Grant 2023GD007, and Ningbo Science and Technology Bureau, "Science and Technology Yongjiang 2035" Key Technology Breakthrough Program under Grant 2024Z126.

References

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 126–135.

Cai, J.; Zeng, H.; Yong, H.; Cao, Z.; and Zhang, L. 2019. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, 3086–3095.

Chen, B.; Li, G.; Wu, R.; Zhang, X.; Chen, J.; Zhang, J.; and Zhang, L. 2025. Adversarial diffusion compression for real-world image super-resolution. In *CVPR*, 28208–28220.

Chen, J.; Pan, J.; and Dong, J. 2025. Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In *CVPR*, 28188–28197.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607.

Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2020. Image quality assessment: Unifying structure and texture similarity. *IEEE PAMI*, 44(5): 2567–2581.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2015. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2): 295–307.

Dong, L.; Fan, Q.; Guo, Y.; Wang, Z.; Zhang, Q.; Chen, J.; Luo, Y.; and Zou, C. 2025. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *CVPR*, 23174–23184.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Ignatov, A.; Kobyshev, N.; Timofte, R.; Vanhoey, K.; and Van Gool, L. 2017. Dslr-quality photos on mobile devices with deep convolutional networks. In *ICCV*.

Ji, X.; Cao, Y.; Tai, Y.; Wang, C.; Li, J.; and Huang, F. 2020. Real-world super-resolution via kernel estimation and noise injection. In *CVPRW*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *ICCV*.

Kim, J.; Kwon Lee, J.; and Mu Lee, K. 2016. Accurate image super-resolution using very deep convolutional networks. In *CVPR*.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 4681–4690.

Li, Y.; Zhang, K.; Liang, J.; Cao, J.; Liu, C.; Gong, R.; Zhang, Y.; Tang, H.; Liu, Y.; Demandolx, D.; et al. 2023. Lsdir: A large scale dataset for image restoration. In *CVPR*, 1775–1787.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *ICCV*.

Lin, X.; He, J.; Chen, Z.; Lyu, Z.; Fei, B.; Dai, B.; Ouyang, W.; Qiao, Y.; and Dong, C. 2023. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*.

Liu, A.; Liu, Y.; Gu, J.; Qiao, Y.; and Dong, C. 2022. Blind image super-resolution: A survey and beyond. *IEEE PAMI*, 45(5): 5461–5480.

Meng, C.; Rombach, R.; Gao, R.; Kingma, D.; Ermon, S.; Ho, J.; and Salimans, T. 2023. On distillation of guided diffusion models. In *CVPR*, 14297–14306.

Nguyen, T. H.; and Tran, A. 2024. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. In *CVPR*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Salimans, T.; and Ho, J. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.

Song, Y.; Dhariwal, P.; Chen, M.; and Sutskever, I. 2023. Consistency models. In *ICML*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Sun, L.; Wu, R.; Ma, Z.; Liu, S.; Yi, Q.; and Zhang, L. 2025. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2333–2343.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *AAAI*, volume 37, 2555–2563.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2024a. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 1–21.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*.

Wang, Y.; Yang, W.; Chen, X.; Wang, Y.; Guo, L.; Chau, L.-P.; Liu, Z.; Qiao, Y.; Kot, A. C.; and Wen, B. 2024b. SinSR:

Diffusion-Based Image Super-Resolution in a Single Step. In *CVPR*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4): 600–612.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2024c. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*.

Wei, P.; Xie, Z.; Lu, H.; Zhan, Z.; Ye, Q.; Zuo, W.; and Lin, L. 2020. Component divide-and-conquer for real-world image super-resolution. In *ECCV*, 101–117. Springer.

Wu, R.; Sun, L.; Ma, Z.; and Zhang, L. 2024a. One-step effective diffusion network for real-world image super-resolution. *NeurIPS*, 37: 92529–92553.

Wu, R.; Yang, T.; Sun, L.; Zhang, Z.; Li, S.; and Zhang, L. 2024b. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*.

Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; and Yang, Y. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, 1191–1200.

Yin, T.; Gharbi, M.; Park, T.; Zhang, R.; Shechtman, E.; Durand, F.; and Freeman, W. T. 2024a. Improved Distribution Matching Distillation for Fast Image Synthesis. *arXiv preprint arXiv:2405.14867*.

Yin, T.; Gharbi, M.; Zhang, R.; Shechtman, E.; Durand, F.; Freeman, W. T.; and Park, T. 2024b. One-step diffusion with distribution matching distillation. In *CVPR*, 6613–6623.

Yue, Z.; Liao, K.; and Loy, C. C. 2025. Arbitrary-steps image super-resolution via diffusion inversion. In *CVPR*, 23153–23163.

Yue, Z.; Wang, J.; and Loy, C. C. 2023. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *NeurIPS*.

Zhang, A.; Yue, Z.; Pei, R.; Ren, W.; and Cao, X. 2024. Degradation-guided one-step image super-resolution with diffusion priors. *arXiv preprint arXiv:2409.17058*.

Zhang, K.; Liang, J.; Van Gool, L.; and Timofte, R. 2021. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, 4791–4800.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*, 3836–3847.

Zhang, L.; Zhang, L.; and Bovik, A. C. 2015. A feature-enriched completely blind image quality evaluator. *IEEE TIP*, 24(8): 2579–2591.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018a. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 586–595.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018b. Image super-resolution using very deep residual channel attention networks. In *ECCV*.

Zhu, Y.; Wang, R.; Lu, S.; Li, J.; Yan, H.; and Zhang, K. 2024. Oftsr: One-step flow for image super-resolution with tunable fidelity-realism trade-offs. *arXiv preprint arXiv:2412.09465*.