LEARNING WHAT TO DEGRADE: A RECOVERABILITY-AWARE CURRICULUM FOR RELIABLE SMART-GLASSES SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Images captured by smart glasses often suffer from complex degradations, obscuring small text and fine-grained structures essential for visual assistance. Although super-resolution (SR) can improve perceptual quality, these gains do not necessarily translate into reliable downstream understanding by vision-language models (VLMs). We introduce a high-resolution (HR)-referenced semantic evaluation protocol that compares VLM responses to restored images with those obtained from corresponding HR references. The evaluation reveals two failure modes: generative SR can introduce misleading details, while restoration-oriented SR can leave critical details blurred or ambiguous. To improve detail recovery in restoration-oriented SR, we investigate adaptive degradation scheduling. High reconstruction error may reflect either insufficient model capability or excessive information loss, requiring different adjustments to degradation severity. We therefore propose ReDiC, a Recoverability–Difficulty Curriculum that jointly estimates input recoverability and model-relative restoration difficulty to adaptively control degradation severity during training. The scheduler increases severity under easy and recoverable conditions, maintains challenging but recoverable conditions, and reduces severity when estimated recoverability is low. ReDiC modifies only degradation generation during training and is compatible with existing SR architectures. Experiments on synthetic degradation benchmarks show consistent improvements across NAFNet, SwinIR, and MambaIR, yielding average PSNR gains of 2.36–2.65 dB over their corresponding baselines while improving downstream recognition and HR-referenced semantic agreement. Both automatic and human evaluations on real smart-glasses captures show improved downstream semantic agreement across all three backbones.

1 INTRODUCTION

Images captured by smart glasses often suffer from degradations such as low resolution and blur, obscuring small text and fine-grained structures Engel et al. (2023). The loss of these details can impair downstream visual understanding Wang et al. (2020); Qiu et al. (2025), leading to navigation errors or failed place identification, as illustrated in Figure 1. Super-resolution (SR) offers a way to recover these visual details by reconstructing high-resolution images from low-resolution (LR) inputs.

Improving the appearance of degraded inputs, however, does not necessarily improve the reliability of downstream understanding. Recent GAN- and diffusion-based SR methods exploit strong generative priors to produce sharper and more realistic images Wang et al. (2021); Lin et al. (2024). Yet visually plausible details may deviate from the underlying scene content and lead to incorrect downstream predictions Wu et al. (2024). To examine this gap across generative and restoration-oriented SR, we introduce a high-resolution (HR)-referenced semantic evaluation protocol that compares vision-language model (VLM) responses to restored images with those obtained from the corresponding HR references. As illustrated in Figure 2, our evaluation reveals two failure modes: generative SR can synthesize visually realistic but incorrect details that mislead VLMs, while restoration-oriented SR can leave critical details blurred or ambiguous, providing insufficient visual evidence for reliable downstream understanding.

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

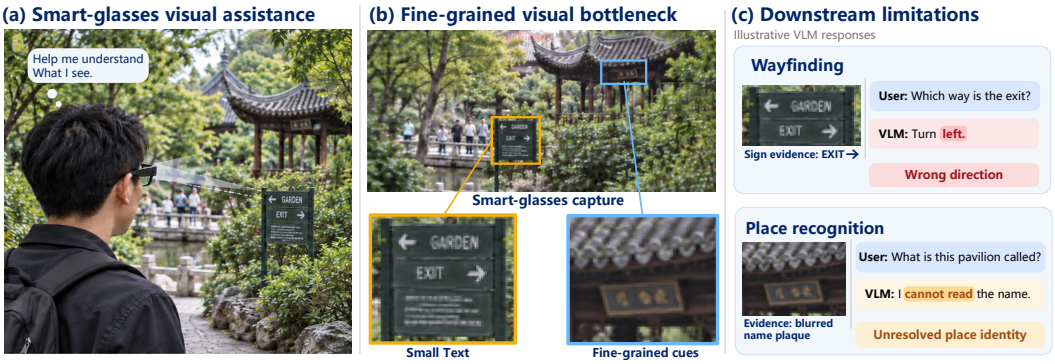

Figure 1: Visual evidence bottlenecks in smart-glasses assistance. (a) A user wearing smart glasses for visual assistance in an everyday environment. (b) A smart-glasses capture with enlarged regions highlighting small text and fine-grained visual cues obscured by limited image quality. (c) Illustrative VLM responses show two potential downstream consequences: incorrect navigation guidance from a misread exit sign and failed place identification from an unreadable name plaque.

Motivated by the insufficient detail recovery observed in restoration-oriented SR, we examine how training degradations can be adapted to improve restoration. Training degradations determine both the information retained in LR inputs and the restoration difficulty faced by the current model. Existing methods sample corruptions from predefined distributions Zhang et al. (2021); Wang et al. (2021), but increasing degradation severity does not necessarily create more useful training challenges. The effects depend on image content: the same degradation may preserve useful information in some regions while removing essential details in others. High reconstruction error may therefore reflect either insufficient model capability or excessive information loss, making error alone an incomplete guide for adjusting degradation severity.

We therefore propose ReDiC (Recoverability–Difficulty Curriculum), which adapts training degradations using a model-independent recoverability proxy and model-relative restoration difficulty. These signals guide an adaptive scheduler that increases severity under easy and recoverable conditions, maintains challenging but recoverable conditions, and reduces severity when estimated recoverability is low. ReDiC modifies only degradation generation during training and is compatible with existing SR architectures.

Our contributions are summarized as follows:

- We reveal a semantic reliability gap in smart-glasses SR through an HR-referenced evaluation protocol, showing that perceptual enhancement does not necessarily improve downstream understanding.
- We propose ReDiC (Recoverability–Difficulty Curriculum), which distinguishes degradation recoverability from model-relative restoration difficulty and adaptively adjusts training degradation.
- We demonstrate that ReDiC consistently improves restoration quality and downstream semantic reliability across multiple SR backbones and real-world smart-glasses captures.

2 Related Work

Real-World Super-Resolution and Degradation Modeling. Representative restoration backbones include NAFNet Chen et al. (2022), SwinIR Liang et al. (2021), and MambaIR Guo et al. (2024). Alongside architectural advances, realistic degradation synthesis is important for real-world SR. BSRGAN Zhang et al. (2021) constructs complex degradations by randomly composing and shuffling multiple operations, while Real-ESRGAN Wang et al. (2021) adopts a high-order degradation process to better approximate practical image corruption. More recently, RealDGen Peng et al. (2025) learns realistic paired data by decoupling image content from degradation characteristics. These methods demonstrate the importance of degradation modeling, but the degradation distribution is typically predefined or learned independently of the current restoration state. In contrast, we

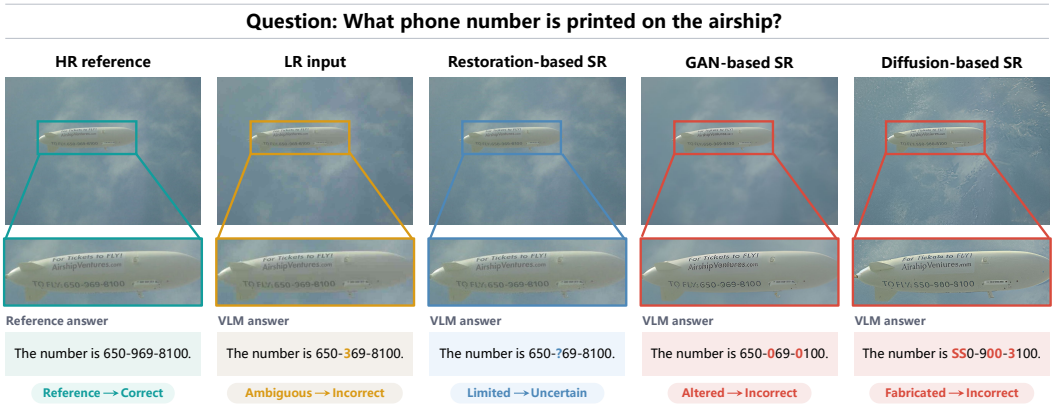

Figure 2: Visual appearance and downstream text understanding in a phone-number reading example. Columns show the HR reference, LR input, and outputs from restoration-, GAN-, and diffusion-based SR methods. Each column includes an enlarged phone-number region and the corresponding VLM response to the same question. The generative SR examples illustrate how sharper-looking text can contain altered digits and yield responses that disagree with the HR-derived answer.

treat degradation generation as a dynamic training process and adapt its severity according to both sample recoverability and model-relative restoration difficulty.

Generative Priors and Semantic Fidelity. Generative SR methods leverage strong image priors to recover visually realistic details under severe degradation Ledig et al. (2017). StableSR Wang et al. (2024) exploits pretrained diffusion priors for blind SR, while DiffBIR Lin et al. (2024) separates degradation removal from generative reconstruction. SeeSR Wu et al. (2024) introduces degradation-aware semantic guidance, SUPIR Yu et al. (2024) scales generative restoration with large diffusion priors, and recent work further explores controllable fidelity–realism trade-offs Zhu et al. (2026). Although these methods substantially improve perceptual realism, details absent from the LR observation may be synthesized from the learned prior rather than recovered from image evidence. Related work such as DA-CLIP uses vision-language representations to guide restoration Luo et al. (2024); Li et al. (2026); Wang et al. (2026). Our focus is complementary: we study whether restored images themselves remain reliable evidence for downstream VLMs, and show that perceptually convincing restoration can introduce unsupported content that alters machine interpretation.

Curriculum Learning and Adaptive Difficulty. Curriculum learning improves optimization by organizing training examples according to difficulty Bengio et al. (2009); Kumar et al. (2010); Graves et al. (2017). Recent work on diffusion models explicitly estimates denoising difficulty across timesteps and constructs an easy-to-hard curriculum from the resulting training dynamics Kim et al. (2025); Xu et al. (2024). However, in image restoration, stronger degradation does not necessarily correspond to more useful difficulty: once relevant information is destroyed, high reconstruction error may reflect intrinsic ambiguity rather than an informative hard sample Dong et al. (2015). We therefore distinguish *model-relative restoration difficulty* from *degradation recoverability*, and adapt degradation severity jointly according to whether a sample remains recoverable and whether the current model has mastered it.

3 Semantic Reliability Gap in Super-Resolution

We assess whether SR improves downstream semantic agreement. Figure 2 provides a qualitative phone-number reading example. The quantitative evaluation compares image descriptions against HR-derived references.

3.1 HR-REFERENCED SEMANTIC EVALUATION

For an HR reference x^{HR}, an LR observation x^{LR}, and an SR output $\hat{x}^{\mathrm{SR}}$, a fixed VLM Φ independently generates descriptions $r = \Phi(x^{\mathrm{HR}})$, $d^{\mathrm{LR}} = \Phi(x^{\mathrm{LR}})$, and $d^{\mathrm{SR}} = \Phi(\hat{x}^{\mathrm{SR}})$. We treat r as an HR-derived semantic reference rather than manually annotated ground truth. Three LLM judges Zheng et al. (2023); Liu et al. (2023) evaluate each LR or SR description against r on semantic consistency, information coverage, factual fidelity, and semantic relevance, each scored on a 0–100 scale. The weighted scores are averaged across judges to obtain $S(d, r)$. We define semantic gain over the LR input as

$$\Delta S = S(d^{\mathrm{SR}}, r) - S(d^{\mathrm{LR}}, r). \tag{1}$$

Positive values indicate improved agreement with the HR-derived reference. Scoring definitions, weights, and judge settings are provided in Appendix A.

3.2 PERCEPTUAL ENHANCEMENT DOES NOT GUARANTEE SEMANTIC RELIABILITY

In Figure 2, sharper-looking generative SR outputs alter digits and yield VLM responses inconsistent with the HR-derived answer, illustrating how perceptual enhancement can introduce misleading evidence.

The quantitative results in Table 1 complement this observation. Real-ESRGAN achieves the lowest LPIPS and DiffBIR the highest PSNR among the evaluated methods, yet their semantic scores decrease by 5.71 and 2.27 points relative to LR, respectively. The restoration-oriented methods instead yield gains of 0.24–0.84 points. These results show that favorable reconstruction or perceptual metrics do not necessarily translate into better agreement with HR-derived semantic references.

Table 1: Reconstruction quality and HR-referenced semantic agreement on real smart-glasses images. Judge Score measures agreement between VLM-generated descriptions and HR-derived references. Δ Score denotes the change relative to the LR input. Best results in each metric column are bold.

Method	Type	PSNR↑	LPIPS↓	Judge↑	Δ Score↑
LR	–	–	–	65.75	–
NAFNet	Restoration	22.93	0.4684	66.46	+0.71
SwinIR	Restoration	22.94	0.4793	65.99	+0.24
MambaIR	Restoration	22.96	0.4761	**66.59**	**+0.84**
Real-ESRGAN	GAN	22.97	**0.2945**	60.04	-5.71
DiffBIR	Diffusion	**23.29**	0.3273	63.48	-2.27

This observation motivates us to focus on restoration-oriented SR and adapt its training degradations. ReDiC distinguishes high error due to limited model capability from high error associated with excessive information loss, using model-relative difficulty and a recoverability proxy to guide degradation severity updates.

4 METHOD

4.1 OVERVIEW

Figure 3 summarizes ReDiC: severity-conditioned degradation generates LR inputs, and recoverability–difficulty estimates guide subsequent severity updates. The curriculum changes only degradation generation, leaving the SR architecture and reconstruction objective unchanged.

4.2 SEVERITY-CONDITIONED DEGRADATION

Given an HR image I^{gt}, we generate its LR observation at curriculum window t as

$$I^{\mathrm{lr}} = \mathcal{G}(I^{\mathrm{gt}}; \phi_t), \qquad \phi_{k,t} \sim p_k(\phi_k \mid s_t), \tag{2}$$

where $\mathcal{G}$ combines blur, noise, compression, and resampling, $\phi_t = \{\phi_{k,t}\}_{k=1}^{K}$ denotes the sampled degradation parameters, and $s_t \in [0, 1]$ is the curriculum-controlled severity. The parameter sampling ranges are continuously interpolated between predefined mild and strong configurations. Increasing s_t therefore shifts the degradation distribution toward stronger corruption while preserving stochastic diversity among samples. The SR scale remains fixed at $\times 2$, so the curriculum adjusts degradation severity without changing the reconstruction scale.

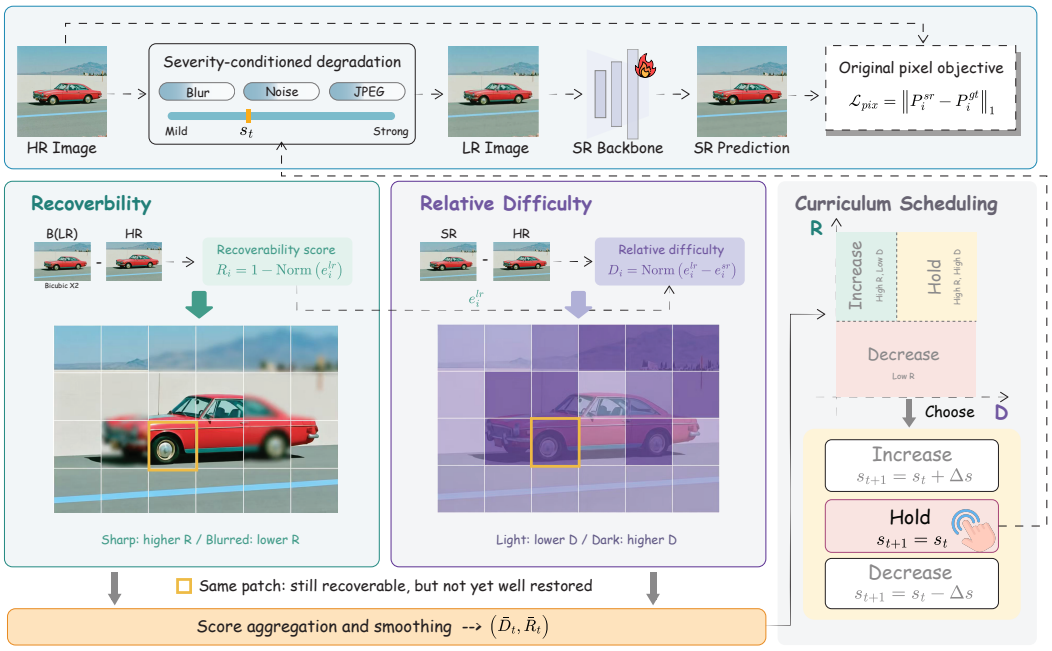

Figure 3: Overview of ReDiC. A severity-conditioned degradation process generates LR training samples with adaptive difficulty. ReDiC estimates recoverability and relative difficulty from the generated samples, and a curriculum scheduler adjusts degradation severity according to the resulting learning state. The SR architecture and reconstruction objective remain unchanged.

4.3 Recoverability–Difficulty Learning-State Estimation

Degradation recoverability. For an LR patch P_i^{lr} with HR target P_i^{gt}, we obtain a fixed reconstruction by bicubic upsampling, $P_i^{\text{b}} = \mathcal{B}(P_i^{\text{lr}})$, and compute its error $e_i^{\text{lr}} = |\Omega_i|^{-1}\|P_i^{\text{b}} - P_i^{\text{gt}}\|_1$, where Ω_i denotes the patch support. We define the recoverability score as

$$R_i = 1 - \text{clip}\left(\frac{e_i^{lr} - e^{low}}{e^{high} - e^{low} + \epsilon}, 0, 1\right), \tag{3}$$

where e^{low} and e^{high} are fixed empirical calibration bounds. A larger R_i indicates lower reconstruction error under fixed interpolation, providing a model-independent proxy for degradation recoverability.

Relative restoration difficulty. Given the SR prediction $P_i^{\text{sr}} = F_\theta(P_i^{\text{lr}})$, we compute its reconstruction error $e_i^{\text{sr}} = |\Omega_i|^{-1}\|P_i^{\text{sr}} - P_i^{\text{gt}}\|_1$ and normalize it by the interpolation baseline:

$$D_i = \text{clip}\left(\frac{e_i^{sr}}{\max(e_i^{lr}, e^{low})}, 0, 1\right). \tag{4}$$

Normalizing by the interpolation error expresses the model's remaining error relative to baseline distortion. High R_i and high D_i indicate relatively low baseline distortion but limited error reduction by the model. Low R_i indicates large baseline distortion and motivates reducing degradation severity.

Robust state aggregation. Within each image, patch-level recoverability and difficulty are aggregated using quantiles q_R and q_D, respectively. The resulting image-level scores are averaged over each curriculum window and smoothed by exponential moving averages to obtain the state $(\bar{R}_t, \bar{D}_t)$ used by the scheduler. Full aggregation equations are provided in Appendix B.

4.4 Recoverability-Aware Curriculum Scheduling

At curriculum window t, ReDiC updates degradation severity using the smoothed state $(\bar{R}_t, \bar{D}_t)$:

$$s_{t+1} = \begin{cases} \mathrm{clip}(s_t - \Delta s, 0, 1), & \bar{R}_t < \tau_R, \\ \mathrm{clip}(s_t + \Delta s, 0, 1), & \bar{R}_t \geq \tau_R \ \wedge \ \bar{D}_t < \tau_D, \\ s_t, & \text{otherwise}, \end{cases} \tag{5}$$

Here, $\Delta s > 0$ is the severity step, and τ_R and τ_D are the recoverability and difficulty thresholds. Recoverability takes priority: low $\bar{R}_t$ triggers retreat, whereas sufficient $\bar{R}_t$ permits progression only when $\bar{D}_t$ is low. Otherwise, severity is maintained. Training retains the original ℓ_1 reconstruction objective:

$$\mathcal{L}_{\mathrm{pix}} = \frac{1}{B} \sum_{i=1}^{B} \left\| F_\theta(P_i^{\mathrm{lr}}) - P_i^{\mathrm{gt}} \right\|_1 . \tag{6}$$

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on 100 validation images from each of LSVT Sun et al. (2019) and DIV2K Agustsson & Timofte (2017); Timofte et al. (2017), covering text-rich and natural scenes, respectively. We additionally collect a real-world smart-glasses dataset comprising 200 images for evaluation. Each HR image is degraded at Mild ($s \in [0.2, 0.4]$), Medium ($s \in [0.4, 0.6]$), and Hard ($s \in [0.6, 0.8]$) levels using the pipeline in Sec. 4.2. Each real capture has a separately photographed reference from an approximately matched viewpoint, providing corresponding scene content without pixel alignment.

Implementation details. We train NAFNet Chen et al. (2022), SwinIR Liang et al. (2021), and MambaIR Guo et al. (2024) for $\times 2$ SR using online degradations and an ℓ_1 reconstruction loss. Within each backbone pair, training images, architecture, and optimization settings are identical; only the training-time degradation severity schedule differs. Additional training, degradation, and evaluation settings are provided in Appendix C.

Evaluation metrics. For paired data, PSNR, LPIPS Zhang et al. (2018), and Judge Score are averaged over both datasets at each severity level, then equally across the three levels. OCR-NED is evaluated only on text-containing LSVT images, using the same OCR system's HR outputs as references. Semantic evaluation follows Sec. 3, with Δ Score measured relative to LR. For the 200 real captures, we report only semantic scores because the references are not pixel-aligned.

Table 2: Restoration and downstream performance under synthetic degradations. OCR-NED is evaluated only on text-containing LSVT images. Avg. denotes an equal average over the Mild, Medium, and Hard degradation levels. Δ Score denotes the Judge Score gain over the LR input. Best results among restoration methods in each metric column are bold.

Method	Mild PSNR ↑	Medium PSNR ↑	Hard PSNR ↑	Avg. PSNR ↑	Avg. LPIPS ↓	Avg. OCR-NED ↓	Avg. Judge ↑	Avg. Δ Score ↑
LR Input	–	–	–	–	–	0.5490	71.04	–
NAFNet	30.92	29.36	25.44	28.57	0.2825	0.5503	73.54	2.50
NAFNet + ReDiC	33.97	31.41	27.42	30.93	0.2500	0.4860	76.47	5.43
SwinIR	31.02	29.48	25.46	28.65	0.2905	0.5369	73.83	2.79
SwinIR + ReDiC	34.17	31.63	27.34	31.05	0.2361	0.4564	**76.87**	**5.83**
MambaIR	31.19	29.54	25.46	28.73	0.2918	0.5390	73.96	2.92
MambaIR + ReDiC	**34.39**	**31.91**	**27.83**	**31.38**	**0.2298**	**0.4372**	76.85	5.81

5.2 Performance under Synthetic Degradations

Restoration quality. Table 2 shows that ReDiC improves PSNR across all three backbones and degradation levels, with average gains of 2.36–2.65 dB over the corresponding baselines. Gains remain 1.88–2.37 dB under Hard degradation, showing that the benefit extends to more severely degraded inputs. Average LPIPS also decreases for every backbone, indicating improved perceptual similarity alongside higher reconstruction fidelity. Under the matched training protocol, where ReDiC changes only degradation scheduling, these consistent improvements support the effectiveness of adaptive training degradations across the evaluated architectures.

Downstream understanding. ReDiC improves both text recognition and HR-referenced semantic agreement across all three backbones. Average OCR-NED decreases by 11.7–18.9% relative to the corresponding baselines, with MambaIR improving from 0.5390 to 0.4372. Average Judge Score increases by 2.89–3.04 points over the corresponding baselines. All ReDiC variants also outperform the LR input on both metrics, with Judge Score gains of 5.43–5.83 points. The comparison with the LR input shows that ReDiC improves measured downstream performance beyond that obtained by processing the degraded images directly.

5.3 Generalization to Real Smart-Glasses Captures

We evaluate ReDiC on 200 real smart-glasses captures to test whether improvements from training with synthetic degradations transfer to real imaging conditions. The captures cover text-rich scenes, distant objects, and fine structures and are used exclusively for evaluation. Since pixel-aligned HR targets are unavailable, Table 3 compares downstream semantic agreement for raw inputs, Bicubic interpolation, and the three backbones with and without ReDiC, following the HR-referenced protocol in Sec. 3.

Table 3: LLM-based semantic evaluation on 200 real-world smart-glasses captures using HR-derived descriptions as references. Cons., Cov., Fid., and Rel. denote consistency, coverage, fidelity, and relevance, respectively. Judge denotes the weighted aggregate of the four dimensions.

Method	Cons.↑	Cov.↑	Fid.↑	Rel.↑	Judge↑
Raw Glasses	60.8	59.5	50.9	66.6	58.6
Bicubic	60.6	59.7	49.9	66.6	58.3
NAFNet	60.3	60.9	50.2	66.5	58.5
NAFNet + ReDiC	**64.1**	61.3	52.2	69.0	60.9
SwinIR	60.7	59.4	51.1	67.1	58.6
SwinIR + ReDiC	63.1	63.2	**53.2**	68.9	**61.2**
MambaIR	59.6	60.5	50.1	67.0	58.2
MambaIR + ReDiC	63.0	**63.5**	52.3	**69.3**	61.1

Baseline restoration. Baseline restoration yields Judge Scores of 58.2–58.6, no higher than the raw input's 58.6 at the reported precision, indicating that restoration alone does not ensure improved downstream semantic agreement. ReDiC raises the scores of NAFNet, SwinIR, and MambaIR to 60.9, 61.2, and 61.1, respectively, exceeding their baselines by 2.4–2.9 points and improving all four semantic dimensions. All three variants also outperform the raw input by 2.3–2.6 points, showing that restoration with ReDiC benefits downstream interpretation compared with using the captured images directly. These consistent gains support the transfer of ReDiC-trained models from synthetic degradations to the evaluated real-world capture conditions.

Human evaluation. We ask 20 participants to rate VLM descriptions of the 200 real captures against their HR-derived references, using the same four semantic dimensions as the LLM judges. Ratings use a 0–100 scale and the same aggregation weights. The averaged LLM scores achieve a Spearman correlation of 0.81 with the aggregated human scores, indicating strong agreement in the ranking of semantic quality. ReDiC improves mean human scores by 2.2, 2.5, and 2.7 points over the NAFNet, SwinIR, and MambaIR baselines, respectively, supporting the semantic gains observed in automatic evaluation.

5.4 Qualitative Analysis

Figure 4 compares restored images and VLM responses, with HR-derived responses serving as semantic references. In Figure Figure 4(a), the VLM identifies the dragon figure or sculpture with all three ReDiC variants, whereas LR and baseline responses omit this object cue. In Figure 4(b), all three variants yield the correct transcription of `CHARTER`, which the LR input and baselines misread. In the real smart-glasses example in Figure 4(c), the main restaurant name is recognized across

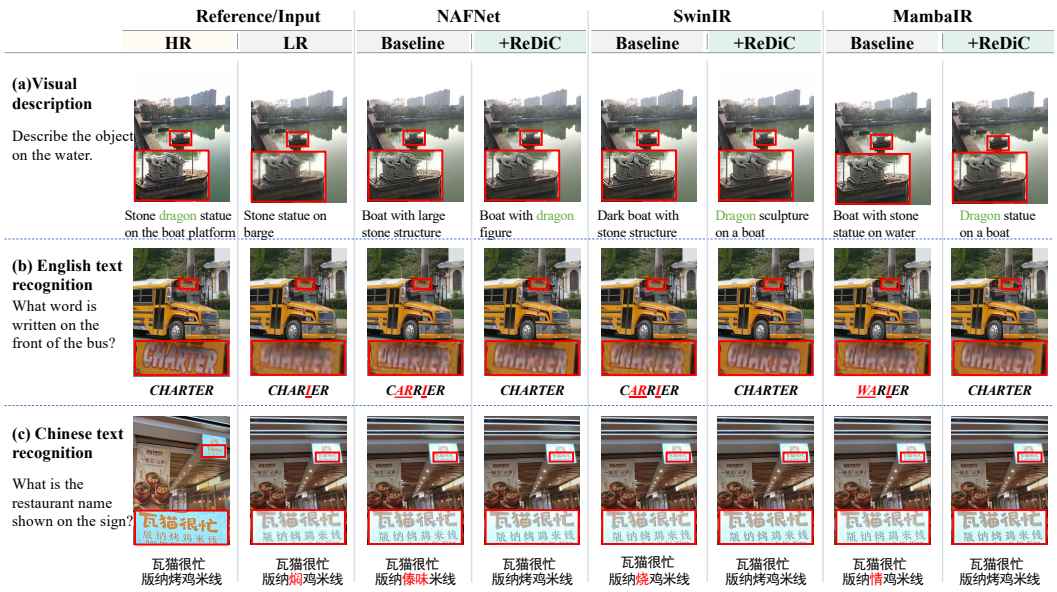

Figure 4: Qualitative comparison of restoration outputs and corresponding VLM responses with and without ReDiC across NAFNet, SwinIR, and MambaIR. Examples cover (a) visual description, (b) English text recognition, and (c) Chinese text recognition on a real smart-glasses capture. Responses to the HR images serve as semantic references. Red boxes indicate the regions shown in enlarged crops.

Table 4: Ablation of recoverability (R) and relative difficulty (D) with MambaIR under synthetic degradations. Checkmarks indicate enabled components. Results are averaged over the three severity levels.

Variant	R	D	Scheduler	PSNR↑	OCR-NED↓	Judge↑
MambaIR	✗	✗	✗	28.73	0.5390	73.96
D-only	✗	✓	✓	30.24	0.4932	74.85
R-only	✓	✗	✓	29.64	0.4733	75.42
ReDiC	✓	✓	✓	**31.38**	**0.4372**	**76.85**

methods, while all ReDiC variants match the HR-derived transcription of the smaller accompanying text. These examples illustrate improvements in object identification and recognition of fine textual details, complementing the quantitative OCR and semantic results.

5.5 Ablation of Recoverability and Difficulty Signals

Table 4 compares a baseline without adaptive scheduling, D-only and R-only curricula, and full ReDiC. Both single-signal variants improve all three metrics over the baseline. D-only achieves higher PSNR than R-only, whereas R-only yields lower OCR-NED and higher Judge Score. This difference suggests that recoverability captures information relevant to downstream performance that model-relative difficulty alone does not fully reflect.

ReDiC achieves the best results across all three metrics, improving PSNR by 1.14–1.74 dB and Judge Score by 1.43–2.00 points over the single-signal variants while further reducing OCR-NED. Each signal thus provides additional gains when the other is already present, supporting their complementary roles in curriculum scheduling. With the pixel reconstruction objective unchanged, these results also show that adapting training degradations can improve downstream semantic agreement.

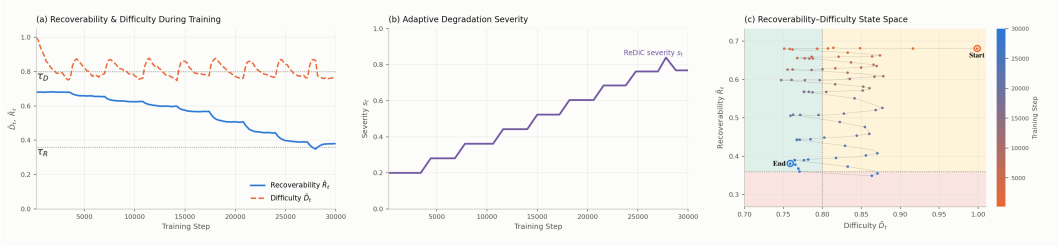

Figure 5: Training dynamics of ReDiC. (a) Recoverability $\bar{R}_t$ and relative restoration difficulty $\bar{D}_t$. (b) Degradation severity s_t. (c) The trajectory in the $(\bar{D}_t, \bar{R}_t)$ state space. Dotted lines mark $\tau_R = 0.36$ and $\tau_D = 0.80$. In (c), the green, yellow, and red regions indicate severity increase, maintenance, and reduction under the scheduling rule, respectively. Point color indicates training step.

5.6 Curriculum Dynamics

Figure 5 illustrates how ReDiC adjusts degradation severity using recoverability and relative restoration difficulty during training.

Figure 5(a) shows repeated increases in $\bar{D}_t$ around severity updates, followed by decreases during periods of constant severity. This pattern is consistent with the model adapting to newly introduced degradation conditions. Meanwhile, $\bar{R}_t$ generally declines as stronger degradations are introduced, reflecting increasing distortion in the training inputs. The two curves thus characterize different aspects of curriculum progression: the changing restoration challenge for the model and the estimated recoverability of the degraded inputs.

The severity trajectory in Figure 5(b) starts at $s_t = 0.20$ and progresses through a series of increases and plateaus. Under the scheduling rule, challenging but sufficiently recoverable conditions are maintained to allow further learning. Late in training, an increase to approximately $s_t = 0.84$ is accompanied by $\bar{R}_t$ falling below $\tau_R = 0.36$. ReDiC then reduces severity to approximately $s_t = 0.77$, after which recoverability returns above the threshold. This reversal illustrates how recoverability feedback can correct curriculum progression when stronger degradation becomes insufficiently recoverable.

Figure 5(c) places these transitions in the $(\bar{D}_t, \bar{R}_t)$ state space. High difficulty occurs in both the upper-right maintenance region and the lower reduction region, where recoverability determines whether the current degradation should be maintained or reduced. The observed increases, plateaus, and late reduction illustrate the complementary roles of the two signals: relative difficulty guides progression as the model adapts, while recoverability provides a criterion for retreat. ReDiC therefore supports reversible curriculum progression beyond a monotonic increase in degradation severity.

6 Conclusion

We studied semantic reliability in smart-glasses super-resolution using an HR-referenced evaluation protocol and showed that improved visual quality does not necessarily lead to more reliable downstream understanding. Generative SR can introduce misleading details, while restoration-oriented SR can leave critical details blurred or ambiguous. To improve detail recovery in restoration-oriented SR, we proposed ReDiC, which adapts training degradation severity using a model-independent recoverability proxy and model-relative restoration difficulty. ReDiC modifies only degradation generation during training, without changing the SR architecture or reconstruction objective. Experiments across NAFNet, SwinIR, and MambaIR demonstrate average PSNR gains of 2.36–2.65 dB over their corresponding baselines under synthetic degradations, alongside improved text recognition and HR-referenced semantic agreement. Automatic and human evaluations on 200 real smart-glasses captures further demonstrate improved downstream semantic agreement across all three backbones.

AI USE STATEMENT

We used ChatGPT to assist with manuscript preparation, including translation and language editing. The use of vision-language models and LLM judges for semantic evaluation is described in Section 3 and Appendix A. We reviewed and revised the AI-assisted content and take full responsibility for the final manuscript.

ETHICS STATEMENT

This work uses public image datasets and real smart-glasses captures to study super-resolution for visual assistance. The real captures are used exclusively for evaluation. We also conduct human evaluation to assess the semantic quality of image descriptions. All participants provided informed consent, and their responses were collected and analyzed anonymously. Restored images may contain incorrect details that affect downstream recognition and interpretation. Applications involving wearable cameras should also consider the privacy of people and information captured in the images.

REPRODUCIBILITY STATEMENT

The datasets, training settings, and evaluation metrics are described in Section 5.1. The degradation pipeline is detailed in Section 4.2, and the learning-state aggregation procedure is provided in Appendix B. Appendix A describes the semantic evaluation protocol, including reference construction, scoring rules, and judge settings. We plan to publicly release the training and evaluation code.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 1122–1131. IEEE, 2017.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pp. 17–33. Springer, 2022.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2): 295–307, 2015.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pp. 1311–1320. Pmlr, 2017.

Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *European conference on computer vision*, pp. 222–241. Springer, 2024.

Jin-Young Kim, Hyojun Go, Soonwoo Kwon, and Hyun-Gyoon Kim. Denoising task difficulty-based curriculum for training diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 64392–64405, 2025.

M Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23, 2010.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 105–114. IEEE, 2017.

Zhichao Li, Hongshan Yang, Zhibo Wang, Huiyu Xu, Junhong Lai, Yaopeng Wang, Kui Ren, and Chun Chen. On evaluating the robustness of large vision-language models via untargeted modality alignment breaking adversarial attack. In *USENIX Security*, 2026.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *2021 IEEE/CVF international conference on computer vision workshops (ICCVW)*, pp. 1833–1844. IEEE, 2021.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas Schön. Controlling vision-language models for multi-task image restoration. In *International Conference on Learning Representations*, volume 2024, pp. 16226–16246, 2024.

Long Peng, Wenbo Li, Renjing Pei, Jingjing Ren, Jiaqi Xu, Yang Wang, Yang Cao, and Zheng-Jun Zha. Towards realistic data generation for real-world super-resolution. In *International Conference on Learning Representations*, volume 2025, pp. 55077–55101, 2025.

Xinkuan Qiu, Meina Kan, Yongbin Zhou, and Shiguang Shan. Benchmarking multimodal large language models against image corruptions. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9014–9023. IEEE, 2025.

Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1557–1562. IEEE, 2019.

Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 1110–1121. IEEE, 2017.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024.

Pan Wang, Yihao Hu, and Xiujin Liu. Cogsenet: Blind image deblurring with blur-conditioned semantic routing and explicit frequency fusion. *arXiv preprint arXiv:2606.30030*, 2026.

Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. Scene text image super-resolution in the wild. In *European conference on computer vision*, pp. 650–666. Springer, 2020.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1905–1914. IEEE, 2021.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25456–25467. IEEE, 2024.

Huiyu Xu, Yaopeng Wang, Zhibo Wang, Zhongjie Ba, Wenxin Liu, Lu Jin, Haiqin Weng, Tao Wei, and Kui Ren. Profake: Detecting deepfakes in the wild against quality degradation with progressive quality-adaptive learning. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 2207–2221, 2024.

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

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4771–4780. IEEE, 2021.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Yuanzhi Zhu, Ruiqing Wang, Shilin Lu, Hanshu Yan, Junnan Li, and Kai Zhang. One-step flow for image super-resolution with tunable fidelity-realism trade-offs. In *International Conference on Learning Representations*, volume 2026, pp. 91476–91505, 2026.

A HR-REFERENCED SEMANTIC EVALUATION DETAILS

A.1 REFERENCE CONSTRUCTION AND INDEPENDENT EVALUATION

For each observation, we consider an HR reference x^{HR}, its corresponding LR image x^{LR}, and the restored output $\hat{x}^{\mathrm{SR}}$. The same fixed VLM Φ is applied independently to all three images to obtain

$$r = \Phi(x^{\mathrm{HR}}), \qquad d^{\mathrm{LR}} = \Phi(x^{\mathrm{LR}}), \qquad d^{\mathrm{SR}} = \Phi(\hat{x}^{\mathrm{SR}}). \tag{7}$$

The HR-derived description r provides a common semantic reference for examining content omitted, recovered, or altered in downstream interpretations. Applying the same interpreter to all image variants reduces variation due to interpreter choice.

We evaluate d^{LR} and d^{SR} separately against r, rather than asking a judge to choose between them. Both candidates are therefore assessed under the same reference-based rubric without an explicit pairwise preference prompt. The question-answering example in Fig. 4 serves as a qualitative illustration; the quantitative scores are computed from image descriptions.

A.2 SCORING DIMENSIONS AND AGGREGATION

Lexical similarity alone cannot capture semantic agreement: a description may mention the correct objects while assigning an incorrect number or attribute. We therefore assess four complementary dimensions.

Semantic consistency evaluates whether the key facts in a candidate description agree with the HR-derived reference.

Information coverage evaluates how much of the semantic content expressed in the reference is retained in the candidate.

Factual fidelity evaluates contradictions and claims unsupported by the reference, providing a reference-based indicator of potential semantic hallucination.

Attribute alignment evaluates agreement on fine-grained properties expressed in the reference, including text, numbers, colors, and spatial relations.

For a candidate description d and reference r, judge m assigns each dimension a score in $[0, 100]$, with higher scores indicating better agreement. Its aggregate score is

$$\begin{aligned} S^{(m)}(d, r) = 0.4 S^{(m)}_{\mathrm{cons}}(d, r) + 0.25 S^{(m)}_{\mathrm{cov}}(d, r) \\ + 0.25 S^{(m)}_{\mathrm{fid}}(d, r) + 0.1 S^{(m)}_{\mathrm{rel}}(d, r). \end{aligned} \tag{8}$$

We average these scores across M judges:

$$S(d, r) = \frac{1}{M} \sum_{m=1}^{M} S^{(m)}(d, r). \tag{9}$$

The semantic gain ΔS is defined in Eq. 1. A positive value indicates improved reference agreement after SR, whereas a negative value indicates reduced agreement relative to the LR input.

A.3 JUDGE IMPLEMENTATION AND INTERPRETATION

We use three distinct LLM judges, setting $M = 3$: GLM-5, MiniMax-M2.1, and Qwen3.5-397B-A17B. All judges receive the same evaluation instructions and structured output format and use low-temperature decoding. We collect the dimension-wise scores and compute the weighted sums and cross-judge average externally, rather than using a model-generated overall score. This ensures that the reported scores follow the same aggregation rule for every judge.

The HR-derived reference is a model-generated interpretation rather than manually verified ground truth. It may omit visible information or contain errors, and a claim absent from the reference is not necessarily false in the image. Accordingly, S and ΔS measure semantic agreement under the fixed interpreter and judging protocol; they do not directly establish the factual correctness of every visual detail or guarantee reliable downstream behavior.

B Learning-State Aggregation

This section details how the patch-level recoverability and relative difficulty scores defined in Sec. 4.3 are aggregated into the state used for curriculum scheduling.

B.1 Within-Image Aggregation

Let $\mathcal{P}_j$ denote the set of sampled patches from image j. To summarize the spatially heterogeneous patch-level scores, we compute

$$R_j = Q_{q_R}\big(\{R_i\}_{i\in\mathcal{P}_j}\big), \qquad D_j = Q_{q_D}\big(\{D_i\}_{i\in\mathcal{P}_j}\big), \tag{10}$$

where $Q_q(\cdot)$ denotes the q-th quantile operator, and q_R and q_D specify the quantiles used for recoverability and relative difficulty, respectively.

B.2 Window-Level Aggregation

For curriculum window t containing N_t images, we average the image-level scores:

$$\widetilde{R}_t = \frac{1}{N_t}\sum_{j=1}^{N_t} R_j, \qquad \widetilde{D}_t = \frac{1}{N_t}\sum_{j=1}^{N_t} D_j. \tag{11}$$

This yields a window-level estimate of the current degradation condition and restoration difficulty.

B.3 Temporal Smoothing

To reduce short-term fluctuations, we maintain exponential moving averages:

$$\begin{aligned} \bar{R}_t &= \beta\bar{R}_{t-1} + (1-\beta)\widetilde{R}_t, \\ \bar{D}_t &= \beta\bar{D}_{t-1} + (1-\beta)\widetilde{D}_t, \end{aligned} \tag{12}$$

where $\beta \in [0, 1)$ is the smoothing coefficient. The resulting state $(\bar{R}_t, \bar{D}_t)$ is passed to the curriculum scheduler to determine whether degradation severity should increase, remain unchanged, or decrease.

C Training and Degradation Settings

NAFNet, SwinIR, and MambaIR are initialized from their respective pretrained $\times 2$ checkpoints and use the same fine-tuning configuration. The training degradation pipeline includes blur, noise, compression, and resampling, with operator parameters sampled according to the current curriculum severity. The SR scale remains fixed at $\times 2$ throughout training. We apply degradations to full training images and then extract spatially aligned LR–HR patch pairs of size 64×64 and 128×128, respectively. The extracted patch pairs are shuffled and augmented with the same rotation applied to both members of each pair. All three backbones are fine-tuned with a batch size of 16.