DEMEMSR: TRANSFERABLE DEGRADATION REPRESENTATION LEARNING AND STABLE–DYNAMIC LR MEMORY FUSION FOR DIFFUSION-BASED REAL-WORLD IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based real-world image super-resolution can benefit from both degradation-aware conditioning and adaptive interaction between low-resolution (LR) observations and evolving noisy latents. However, contrastive degradation learning may distinguish different degradation configurations using only a subset of degradation cues, without adequately retaining the degradation information relevant to restoration. Moreover, although bidirectional LR–latent interaction allows LR guidance to adapt to the evolving denoising state, repeatedly updating the LR stream may cause it to drift from the original observation, weakening its role as a faithful structural anchor. To address these issues, we present DeMemSR, a degradation-aware diffusion Transformer framework that combines Transferable Degradation Representation Learning (TDRL) with Stable–Dynamic LR Memory Fusion (SD-LRMF). TDRL combines contrastive degradation learning with cross-content re-degradation. While contrastive learning encourages the projected features to distinguish degradation configurations, re-degradation encourages the encoder to retain degradation information beyond what is needed for discrimination. A degradation-aware conditioner then uses the projected features to provide layer-specific guidance to the diffusion Transformer. SD-LRMF explicitly separates LR guidance into an immutable Stable LR Memory derived from the initial observation and an interaction-updated Dynamic LR State. Learnable layer-wise fusion re-anchors the dynamic state to the stable memory across network depth, preserving observation fidelity while retaining the flexibility of bidirectional interaction. Experiments on four real-world benchmarks show that DeMemSR achieves leading no-reference perceptual quality, including the highest MUSIQ and LIQE scores among the evaluated methods on all four datasets.

1 INTRODUCTION

Real-world image super-resolution (SR) reconstructs high-resolution (HR) images from low-resolution (LR) images with unknown degradations, such as noise, blur and compression. In recent years, Diffusion models have made significant progress. Diffusion models provide strong generative priors for recovering plausible details, but effective restoration also requires informative degradation representations and reliable guidance from the input LR image. This motivates learning degradation representations that transfer across image content and preserving the initial LR representation as a stable reference across network depth.

Learning transferable degradation representations is challenging because image content and degradation are entangled in LR images. Contrastive degradation methods such as DASR (Wang et al., 2021a) learn representations that distinguish degradation configurations, while DRAT (Chen et al., 2025) and CdCL (Yuan et al., 2025) reduce interference from image content. However, reducing content interference and increasing degradation discriminability do not necessarily ensure that the learned representation captures degradation characteristics relevant to restoration. Discrimination may rely on only a subset of cues, so the model may not learn other degradation information needed for restoration. Cross-content re-degradation, as explored in ReDSR (Liu et al., 2024), provides

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.

complementary functional supervision by requiring a representation extracted from one LR image to re-degrade another clean image. Therefore, we combine cross-content contrastive learning with auxiliary re-degradation supervision to encourage both degradation discrimination and the retention of transferable degradation information.

Beyond degradation information, diffusion-based SR relies on the input LR image to guide the detail recovery. StableSR (Wang et al., 2024) and SeeSR (Wu et al., 2024b) incorporate LR guidance into pretrained diffusion models, while DiT4SR (Duan et al., 2025) enables joint interaction among the LR, noisy-latent, and text streams through attention and cross-stream convolution. During this interaction, the LR representation guides restoration but is also repeatedly updated, potentially weakening its role as a stable reference to the input LR image. Although DiT4SR includes within-block LR residual connections, these connections use the current LR state rather than preserve an immutable initial LR representation across network depth. Therefore, SD-LRMF preserves the initial LR representation as an immutable Stable LR Memory and fuses it with the interaction-updated Dynamic LR State after intermediate blocks, providing a persistent reference as the LR state evolves.

Building on these two designs, we develop DeMemSR, a degradation-aware diffusion Transformer for real-world image super-resolution. DeMemSR combines transferable degradation representation learning with stable–dynamic LR memory fusion to provide degradation guidance and a persistent reference to the initial LR representation.

Our main contributions are summarized as follows:

- We introduce Transferable Degradation Representation Learning (TDRL), which combines cross-content contrastive learning with auxiliary re-degradation supervision to encourage degradation discrimination and the retention of transferable degradation information.
- We propose Stable–Dynamic LR Memory Fusion (SD-LRMF), which preserves the initial LR representation as an immutable memory and fuses it with evolving LR features to maintain persistent spatial guidance.
- We develop DeMemSR, a degradation-aware diffusion Transformer for real-world super-resolution. It incorporates the learned degradation representation through layer-specific conditioning and maintains LR guidance through SD-LRMF. DeMemSR achieves leading results on no-reference perceptual quality metrics across four real-world benchmarks.

2 Related Work

2.1 Degradation Representation Learning

Degradation representations enable SR models to adapt to unknown degradation processes. DASR (Wang et al., 2021a) learns contrastive degradation features to condition an SR network. Subsequent studies address interference from image content: DRAT (Chen et al., 2025) subtracts auxiliary content features from mixed content–degradation representations, while CdCL (Yuan et al., 2025) introduces content-decoupled contrastive learning. Complementary approaches learn representations through degradation transfer. ReDSR (Liu et al., 2024) uses a representation extracted from one LR image to re-degrade another HR image, and UDM (Yang et al., 2025) combines degradation synthesis, compression-based regularization, and contrastive learning to support cross-image degradation transfer. Our TDRL uses cross-content contrastive learning to promote degradation discrimination in the projected features, while auxiliary re-degradation encourages the underlying encoder features to retain degradation information that transfers across image content.

2.2 Diffusion-Based Real-World Image Super-Resolution

Diffusion-based SR methods combine generative priors with guidance from degraded observations. StableSR (Wang et al., 2024) introduces a time-aware encoder and controllable feature fusion, PASD (Yang et al., 2024) uses pixel-aware cross-attention, and SeeSR (Wu et al., 2024b) provides degradation-robust semantic prompts. DiffBIR (Lin et al., 2024) conditions generation on an initial restoration, while SUPIR (Yu et al., 2024) explores large-scale training and restoration-guided sampling. Degradation information has also been incorporated explicitly into diffusion restoration:

DeeDSR (Bi et al., 2024) integrates pretrained degradation representations through a degradation-aware adapter, and S3Diff (Zhang et al., 2024) uses degradation information to guide low-rank adaptation for one-step SR. With a Transformer backbone, DiT4SR (Duan et al., 2025) enables interaction among LR, noisy-latent, and text streams through joint attention and cross-stream convolution. Our DeMemSR extends DiT4SR by incorporating learned degradation representations and retaining the initial LR features as a persistent reference.

3 METHOD

3.1 OVERALL FRAMEWORK

As shown in Figure 1(a), given an LR image I_{LR}, DeMemSR reconstructs an HR image I_{SR} through degradation-aware conditioning and stable LR memory fusion. Following DiT4SR (Duan et al., 2025), restoration is performed in the latent space of a variational autoencoder (VAE) using a rectified-flow Transformer. The LR image is resized to the target resolution and encoded by the VAE, followed by patch and positional embeddings to obtain the initial LR tokens L_0. Text tokens and a pooled text embedding are extracted from a caption generated from the LR image.

To provide degradation information, TDRL trains a Cross-Content Degradation Encoder (CCDE), as illustrated in Figure 1(b), and a projection head using cross-content contrastive learning and auxiliary re-degradation supervision. After this training, the encoder and projection head are frozen, and the auxiliary re-degradation generator is discarded. For each input image, CCDE produces a representation z, which is mapped to a projected representation q. As illustrated in Figure 1(c), DAC transforms q into layer-specific offsets that are added to the existing text–timestep conditioning embeddings, providing degradation guidance to each Transformer block.

As illustrated in Figure 1(d), SD-LRMF preserves L_0 as an immutable Stable LR Memory throughout each Transformer forward pass. After each intermediate block, the interaction-updated LR features are fused with this memory, and the resulting Dynamic LR State is passed to the next block. This allows the LR stream to evolve through feature interaction while retaining a persistent reference to the initial observation. The Transformer guides iterative restoration of the noisy latent, and the final latent is decoded by the VAE to obtain I_{SR}. In the following sections, ℓ denotes the Transformer layer index, and t denotes the diffusion timestep.

3.2 TRANSFERABLE DEGRADATION REPRESENTATION LEARNING

TDRL learns degradation representations through cross-content discrimination and auxiliary re-degradation supervision. A shared Cross-Content Degradation Encoder (CCDE) extracts an encoder feature z, and a projection head maps it to a normalized representation q. Contrastive learning encourages q to distinguish degradation configurations across different image contents, while an auxiliary re-degradation generator uses z to transfer degradation characteristics to another clean image.

Controlled content–degradation quadruplets. Given two clean images x_1 and x_2 with different contents, we independently sample two degradation configurations θ_1 and θ_2. Each configuration is applied to both images:

$$y_{ij} = \mathcal{D}(x_i; \theta_j, \varepsilon_{ij}), \qquad i, j \in \{1, 2\}, \tag{1}$$

where $\mathcal{D}$ is the degradation process and ε_{ij} denotes an independent stochastic realization. The resulting quadruplet is $\mathcal{Q} = \{y_{11}, y_{12}, y_{21}, y_{22}\}$. Images y_{1j} and y_{2j} share the same degradation configuration across different contents, whereas y_{i1} and y_{i2} contain the same content under different configurations.

We adopt the high-order degradation process of Real-ESRGAN (Wang et al., 2021b), including blur, resizing, noise, and JPEG compression. Each configuration specifies the degradation settings and operation order. These settings are shared across the two contents, while noise realizations are sampled independently. Configuration labels are used only to construct the training supervision and are not provided to the encoder.

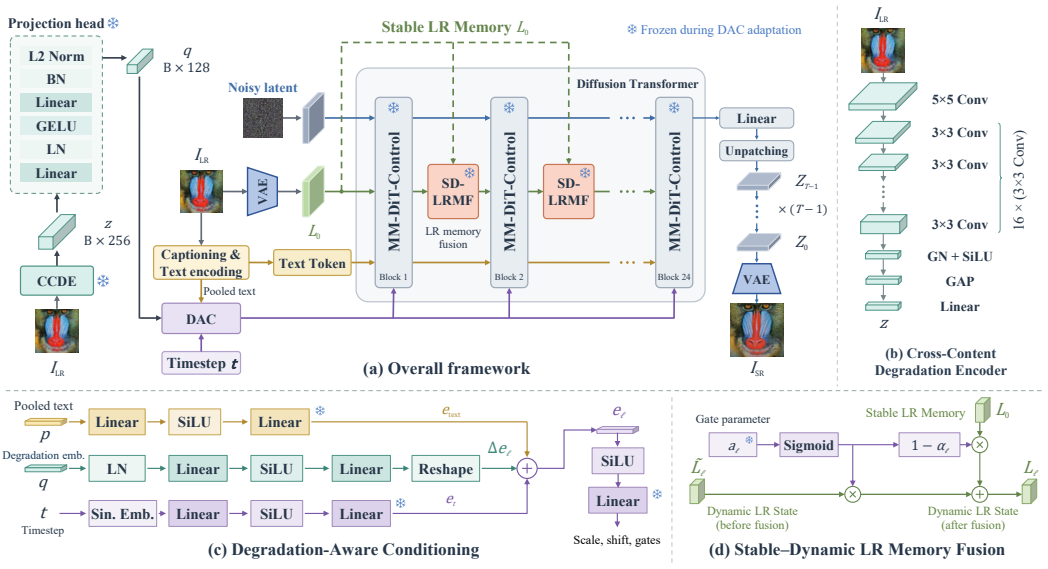

Figure 1: Overview of DeMemSR. (a) Overall framework. (b) Cross-Content Degradation Encoder (CCDE). (c) Degradation-Aware Conditioning (DAC). (d) Stable–Dynamic LR Memory Fusion (SD-LRMF).

Degradation encoding and projection. All four LR images are processed by the shared encoder E and projection head P:

$$z_{ij} = E(y_{ij}) \in \mathbb{R}^{256}, \qquad q_{ij} = \frac{P(z_{ij})}{\|P(z_{ij})\|_2} \in \mathbb{R}^{128}, \tag{2}$$

where z_{ij} is the encoder feature used by the auxiliary re-degradation generator, while q_{ij} is the normalized projected representation supervised by the contrastive objective and subsequently used for SR conditioning. Figure 1(b) illustrates the encoder, with implementation details provided in Appendix A.1.

Cross-content degradation discrimination. Representations extracted from different contents with the same degradation configuration form positive pairs. Different configurations form negative pairs, including those applied to the same content. For a batch of representations indexed by $\mathcal{A}$, let $\mathcal{P}(a)$ denote the positive samples for anchor a. The supervised contrastive loss (Khosla et al., 2020) is

$$\mathcal{L}_{\text{con}} = -\frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \frac{1}{|\mathcal{P}(a)|} \sum_{p \in \mathcal{P}(a)} \log \frac{\exp(q_a^\top q_p/\tau)}{\sum_{b \in \mathcal{A}\setminus\{a\}} \exp(q_a^\top q_b/\tau)}, \tag{3}$$

where τ is the temperature. Within each quadruplet, (q_{11}, q_{21}) and (q_{12}, q_{22}) are positive pairs. This construction encourages degradation discrimination across content and prevents shared image content alone from determining a positive match.

Auxiliary cross-content re-degradation. Degradation discrimination can rely on a subset of cues and does not necessarily preserve other degradation characteristics. Following the cross-content re-degradation principle of ReDSR (Liu et al., 2024), we introduce an auxiliary re-degradation generator G. Given the encoder feature z_{ij} and the other clean image x_{3-i}, the generator produces

$$\widehat{y}_{(3-i)j} = G(x_{3-i}, z_{ij}). \tag{4}$$

For example, z_{11}, extracted from y_{11}, is combined with x_2 to generate $\widehat{y}_{21}$, using y_{21} as the target. The generator receives neither the target LR image nor the degradation configuration. The re-degradation objective averages the four transfers:

$$\mathcal{L}_{\text{redeg}} = \frac{1}{4} \sum_{i=1}^{2} \sum_{j=1}^{2} d_{\text{deg}}\big(\widehat{y}_{(3-i)j}, y_{(3-i)j}\big), \tag{5}$$

$$d_{\text{deg}} = d_{\text{LF}} + 0.1 d_{\text{VGG}} + 0.05 d_{\mu} + 0.1 d_{\sigma} + 0.05 d_{\text{spec}}, \tag{6}$$

where d_{LF}, d_{VGG}, d_{μ}, d_{σ}, and d_{spec} measure low-frequency reconstruction differences, VGG perceptual differences, differences in channel-wise means and standard deviations of high-frequency residuals, and radial-spectrum differences, respectively. The low-frequency term supervises image structure, while the statistical and spectral terms compare degradation characteristics without requiring identical random noise realizations. The detailed formulation is provided in Appendix A.2.

Representation learning objective. We first train the encoder and projection head using the contrastive objective, then warm up the re-degradation generator with the encoder and projection head frozen. Joint fine-tuning subsequently optimizes

$$\mathcal{L}_{\mathrm{TDRL}} = \mathcal{L}_{\mathrm{con}} + \lambda_k \mathcal{L}_{\mathrm{redeg}}, \qquad \lambda_k = \min\left(1, \frac{k}{1000}\right), \tag{7}$$

where k is the joint fine-tuning update index. Both objectives update the encoder, while only the contrastive objective directly supervises the projection head. Re-degradation therefore encourages information retention in z, without requiring its projection q to preserve every degradation characteristic. After representation learning, the encoder and projection head are frozen, the re-degradation generator is discarded, and q is passed to DAC for restoration conditioning. The detailed training schedule is provided in Appendix A.3.

3.3 Degradation-Aware Conditioning

DAC combines text, degradation, and timestep information to generate modulation parameters for the diffusion Transformer. As illustrated in Figure 1(c), it comprises three conditioning branches, layer-wise condition fusion, and modulation parameter generation.

Given the pooled text representation p and diffusion timestep t, the text and timestep branches produce

$$e_{\mathrm{text}} = f_{\mathrm{text}}(p), \qquad e_t = f_t(\mathrm{SinEmb}(t)), \tag{8}$$

where f_{text} and f_t denote the text and timestep embedding networks, respectively, and SinEmb denotes sinusoidal timestep encoding. Each network consists of two linear layers separated by a SiLU activation. The degradation branch takes the normalized representation q extracted by the frozen CCDE and projection head, and predicts layer-specific degradation embeddings:

$$[\Delta e_1; \ldots; \Delta e_K; \Delta e_{\mathrm{out}}] = \mathrm{Reshape}(W_2\,\mathrm{SiLU}(W_1\,\mathrm{LN}(q) + b_1) + b_2)\,, \tag{9}$$

where LN denotes layer normalization, and W_1, W_2, b_1, b_2 are learnable parameters. The output is reshaped into $K+1$ embeddings of dimension d: one embedding Δe_ℓ for each of the K Transformer blocks and an additional embedding Δe_{out} for output normalization.

For each block, DAC combines the shared text and timestep embeddings with the corresponding degradation embedding:

$$e_\ell = e_{\mathrm{text}} + e_t + \Delta e_\ell. \tag{10}$$

The fused condition e_ℓ passes through a SiLU activation and a linear projection to generate scale, shift, and gate parameters:

$$e_\ell \xrightarrow{\mathrm{SiLU}} \mathrm{SiLU}(e_\ell) \xrightarrow{\mathrm{Linear}} (\mathbf{s}_\ell, \mathbf{b}_\ell, \mathbf{g}_\ell), \tag{11}$$

where $\mathbf{s}_\ell$, $\mathbf{b}_\ell$, and $\mathbf{g}_\ell$ denote the scale, shift, and gate parameters, respectively. The scale and shift parameters modulate layer-normalized token features within the Transformer block, while the gates control the contributions of the attention and feed-forward outputs to their respective residual connections. Through these operations, the fused text, degradation, and timestep information guides feature processing throughout the restoration process. The additional embedding Δe_{out} adjusts the condition used for output normalization. The three-stream attention structure and rectified-flow objective remain unchanged.

The final linear layer of the degradation branch is initialized to zero, preserving the original conditioning behavior at the start of DAC adaptation. Only the degradation branch is trained during this stage; the SR model, including its text and timestep branches and modulation projections, remains frozen, as do CCDE and its projection head. At inference, DAC derives degradation conditions directly from the LR input without requiring explicit degradation parameters.

3.4 Stable–Dynamic LR Memory Fusion

Stable–Dynamic LR Memory Fusion (SD-LRMF) complements global degradation conditioning with spatial guidance from the LR observation. SD-LRMF preserves the initial LR representation L_0 as an immutable *Stable LR Memory* and uses it to re-anchor the interaction-updated *Dynamic LR State* after each intermediate Transformer block. As illustrated in Figure 1(d), the fusion is formulated as

$$\alpha_\ell = \mathrm{sigmoid}(a_\ell), \qquad L_\ell = \alpha_\ell \widetilde{L}_\ell + (1-\alpha_\ell) L_0, \quad \ell = 1, \ldots, K-1, \tag{12}$$

where a_ℓ is a learnable scalar associated with the ℓ-th Transformer block, and $\widetilde{L}_\ell$ and L_ℓ denote the Dynamic LR State before and after fusion, respectively. The fused state L_ℓ is passed to the next block. The final block requires no fusion gate because its LR output is not used by a subsequent block; SD-LRMF therefore introduces only $K-1=23$ learnable scalar parameters.

By repeatedly accessing the same Stable LR Memory, SD-LRMF provides a persistent reference to the LR observation while allowing the Dynamic LR State to evolve across network depth.

3.5 Training and Inference

After representation learning, we freeze CCDE and its projection head and discard the auxiliary re-degradation generator. The restoration framework is then trained in two stages.

Stage I: Memory fusion training. We jointly optimize the inherited SR control modules and SD-LRMF gates while keeping the pretrained generative backbone frozen. This stage learns to restore image details with spatial guidance from the LR observation, while incorporating the Stable LR Memory into the evolving Dynamic LR State.

Stage II: Degradation-aware adaptation. We freeze the restoration model obtained in Stage I and train only the degradation branch of DAC. The frozen CCDE and projection head extract the degradation representation q, which DAC maps to layer-specific conditioning offsets. This stage introduces degradation-aware conditioning into the restoration model while preserving the learned memory fusion mechanism.

Both stages use the inherited rectified-flow construction. For HR latent u_x, Gaussian noise ϵ, and noise level σ_t, the implemented preconditioning and loss are

$$u_t = (1-\sigma_t)u_x + \sigma_t \epsilon, \qquad \widehat{u}_x = u_t - \sigma_t f_\Theta(u_t, t; y, p, q), \qquad \mathcal{L}_{\mathrm{SR}} = \mathbb{E}\left[\frac{\|\widehat{u}_x - u_x\|_2^2}{D_u}\right], \tag{13}$$

where D_u denotes the total number of latent elements. The degradation condition q is introduced only in Stage II. At inference, the frozen CCDE and projection head extract q from the LR input for degradation-aware restoration. The auxiliary re-degradation generator is discarded after representation learning.

4 Experiments

We evaluate DeMemSR on four real-world image super-resolution benchmarks. We first describe the experimental settings and compare DeMemSR with existing methods through quantitative and qualitative evaluations. We then analyze the learned degradation representations in terms of cross-content consistency and degradation discrimination. Finally, we conduct ablation studies to assess the contributions of memory fusion and degradation-aware conditioning.

4.1 Experimental Settings

Datasets. We use DIV2K (Agustsson & Timofte, 2017), DIV8K (Gu et al., 2019), FFHQ (Karras et al., 2019), Flickr2K (Lim et al., 2017), Flickr8K (Hodosh et al., 2013), NKUSR8K (Duan et al., 2025), OSTv2 (Wang et al., 2018), and Unsplash2K (Kim & Son, 2021) as training data. Following the high-order degradation pipeline of Real-ESRGAN (Wang et al., 2021b), we synthesize LR–HR

Table 1: No-reference comparison on DRealSR, RealSR, RealLR200, and RealLQ250 ($4\times$). Bold and underlined values indicate the best and second-best results, respectively.

Dataset	Method	MUSIQ↑	MANIQA↑	CLIP-IQA↑	LIQE↑	Q-Align↑
DRealSR	StableSR (IJCV 2024)	58.7464	0.3868	0.6277	3.2507	3.0348
	PASD (ECCV 2024)	50.9877	0.3738	0.5077	2.7904	3.0074
	SeeSR (CVPR 2024)	65.0905	**0.5129**	0.6909	<u>4.1262</u>	3.5869
	SUPIR (CVPR 2024)	55.1819	0.4337	0.6260	3.2408	3.2761
	DiT4SR (ICCV 2025)	64.6649	0.4423	0.6682	3.9415	3.4304
	FiDeSR (CVPR 2026)	65.3565	0.4585	<u>0.6924</u>	3.9622	<u>3.6243</u>
	StrSR-Z (ECCV 2026)	61.0322	0.4487	0.6377	3.4665	3.3426
	GDPO-SR (CVPR 2026)	<u>65.5502</u>	0.4678	**0.7023**	4.0389	3.5777
	DeMemSR (Ours)	**67.2561**	<u>0.4896</u>	0.6844	**4.2813**	**3.6392**
RealSR	StableSR (IJCV 2024)	64.8809	0.4228	0.6226	3.4780	3.2146
	PASD (ECCV 2024)	59.8234	0.3954	0.4865	3.1668	3.2885
	SeeSR (CVPR 2024)	<u>69.8255</u>	**0.5437**	0.6704	<u>4.1363</u>	**3.7188**
	SUPIR (CVPR 2024)	60.0190	0.4697	0.6296	3.4678	3.4923
	DiT4SR (ICCV 2025)	67.8713	0.4650	0.6375	4.0064	3.5197
	FiDeSR (CVPR 2026)	69.5767	0.4611	**0.6821**	4.0382	3.6889
	StrSR-Z (ECCV 2026)	66.1677	<u>0.5070</u>	0.6489	3.6885	3.6576
	GDPO-SR (CVPR 2026)	69.2788	0.4634	<u>0.6754</u>	4.1060	3.6222
	DeMemSR (Ours)	**70.2797**	0.4963	0.6708	**4.2521**	<u>3.7105</u>
RealLR200	StableSR (IJCV 2024)	63.1971	0.3726	0.5943	3.3799	3.9645
	PASD (ECCV 2024)	61.9639	0.3823	0.5663	3.3188	4.0626
	SeeSR (CVPR 2024)	69.5970	**0.4977**	0.6835	4.0392	4.2371
	SUPIR (CVPR 2024)	61.2643	0.4131	0.5919	3.3321	4.1364
	DiT4SR (ICCV 2025)	70.4876	0.4772	<u>0.6953</u>	<u>4.3506</u>	4.2676
	FiDeSR (CVPR 2026)	<u>70.6201</u>	0.4591	0.6880	4.1467	<u>4.3795</u>
	StrSR-Z (ECCV 2026)	67.9139	0.4660	0.6626	3.7763	4.2422
	GDPO-SR (CVPR 2026)	70.5114	0.4471	0.6892	4.1915	4.3503
	DeMemSR (Ours)	**71.6738**	<u>0.4945</u>	**0.7174**	**4.4749**	**4.3857**
RealLQ250	StableSR (IJCV 2024)	57.0034	0.2930	0.5143	2.7207	3.5416
	PASD (ECCV 2024)	62.0208	0.3612	0.5559	3.1592	3.9288
	SeeSR (CVPR 2024)	70.3772	**0.4895**	0.7061	3.9787	4.1418
	SUPIR (CVPR 2024)	62.7464	0.3442	0.5449	3.2818	4.1334
	DiT4SR (ICCV 2025)	<u>72.0244</u>	0.4715	<u>0.7156</u>	<u>4.4309</u>	4.2981
	FiDeSR (CVPR 2026)	71.6198	0.4508	0.7032	4.1016	<u>4.3874</u>
	StrSR-Z (ECCV 2026)	68.6228	0.4322	0.6582	3.7427	4.2613
	GDPO-SR (CVPR 2026)	71.0973	0.4293	0.6988	4.1324	4.3601
	DeMemSR (Ours)	**72.8239**	<u>0.4882</u>	**0.7282**	**4.5435**	**4.4174**

training pairs with resolutions of 128×128 and 512×512, respectively. For degradation-encoder pretraining, we construct 16,386 quadruplets. Each quadruplet combines two image contents with two degradation configurations, producing four LR observations. The resulting 16,386 quadruplets are split into 14,747 training samples and 1,639 validation samples.

We evaluate all methods at a $4\times$ upscaling factor on four real-world benchmarks. The paired evaluation subsets contain 93 LR–HR crop pairs from DRealSR (Wei et al., 2020) and 100 pairs from RealSR (Cai et al., 2019). We additionally evaluate on RealLR200 and RealLQ250, which contain 200 and 250 real-world LR images without HR references, respectively.

Implementation details. DeMemSR is built upon the SD3.5 Medium backbone (Esser et al., 2024) and the architecture of DiT4SR (Duan et al., 2025). Following degradation-encoder pretraining (Appendix A.3), SR adaptation proceeds in two stages. We first train the inherited SR control

modules and the SD-LRMF gates for 50,000 updates, then freeze the resulting restoration model and optimize only the degradation branch of DAC for an additional 5,000 updates. Detailed SR optimization settings are provided in Appendix B.2.

Evaluation metrics and compared methods. We report five no-reference image-quality metrics on all four datasets: MUSIQ (Ke et al., 2021), MANIQA (Yang et al., 2022), CLIP-IQA (Wang et al., 2023), LIQE (Zhang et al., 2023), and Q-Align (Wu et al., 2024a).

We compare DeMemSR with StableSR, PASD, SeeSR, SUPIR, DiT4SR, FiDeSR (Kim et al., 2026), StrSR (Wang et al., 2026), and GDPO-SR (Yi et al., 2026). The quantitative comparison uses StrSR-Z, while the qualitative comparison uses StrSR-F. We use their publicly released pretrained weights and follow their official inference procedures. All methods are evaluated on the same test images at a $4\times$ upscaling factor using a common implementation of the evaluation metrics (Appendix C).

4.2 Comparison with Existing Methods

Quantitative comparisons. Table 1 reports quantitative results on DRealSR, RealSR, RealLR200, and RealLQ250 using five no-reference image-quality metrics. DeMemSR achieves the highest MUSIQ and LIQE scores on all datasets. It also ranks first in Q-Align on DRealSR and second on RealSR. On RealLR200 and RealLQ250, it ranks first in four of the five metrics and second in MANIQA. Compared with DiT4SR, DeMemSR improves all five metrics across all benchmarks.

Qualitative comparisons. Figure 2 compares three real-world examples. In the foliage example, PASD and SUPIR produce blurred details, while FiDeSR and StrSR-F leave leaf boundaries less distinct. DeMemSR produces clearer contours and finer surface textures than DiT4SR. In the stone-wall example, PASD, SeeSR, SUPIR, and StrSR-F produce less clearly defined masonry structures, while DiT4SR and FiDeSR show weaker mortar lines. DeMemSR achieves a more visually coherent result, with distinct joints in both illuminated and shaded regions. In the flower example, PASD, SUPIR, StrSR-F, and FiDeSR produce relatively blurred petals, and FiDeSR additionally introduces a curved artifact around the bud. Compared with DiT4SR, DeMemSR produces finer petal textures and smoother color transitions, yielding a more natural appearance.

4.3 Analysis of Degradation Representations

We analyze the learned degradation representations from two perspectives: cross-content consistency and degradation discrimination. Using HR images from DRealSR and RealSR, we construct controlled samples by applying the same degradation configuration to different image contents and different configurations to the same content. We measure the cosine similarity between representations of different contents with the same degradation, denoted by s_+, and representations of the same content with different degradations, denoted by s_-. A larger gap $s_+ - s_-$ indicates stronger separation between these two types of pairs.

Table 2 compares the encoder feature z and its projection q from the same jointly trained model. Here, ranking accuracy is the fraction of synthetic anchors satisfying $s_+ > s_-$; it is not classification accuracy on native real-world degradations. For q, the average same-degradation similarity remains above 0.93 on both datasets, whereas the different-degradation similarity becomes negative. These results show that the projection strengthens degradation discrimination while maintaining high cross-content consistency. Moreover, q ranks same-degradation pairs above different-degradation pairs for 100.0% and 99.5% of the evaluated anchors, respectively. This analysis supports the use of q as a degradation representation for DAC (Appendix D).

4.4 Ablation Studies

Effect of SD-LRMF and DAC. We evaluate the contributions of SD-LRMF and DAC on RealLR200 using MUSIQ and Q-Align. As shown in Table 3, removing both components yields the baseline. Adding SD-LRMF increases MUSIQ from 70.9365 to 71.5856 and Q-Align from 4.3492 to 4.3789. Adding DAC alone improves MUSIQ to 71.3533, with a marginal increase in Q-Align to 4.3502. The full model achieves the highest scores of 71.6738 and 4.3857, respectively, suggesting complementary benefits from the two components.

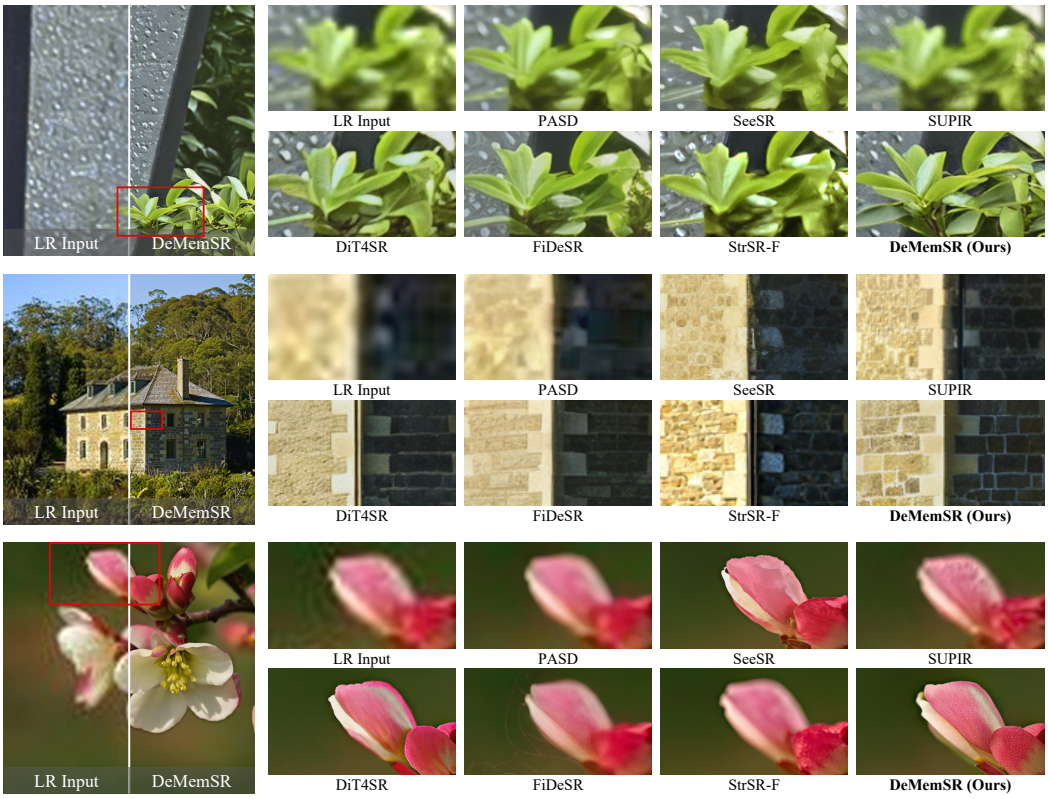

Figure 2: Qualitative comparison at $4\times$ SR. The top row is from RealSR; the middle and bottom rows are from RealLQ250.

Table 2: Feature analysis of evaluation HR images.

Dataset	Feature	s_+	s_-	Rank. acc. ↑
DRealSR	z (256-D)	0.9915	0.6041	100.0%
	q (128-D)	0.9431	−0.1414	100.0%
RealSR	z (256-D)	0.9899	0.6068	100.0%
	q (128-D)	0.9356	−0.1212	99.5%

Table 3: Ablation on RealLR200 ($4\times$). ✓/×: enabled/disabled. Best results in bold.

SD-LRMF	DAC	MUSIQ ↑	Q-Align ↑
×	×	70.9365	4.3492
✓	×	71.5856	4.3789
×	✓	71.3533	4.3502
✓	✓	**71.6738**	**4.3857**

5 Conclusion

We presented DeMemSR, a framework for real-world image super-resolution that combines transferable degradation representation learning with stable–dynamic LR memory fusion. TDRL learns degradation representations through cross-content contrastive learning and auxiliary re-degradation supervision, and DAC incorporates these representations into the diffusion Transformer through layer-wise conditioning. SD-LRMF fuses the evolving LR features with a fixed memory of the initial observation, maintaining a persistent source of spatial guidance. Experiments on four real-world benchmarks demonstrate competitive restoration performance, with the highest MUSIQ and LIQE scores among the evaluated methods and improved visual quality. Representation analysis supports cross-content consistency and degradation discrimination under controlled degradations, while component ablations on RealLR200 suggest complementary benefits from SD-LRMF and DAC. These results highlight the potential of combining learned degradation guidance with stable LR information for real-world super-resolution.

Reproducibility Statement

The formulations of TDRL, DAC, and SD-LRMF are provided in Sections 3.2–3.4. Appendix A details the degradation encoder, projection head, auxiliary generator, loss functions, and pretraining schedule. Appendix B documents the training data, degradation synthesis, and SR optimization settings. Appendix C specifies the evaluation datasets, metrics, preprocessing, and method-specific inference configurations. Appendix D describes the feature-analysis protocol. We will make our source code publicly available upon acceptance of the paper.

Statement of AI Use

AI tools were used to polish the language of this manuscript, improving grammar, clarity, and readability.

References

Eirikur Agustsson and Radu Timofte. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 126–135, July 2017.

Chunyang Bi, Xin Luo, Sheng Shen, Mengxi Zhang, Huanjing Yue, and Jingyu Yang. DeeDSR: Towards Real-World Image Super-Resolution via Degradation-Aware Stable Diffusion, 2024. URL https://arxiv.org/abs/2404.00661.

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. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Cai_Toward_Real-World_Single_Image_Super-Resolution_A_New_Benchmark_and_a_ICCV_2019_paper.html.

Sen Chen, Hongying Liu, Chaowei Fang, Fanhua Shang, Yuanyuan Liu, Liang Wan, Dongmei Jiang, and Yaowei Wang. Unsupervised Degradation Representation Aware Transform for Real-World Blind Image Super-Resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2177–2185, 2025. doi: 10.1609/aaai.v39i2.32216. URL https://ojs.aaai.org/index.php/AAAI/article/view/32216.

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 (ICCV)*, pp. 18948–18958, October 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. DIV8K: DIVerse 8K Resolution Image Dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3512–3516, 2019. doi: 10.1109/ICCVW.2019.00435. URL https://www.research-collection.ethz.ch/items/9535ac6e-9215-4bdc-9dc7-0c717247aaa0.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, 47: 853–899, 2013. URL https://hockenmaier.cs.illinois.edu/Framing_Image_Description/KCCA.html.

Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, June 2019.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-Scale Image Quality Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5148–5157, October 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Aro Kim, Myeongjin Jang, Chaewon Moon, Youngjin Shin, Jinwoo Jeong, and Sang-hyo Park. FiDeSR: High-Fidelity and Detail-Preserving One-Step Diffusion Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38270–38280, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Kim_FiDeSR_High-Fidelity_and_Detail-Preserving_One-Step_Diffusion_Super-Resolution_CVPR_2026_paper.html.

Younggeun Kim and Donghee Son. Noise Conditional Flow Model for Learning the Super-Resolution Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 424–432, June 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 136–144, July 2017.

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 *Computer Vision – ECCV 2024*, pp. 430–448. Springer, 2024. doi: 10.1007/978-3-031-73202-7_25. URL https://link.springer.com/chapter/10.1007/978-3-031-73202-7_25.

Hongda Liu, Longguang Wang, Ye Zhang, Kaiwen Xue, Shunbo Zhou, and Yulan Guo. Preserving Full Degradation Details for Blind Image Super-Resolution, 2024. URL https://arxiv.org/abs/2407.01299v1.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015. URL https://www.robots.ox.ac.uk/~vgg/publications/2015/Simonyan15/.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for Assessing the Look and Feel of Images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023. doi: 10.1609/aaai.v37i2.25353. URL https://ojs.aaai.org/index.php/AAAI/article/view/25353.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024. doi: 10.1007/s11263-024-02168-7. URL https://link.springer.com/article/10.1007/s11263-024-02168-7.

Jingkai Wang, Yixin Tang, Jue Gong, Jiatong Li, Shu Li, Libo Liu, Jianliang Lan, Yutong Liu, and Yulun Zhang. Spectral and Trajectory Regularization for Diffusion Transformer Super-Resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 17060 of *Lecture Notes in Computer Science*, pp. 94–113, 2026. doi: 10.1007/978-3-032-37092-1_6. URL https://arxiv.org/abs/2603.06275.

Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised Degradation Representation Learning for Blind Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10581–10590, June 2021a.

Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 606–615, June 2018.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training Real-World Blind Super-Resolution With Pure Synthetic Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1905–1914, October 2021b.

Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component Divide-and-Conquer for Real-World Image Super-Resolution. In *Computer Vision – ECCV 2020*, pp. 101–117. Springer, 2020. doi: 10.1007/978-3-030-58598-3_7. URL https://link.springer.com/chapter/10.1007/978-3-030-58598-3_7.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029. PMLR, 21–27 Jul 2024a. URL https://proceedings.mlr.press/v235/wu24ah.html.

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 (CVPR)*, pp. 25456–25467, June 2024b.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-Dimension Attention Network for No-Reference Image Quality Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1191–1200, June 2022.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-Aware Stable Diffusion for Realistic Image Super-resolution and Personalized Stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 74–91, 2024. doi: 10.1007/978-3-031-73247-8_5. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/01705.pdf.

Wenbo Yang, Zhongling Wang, and Zhou Wang. Towards a Universal Image Degradation Model via Content-Degradation Disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12966–12975, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yang_Towards_a_Universal_Image_Degradation_Model_via_Content-Degradation_Disentanglement_ICCV_2025_paper.html.

Qiaosi Yi, Shuai Li, Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, and Lei Zhang. GDPO-SR: Group Direct Preference Optimization for One-Step Generative Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2177–2187, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yi_GDPO-SR_Group_Direct_Preference_Optimization_for_One-Step_Generative_Image_Super-Resolution_CVPR_2026_paper.html.

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 (CVPR)*, pp. 25669–25680, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Yu_Scaling_Up_to_Excellence_Practicing_Model_Scaling_for_Photo-Realistic_Image_CVPR_2024_paper.pdf.

Jiang Yuan, Ji Ma, Bo Wang, and Weiming Hu. Content-decoupled Contrastive Learning-based Implicit Degradation Modeling for Blind Image Super-Resolution. *IEEE Transactions on Image Processing*, 34:4751–4766, 2025. doi: 10.1109/TIP.2025.3558442. URL https://arxiv.org/abs/2408.05440.

Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-Guided One-Step Image Super-Resolution with Diffusion Priors, 2024. URL https://arxiv.org/abs/2409.17058.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14071–14081, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Blind_Image_Quality_Assessment_via_Vision-Language_Correspondence_A_Multitask_Learning_CVPR_2023_paper.html.

APPENDIX

A DEGRADATION-ENCODER PRETRAINING

A.1 ENCODER AND AUXILIARY GENERATOR

CCDE is a residual convolutional network (Figure 1(b)). A stride-two 5×5 convolution is followed by eight residual blocks containing sixteen 3×3 convolutions. Four stages use widths 48, 96, 192, and 384; the first block in each stage downsamples and has a 1×1 shortcut projection. GroupNorm and SiLU precede global average pooling, a $384 \to 256$ linear layer, and LayerNorm. The sixteen-convolution count excludes the stem and shortcut projections. The projection head P applies Linear, LayerNorm, GELU, Linear, and BatchNorm to map z from 256 to 128 dimensions, followed by ℓ_2 normalization. The encoder has 6,386,928 parameters and the projection head has 99,200, totaling 6,486,128.

The auxiliary re-degradation generator takes an HR crop and the raw degradation feature z. It uses a convolutional encoder with two spatial downsampling operations and feature-wise affine modulation in four residual blocks, producing a 128×128 RGB image from a 512×512 HR input. The base generator width is 32. This generator provides auxiliary supervision during encoder pretraining.

The roles of the feature stages are intentionally distinguished:

- $z = E(y)$ is the 256-dimensional raw encoder feature used by the auxiliary generator.
- $q = P(z)/\left\|P(z)\right\|_2$ is the 128-dimensional projected feature used by supervised contrastive learning.
- Gradients from both objectives update the encoder during joint pretraining. Only the contrastive objective directly trains the projection head.

A.2 RE-DEGRADATION LOSS DETAILS

This section expands the distance d_{deg} in Equation 6, which is used to compute the re-degradation loss in Equation 5. For a generated LR image $\hat{y}$ and its target y, the five components are

$$\begin{aligned} d_{\mathrm{deg}}(\hat{y}, y) = {} & \left\|B(\hat{y}) - B(y)\right\|_1 + \frac{0.1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \left\|\phi_j(\hat{y}) - \phi_j(y)\right\|_1 \\ & + 0.05 \left\|\mu(H(\hat{y})) - \mu(H(y))\right\|_1 + 0.1 \left\|\sigma(H(\hat{y})) - \sigma(H(y))\right\|_1 \\ & + 0.05 \left\|R(H(\hat{y})) - R(H(y))\right\|_1 , \end{aligned} \tag{14}$$

where $\left\|\cdot\right\|_1$ denotes the mean absolute value over the corresponding tensor elements. The terms follow the order d_{LF}, d_{VGG}, d_μ, d_σ, and d_{spec} in Equation 6, with unchanged weights.

Low-frequency and statistical terms. $B(\cdot)$ is a fixed Gaussian filter with 9×9 support, standard deviation 2, and reflection padding, applied independently to each color channel. The high-frequency residual is $H(v) = v - B(v)$. The operators μ and σ compute the spatial mean and population standard deviation of this residual separately for each channel. The low-frequency term compares image structure, while the two statistical terms compare the mean and variation of the high-frequency residuals.

Perceptual term. $\phi_j(v)$ is the output at zero-based feature-layer index j of a frozen ImageNet-pretrained VGG16 (Simonyan & Zisserman, 2015), with $\mathcal{S} = \{3, 8, 15\}$. Before feature extraction, RGB inputs are normalized channel-wise using mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. The VGG distance averages the mean absolute feature differences over these three layers. All target quantities are detached from the computation graph. VGG parameters remain frozen, while gradients through the prediction branch propagate to the generated image.

Spectral term. R applies an orthonormal real two-dimensional Fourier transform to each channel of the high-frequency residual, takes its magnitude, and applies $\log(1+\cdot)$. The resulting coefficients

are averaged within 32 uniformly spaced radial frequency bins per channel. The spectral term compares these radial profiles. Together with the low-frequency and statistical terms, it reduces reliance on pixel-wise agreement between independently sampled noise realizations.

A.3 Optimization Phases

The degradation encoder is trained in three successive phases, following Section 3.2. Each phase lasts 5,000 updates for the encoder used in our experiments.

Phase A: Contrastive pretraining. We train CCDE and its projection head using $\mathcal{L}_{\text{con}}$ in Equation 3, with temperature $\tau = 0.07$ and a learning rate of 2×10^{-4}. The auxiliary re-degradation generator is not used in this phase.

Phase B1: Generator warm-up. We freeze the pretrained encoder and projection head and keep them in evaluation mode. Only the auxiliary generator is optimized using $\mathcal{L}_{\text{redeg}}$ in Equation 5, with a learning rate of 2×10^{-4}.

Phase B2: Joint fine-tuning. Starting from the encoder and projection head obtained in Phase A and the generator obtained in Phase B1, we jointly optimize all three modules with $\mathcal{L}_{\text{TDRL}}$ in Equation 7. Both losses update the encoder; the contrastive loss updates the projection head, and the re-degradation loss updates the generator. The encoder and projection head share a learning rate of 2×10^{-5}, while the generator uses 2×10^{-4}. The weight λ_k increases linearly from zero to one over the first 1,000 joint updates and remains one thereafter.

Batch construction and optimization. Each contrastive batch contains eight quadruplets (32 LR images). In Phases B1 and B2, re-degradation is computed on two quadruplets per batch, using all four cross-content transfer directions for each. We use AdamW with weight decay 10^{-4}, 500 updates of linear learning-rate warm-up followed by cosine decay, gradient clipping with maximum norm 1, FP16 training, and random seed 1024.

B Training Data and Implementation

B.1 Training Data Inventory

Table 4 lists the source collections and image counts in the training pool. The inventory contains 32,772 source images, matching the number of offline training pairs. The inherited pair-generation pipeline uses 512×512 crops and unsharp masking before generating the degraded LR images. VAE latents and text embeddings are cached for SR training.

Table 4: Source collection inventory for the 32,772-image training pool. Dataset names with version suffixes follow the local source collection names.

Collection	Images
DIV2K (Agustsson & Timofte, 2017)	800
DIV8K (Gu et al., 2019)	1,500
FFHQ (Karras et al., 2019)	10,000
Flickr2K (Lim et al., 2017)	2,650
Flickr8K (Hodosh et al., 2013)	6,000
NKUSR8K (Duan et al., 2025)	1,000
OSTv2 (Wang et al., 2018)	10,324
Unsplash2K (Kim & Son, 2021)	498
Total	32,772

For encoder pretraining, a deterministic shuffle pairs the 32,772 source crops into 16,386 disjoint content pairs. Each pair receives two configurations, yielding four LR images. A 90:10 split assigns 14,747 quadruplets to training and 1,639 to validation. No source content is shared between these

Table 5: Principal degradation settings used for the crossed encoder training samples. Kernel mixture probabilities are listed in the order isotropic Gaussian, anisotropic Gaussian, generalized isotropic, generalized anisotropic, isotropic plateau, anisotropic plateau.

Setting	First stage	Second stage
Kernel support	21×21	11×11
Kernel mixture	$(.45, .25, .12, .03, .12, .03)$	Same
Sinc-kernel probability	0.1	0.1
Blur σ range	$[0.2, 3.0]$	$[0.2, 1.5]$
Generalized-kernel shape range	$[0.5, 4.0]$	$[0.5, 4.0]$
Plateau-kernel shape range	$[1.0, 2.0]$	$[1.0, 2.0]$
Apply blur probability	1.0	0.8
Resize up/down/keep probabilities	$(0.2, 0.7, 0.1)$	$(0.3, 0.4, 0.3)$
Resize scale range	$[0.15, 1.5]$	$[0.3, 1.2]$
Gaussian noise probability	0.5	0.5
Gaussian noise range	$[1, 30]$	$[1, 25]$
Poisson scale range	$[0.05, 3.0]$	$[0.05, 2.5]$
Grayscale noise probability	0.4	0.4
JPEG quality range	$[30, 95]$	$[30, 95]$

two encoder splits. This split does not imply that the SR model has an independent validation split over the same source pool, and it is not an audit of near-duplicates across public datasets.

The encoder synthesis uses the inherited high-order degradation family (Wang et al., 2021b): blur, random resize, Gaussian or Poisson noise, and JPEG compression, followed by a second degradation stage and a final filtering/resizing operation. Table 5 records its principal parameters. Noise is independently realized for different contents even when they share a configuration. For a given configuration, the other sampled parameters and the operation order are shared across content.

The final sinc-filter probability is 0.8 and the downscaling factor is four. The source crops are already USM-sharpened, so encoder data generation does not sharpen them a second time. The synthetic feature-analysis data in Appendix D instead start from the evaluation HR images and apply USM once.

B.2 SR Framework and Optimization

The SR implementation uses the SD3.5 Medium backbone with 24 Transformer blocks, hidden width 1536, and 2×2 latent patches. We preserve the DiT4SR three-stream attention, LR Residual, and cross-stream convolution. SD-LRMF stores the initial LR tokens rather than introducing another attention branch. Using the one-based block numbering of Figure 1, its 23 scalar gates are initialized in three depth groups: $\alpha_\ell = 0.2$ for blocks 1–8, 0.5 for blocks 9–16, and 0.8 for blocks 17–23. The corresponding logits are initialized by the inverse sigmoid. Training does not impose a monotonicity constraint.

For SR training, the pretrained encoder and projection head are frozen, and the auxiliary re-degradation generator is discarded. During degradation adaptation, the encoder and projection head operate in evaluation mode, including the projection head's normalization statistics, and provide the 128-dimensional projected feature q to DAC. Table 6 summarizes the two SR optimization stages and their training settings.

The gate optimizer has zero weight decay. Timesteps are sampled from the inherited logit-normal density with mean zero and standard deviation one. The clean-latent preconditioning in Equation 13 is used with unit loss weighting; the objective is not an additional pixel-space or adversarial SR loss. DAC's new degradation MLP consists of LayerNorm over 128 dimensions, a $128 \to 512$ linear layer, SiLU, and a $512 \to (25 \cdot 1536)$ linear layer. It has 19,765,504 parameters. The last layer's weights and biases are zero-initialized. Reshaping gives 25 output vectors, one per block plus one for output normalization. These are separate coordinates of one shared MLP, rather than 25 independent MLPs. The inherited text and timestep mappings shown in Figure 1(c) remain frozen during SR adaptation.

Table 6: Two SR optimization stages. The independent encoder pretraining phases are not counted in these SR update totals.

Setting	Memory-fusion base	Degradation adaptation
Updates	50,000	5,000 additional
Trainable modules	SR control modules, memory gates	Degradation conditioner only
Generative backbone	Frozen	Frozen
Existing SR control and gates	Trainable	Frozen
Degradation encoder / projection	Not used	Frozen
Effective batch size	32	32
Control learning rate	5×10^{-5}	–
Gate learning rate	5×10^{-4}	–
Conditioner learning rate	–	10^{-4}
Optimizer	AdamW	AdamW
Learning-rate schedule	Constant	Constant
Text dropout	0.2	0.2
Precision	FP16	FP16

C Evaluation Protocols

C.1 Datasets and Metrics

Paired evaluation uses 93 DRealSR and 100 RealSR crops with 128×128 LR and 512×512 HR resolution. RealLR200 and RealLQ250 are the 200-image and 250-image unpaired evaluation collections, respectively. All four datasets are evaluated using the same five no-reference image-quality metrics: MUSIQ, MANIQA, CLIP-IQA, LIQE, and Q-Align.

Image-wise metrics are averaged over all images in the corresponding set. CLIP-IQA and Q-Align are written with their standard names; they correspond to the `clipiqa` and `qalign` fields in the stored metric tables.

C.2 Inference Settings

Table 7 summarizes important differences between methods. These are the configurations used for the recorded comparisons, not a claim that each setting is optimal for its method. Sampling steps alone are not a hardware-normalized measure of latency. Prompts, color correction, precision, tiling, and generative backbones differ across methods.

The reported runs do not share a common fixed per-image seed. In particular, mixed batch execution in some baseline runs is not a deterministic matched-seed comparison. Our main unpaired evaluation uses latent tiles of size 64 with overlap 24 and minimum processing size 512. For the degradation encoder, paired LR crops are used at native 128×128 resolution. For variable-size unpaired images, only the RGB copy sent to the degradation encoder is bicubic-resized to 128×128; the restoration input follows the regular SR pipeline.

FiDeSR follows the released inference path with an empty prompt, LF/HF merge ratio 0.5, latent tile size 96 and overlap 32, and VAE encoder/decoder tile sizes 1024/224. For narrow inputs, the latent tile is reduced before constructing its blending mask. GDPO-SR uses whole-image VAE inference with memory-efficient attention and the released RAM/DAPE prompting path. Its UNet uses the original attention implementation. Both wrappers restore the final dimensions to exactly four times the original LR size after preprocessing; these size adaptations affect variable-size unpaired inputs. Neither run fixes a sampling seed.

D Additional Feature Analysis

For the diagnostic in Table 2, we apply the encoder's synthetic degradation distribution to evaluation HR content with seed 1024. HR images receive USM with weight 0.5, threshold 10, and window size 51, matching the source sharpening convention. DRealSR has an odd image count; one file (`Canon_40_x1.png`) is excluded before forming 46 content pairs from 92 images. RealSR yields

Table 7: Recorded inference configurations.

Method	Steps	Guidance	Relevant configuration
StableSR	200	–	Released 117k SR and 11k CFW weights; DDPM; CFW weight 0.5; AdaIN.
PASD	20	9.0	100k weights; CoCa prompts; wavelet color correction; noise timestep 900.
SeeSR	50	5.5	DAPE prompts; LR initialization at timestep 999; AdaIN.
SUPIR	50	$1 \rightarrow 4$	v0Q weights; RestoreEDM sampling; minimum side 1024; wavelet correction.
DiT4SR	40	8.0	Released weights; noise initialization; FP16; AdaIN.
FiDeSR	1	–	SD2.1; timestep 1; empty prompt; FP16; LF/HF scales 0.2/0.2; AdaIN.
GDPO-SR	1	–	SD2.1; diffusion/noise timesteps 100/250; FP32; RAM/DAPE prompts; AdaIN.
StrSR-Z	1	–	Z-Image variant; Qwen conditioning; BF16; tile size 1024, overlap 64.
StrSR-F	1	–	FLUX variant; Qwen conditioning; BF16; tile size 1024, overlap 64.
DeMemSR	40	8.0	Noise initialization; FP16; AdaIN; degradation scale $s = 1$.

50 pairs from all 100 images. Each pair produces a four-image degradation quadruplet, giving 184 and 200 anchor comparisons, respectively.

All features are ℓ_2-normalized for cosine similarity evaluation, including raw 256-dimensional features. The complete encoder comparison appears in Table 8. The earlier locally trained encoder is included as a historical comparator; it differs in training from the final encoder. Moreover, DASR and this comparator are evaluated in their raw 256-dimensional feature spaces, while q is a contrastively optimized 128-dimensional projection.

Table 8: Complete feature diagnostic. Scores are means over synthetic quadruplets formed from evaluation HR content. Feature stages and training differ between the encoder rows.

Content source	Encoder / feature	s_+	s_-	Gap	Ranking accuracy
DRealSR	DASR, raw 256-D	0.9799	0.9601	0.0198	55.4%
	Earlier local encoder, 256-D	0.3910	0.1817	0.2092	72.8%
	DeMemSR, z (256-D)	0.9915	0.6041	0.3874	100.0%
	DeMemSR, q (128-D)	0.9431	-0.1414	1.0846	100.0%
RealSR	DASR, raw 256-D	0.9918	0.9627	0.0291	70.0%
	Earlier local encoder, 256-D	0.4238	0.1477	0.2761	80.0%
	DeMemSR, z (256-D)	0.9899	0.6068	0.3831	100.0%
	DeMemSR, q (128-D)	0.9356	-0.1212	1.0568	99.5%

The ranking statistic asks whether an anchor is more similar to a same-configuration/different-content image than to a different-configuration/same-content image. It is not a multiclass classification task.