SPIDER: Boosting Blind Face Restoration via Simultaneous Prior Injection and Degradation Removal

Anonymous Author(s)
Affiliation
Address
email

Abstract

Existing blind face restoration (BFR) methods suffer from drastic performance drop under severe degradations. A common strategy is to first remove degradations and then restore the face by fully harnessing generative prior. However, this sequential pipeline risks discarding subtle but crucial cues from already limited low-quality (LQ) inputs. To address this, we ingeniously introduce a new learning paradigm: simultaneous prior injection and degradation removal (SPIDER). Unlike existing approaches, SPIDER injects semantic prior before degradation removal, thereby preserving identity-relevant features and mitigating the impact of corrupted LQ features. SPIDER consists of two key modules: (1) a prior injection module that distills purified degradation-unaware semantic control tokens from vision-language models, and (2) a degradation removal module equipped with an image-to-text degradation mapper and a degradation remover that refines distorted features into robust representations. Extensive experiments on both synthetic and real-world datasets, including challenging surveillance scenarios, demonstrate SPIDER's clear superiority over state-of-the-art BFR methods.

1 Introduction

Blind Face Restoration (BFR) is a challenging task that aims to recover high-quality (HQ) face images from low-quality (LQ) ones that suffer from unknown and complex degradations such as low resolution [3, 6], blur [48], noise [13, 28], and JPEG compression [5]. This is an inherently ill-posed problem as the information loss caused by the degradations leads to an overwhelming number of plausible HQ solutions consistent with the same LQ input. To mitigate the ill-posedness, recent studies have explored various prior-based methods to produce high-fidelity outputs.

As illustrated in Figure 1, existing prior-based BFR methods fall into three main paradigms: 1) Continuous generative prior (e.g., GFPGAN [36], which learns accurate latent codes via GAN inversion to reconstruct HQ faces with high fidelity; 2) Discrete generative prior (e.g., Codeformer [52], DAEFR [30]), which uses vector quantization to map degraded inputs into semantic tokens and harness a fixed HQ codebook for high-quality restoration; and 3) Diffusion-based conditional generation (e.g., DiffBIR [18], FaithDiff [2]), which reframes the restoration into conditional generation employing the powerful expressiveness of diffusion prior to achieve significant improvements in fine detail, perceptual fidelity, and overall realism. Many SOTA methods [18, 2, 37, 41] belong to the third paradigm and achieve promising results on mild to moderate degradations. However, under severe or extreme degradation, whether synthetic or real-world, they often introduce artifacts or even fail catastrophically in the results, as demonstrated in Figure 2. Taking extreme surveillance degradations in the fourth row as an example, since aliasing and jagged artifacts are not presented in the synthesized training data, the existing models mistakenly use the corrupted signals as the actual

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

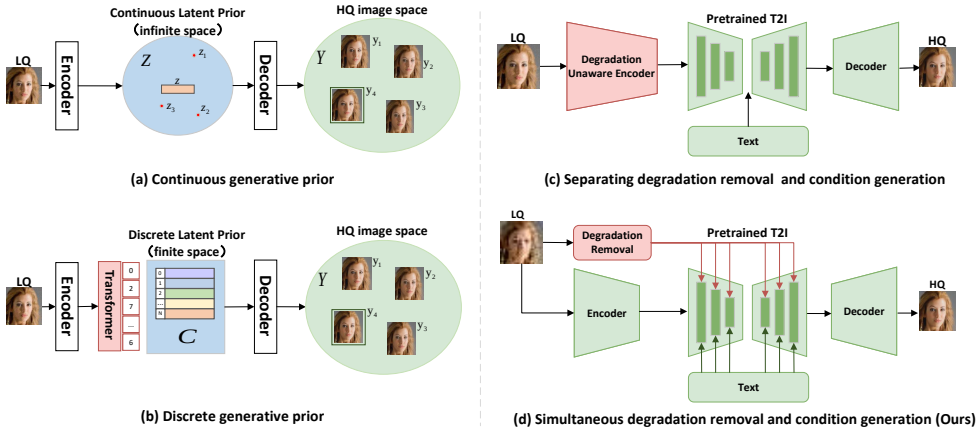

Figure 1: Comparison with existing paradigms for blind face restoration.

features, resulting in erroneous details in the results. We argue that the noises induced by the severe degradations are the primary cause of restoration failure. To address this, many methods [18, 37] first remove degradations explicitly or implicitly and then restore the face by leveraging powerful generative priors (Figure 1 (c)). However, this sequential pipeline risks discarding subtle but crucial cues from already limited LQ inputs, further elevating the ill-posedness of BFR.

To overcome this, we ingeniously propose Simultaneous Prior Injection and Degradation Removal (SPIDER), a new learning paradigm to enhance face restoration (Figure 1 (d)). Rather than removing degradation first, SPIDER injects semantic prior (i.e., combined with diffusion prior) before degradation removal. Intuitively, this design not only enriches the representation of relevant facial content but also amplifies both signal and noise. This amplification enables the subsequent degradation removal module (DRM) to more effectively differentiate between informative structures and unwanted noise, resulting in substantially improved restoration fidelity (Figure 2).

Specifically, SPIDER consists of two key components. The Prior Injection Module distills degradation-unaware semantic tokens using a vision-language model (VLM), such as LLaVA [19], to generate rich textual descriptions from degraded images. These semantic priors are subsequently injected into the diffusion generation pipeline at multiple levels, providing robust and context-aware guidance. The DRM comprises an image-to-text degradation mapper and a degradation remover, which together project noisy visual representations into a purified textual embedding space aligned with the injected prior. This design leverages the noise-resilience of the textual modality and performs degradation filtering through semantic alignment, which is more robust to perturbations than direct visual-space restoration. By jointly integrating semantic prior injection and degradation removal via our proposed decoupled cross-attention (DCA) mechanism, SPIDER delivers state-of-the-art restoration results under severe degradations in both synthetic and real-world scenarios.

SPIDER achieves state-of-the-art results on both existing synthetic and real-world benchmarks and our newly introduced SCface dataset [8] of extreme surveillance face images. Beyond its superior BFR performance, SPIDER pioneers a novel learning paradigm *injecting prior before degradation removal* that can be extended to a wide range of restoration tasks beyond blind face restoration.

2 Related Work

2.1 Blind Face Restoration

Recent BFR approaches mainly leverage generative prior to reconstruct faces with high realism and faithful details. Representative latent-prior-based methods such as GFPGAN [36] and GPEN [44] encode LQ face images into semantically faithful latent codes, enabling faithful reconstruction of their HQ counterparts using StyleGAN-based generative prior [11]. Despite improvements in fidelity, these methods often introduce artifacts when the input images exhibit complex degradations not covered by the training data. State-of-the-art methods like Codeformer [52], RestoreFormer [39],

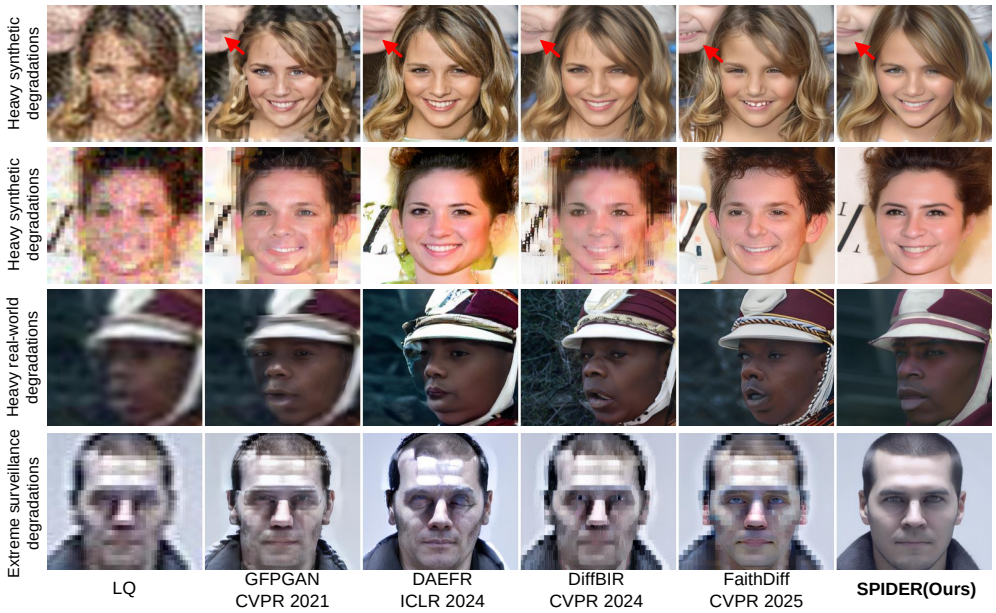

Figure 2: Comparisons with representative face restoration approaches on both synthetic (CelebA-Test [10]) and real-world (WIDER [42], SCface [8]) LQ images under various degradations

.

and DAEFR [30] utilize discrete HQ codebook to generate high-fidelity face details by exploiting Vector-Quantized (VQ) dictionary learning [7, 31]. However, the fixed-size codebook inherently limits the expressiveness ability of such discrete prior, which can hinder the faithful reconstruction of diverse and complex facial structures.

Recent works [45, 46, 34, 18, 2] reframe face restoration as conditional image generation using powerful diffusion prior, significantly advancing BFR quality. VSP [22] introduces prompt-based inference to further refine restoration results. StableSR [34] and TASR [15] finetune the temporal embedding layers to produce time-aware features that adaptively modulate features across the denoising steps. DiffBIR [18] and DR2 [37] perform degradation removal and conditional image generation sequentially: they first remove degradation using an off-the-shelf model, and then refine details. PASD [45] and SUPIR [46] enhance LQ feature extraction with stronger encoders. The latest work FaithDiff [2], employs BSRNet [47] for initial restoration and extracts text embeddings via LLaVA [19]. It further improves this paradigm by jointly training the encoder and diffusion model in an end-to-end fashion, enabling their synergistic evolution and enhancing alignment between the extracted features and the generated content.

Although the above methods have demonstrated strong performance in restoring faces under moderate degradations, they often struggle in real-world scenarios involving severe and complex degradations. This results in visual artifacts, structural distortions, and semantic inconsistencies. A key challenge lies in the model's difficulty in distinguishing intrinsic, reliable facial features from degradation-induced noise. Consequently, synthesizing faces from corrupted or noisy representations can lead to erroneous or unrealistic restoration outcomes. Therefore, effectively removing degradations is a prerequisite for achieving faithful and high-quality restoration.

2.2 Degradation Removal in Blind Image Restoration

Recent blind image restoration methods increasingly focus on learning degradation processes to enhance realism and adaptability. Due to the limitations of handcrafted degradation assumptions, AND [35] introduces an adversarial degradation generator that synthesizes pseudo-degraded images, thereby bridging the domain gap between synthetic and real-world degradations in supervised restoration. DiffBIR [18] and FaithDiff [2] both adopt a two-stage design, where degradation

is first removed and then image quality is refined. TextualDegRemoval [17] leverages textural modality representations to generate clean guidance images under natural degradations such as rain and snow, using them as reference images to enhance blind image restoration. However, despite differences in representation and guidance modality, these methods share a structurally decoupled architecture: degradation removal is treated as a prerequisite, independent of the image generation process. This decoupling hinders joint optimization and frequently leads to unstable outputs when handling occlusion, motion blur, or structural corruption—especially in face restoration, where identity consistency and semantic fidelity are particularly fragile.

2.3 Vision-Language Models

Vision-language models (VLMs) have advanced rapidly with CLIP [26] providing strong semantic prior by aligning image and text embeddings. DA-CLIP [23] models image content and degradations jointly, enabling multi-task restoration. SSP-IR [51] enhances geometric consistency by integrating structural contour information into CLIP-based prior, while FUSION [21] improves cross-modal understanding through deep feature fusion. Apart from general VLMs, researchers have also proposed vision-language architectures specifically designed for face images. FCLIP [4] uses dual-branch learning on FaceCaption-15M for better attribute alignment. Face-MLLM [29] employs a three-stage strategy on a large-scale QA dataset to enhance fine-grained attribute reasoning and instruction following. FaceInsight [14] integrates keypoint detection and attention mechanisms to ensure structural and identity consistency. However, these methods remain unpublished or proprietary, hindering the integration of face-oriented VLMs into BFR task.

3 Proposed Method: SPIDER

3.1 Framework overview

As illustrated in Figure 3, our SPIDER restores HQ face images from their LQ counterparts by simultaneously injecting cross-modal semantic prior and removing degradations. The training process of SPIDER is divided into two stages. In Stage I (Figure 3(b)), we train a degradation removal module (DRM) to remove degradations at the textual level, where the image content and degradation information are loosely coupled, making it more effective to isolate and remove noise. In Stage II (Figure 3(a)), we employ a large vision-language model (i.e., LLaVA) to generate detailed text descriptions of HQ face images, enabling the extraction of fine-grained semantic prior. Meanwhile, the pretrained DRM is used to "erase" feature corruption at multiple scales. These two branches (i.e., prior injection and degradation removal) interact via a decoupled cross attention (DCA) mechanism (Figure 3(c)) integrated into each block of both the UNet [27] and ControlNet [49]. This collaborative design ensures robust guidance and effective noise suppression, ultimately leading to faithful HQ face restoration.

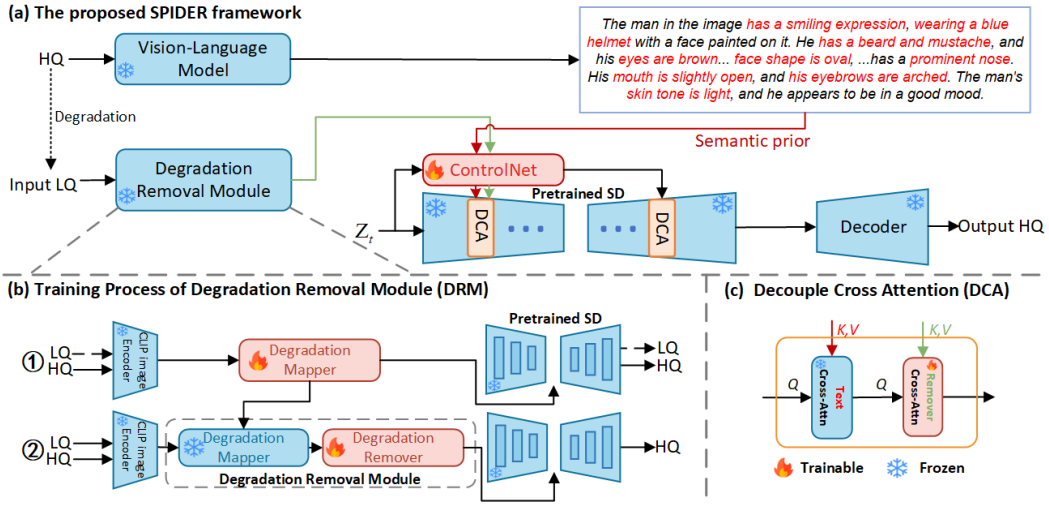

Figure 3: Framework of our proposed SPIDER model.

3.2 Semantic Prior Injection

Many recent super-resolution models leverage prior knowledge provided by vision-language models, such as FaithDiff [2], XPSR [25], and AuthFace [16]. The performance of LLaVA is highly dependent on well-crafted prompts [20]. Considering the highly structured nature of human faces, we modify the prompt design to focus the output on detailed facial component structures. Experiments in XPSR [25] and SPIRE [24] demonstrate that LLaVA can implicitly infer both the type and severity of noise artifacts from facial images, maintaining robust performance even under noisy conditions. In our study, we extract structural descriptions of facial features from LLaVA and encode them using CLIP text encoder to obtain embeddings and then further integrate these embeddings with DRM outputs through a DCA module to guide the image restoration process.

3.3 Degradation Removal Module (DRM)

Effective degradation removal is essential for recovering high-fidelity HQ face images from LQ inputs, as corrupted signals can mislead the blind face restoration (BFR) model. To address this, we propose a novel cross-modal mapping module, DRM, which directly transforms degraded face images into noise-suppressed textual representations. This cross-modal strategy is more effective than conventional image-space denoising, as the textual embedding space exhibits a natural decoupling between semantic content and degradation-induced noise.

As shown in Figure 3, the proposed DRM comprises two key components: (1) a Degradation Mapper that projects CLIP image embeddings into implicit textual representations, preserving rich visual semantics that are often lost in explicit textual descriptions; and (2) a Degradation Remover that purifies these representations by filtering out degradation-specific artifacts. The resulting clean textual features maintain high semantic fidelity to the original image content while eliminating noise patterns, thereby providing reliable guidance for subsequent image generation.

Degradation Mapper. Following [40], we use a CLIP-based cross-modal projection that maps visual features into a text-aligned embedding space. Specifically, given a degraded input image X, we first extract its visual features using a CLIP image encoder E, and then project them into the textual embedding space through a learnable Mapper $\mathcal{P}_{\text{mapper}}$:

$$F_{\text{mapper}} = \mathcal{P}_{\text{mapper}}(E(X)), \quad F_{\text{mapper}} \in \mathbb{R}^{N \times D}, \tag{1}$$

where D is the dimensionality of the textual word embeddings, and N is the number of learned tokens (set to 30) to preserve rich visual details while maintaining computational efficiency.

Degradation Remover. While F_{mapper} encodes high-level textual representations, it also carries noise and degradation-specific artifacts that hinder restoration. To address this, we introduce a Degradation Remover $\mathcal{P}_{\text{remover}}$ to purify the token embeddings:

$$F_{\text{remover}} = \mathcal{P}_{\text{remover}}(F_{\text{mapper}}), \quad F_{\text{remover}} \in \mathbb{R}^{N \times D}, \tag{2}$$

where N is consistent with the Mapper design and the cleaned representation F_{remover} serves as the final conditioning input to the DCA module (Figure 3(c)).

3.4 Training and Inference

Training Stage I: Learning Degradation-Unaware Textual Representations. The Degradation Mapper projects CLIP image features into a text-aligned embedding space, generating a representation F_{mapper} that captures both visual content and degradation patterns in a form compatible with text embeddings. As a result, images of different qualities are encoded into a unified textual representation space. During the training of the Degradation Mapper, the condition F is replaced with F_{mapper} and the training objective is defined as:

$$\mathcal{L}_{\text{stageI}} = \mathbb{E}_{z_0, \epsilon \sim \mathcal{N}(0,I), t} \left[\|\epsilon - \epsilon_\theta(z_t, t, F)\|_2^2 \right], \tag{3}$$

where $\mathbf{z}_0 = \mathcal{E}(X)$ is the latent representation of input image X, encoded by a pretrained VAE encoder $\mathcal{E}(\cdot)$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ represents the added Gaussian noise. At each diffusion timestep t, the noisy latent representation $\mathbf{z}_t$ is constructed from $\mathbf{z}_0$ and $\boldsymbol{\epsilon}$ via the forward diffusion process. The noise prediction model $\epsilon_\theta(\cdot)$ is trained to predict the added noise by minimizing the mean squared error between the ground-truth and predicted noise values. The Degradation Remover is trained using the same diffusion loss $\mathcal{L}_{\text{stageI}}$, but with the condition F replaced by F_{remover}, while keeping the Mapper module frozen. Additional training details are provided in the appendix.

Training Stage II: Simultaneous Prior Injection and Degradation Removal for BFR. After finishing training Degradation Mapper and Degradation Remover, we freeze their weights, and train the blind face restoration model using simultaneous prior injection and degradation removal across multiple feature scales. Specifically, two complementary sources of information are utilized: (1) a high-level semantic prior F_{text} obtained from LLaVA, and (2) a degradation-aware, noise-suppressed embedding F_{remover} produced by the previous stage. They are simultaneously injected into the diffusion model to enable semantically coherent and visually faithful face reconstruction. The training objective follows the standard noise prediction loss used in latent diffusion models:

$$\mathcal{L}_{\text{stageII-BFR}} = \mathbb{E}_{z_0,\epsilon\sim\mathcal{N}(0,I),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, DCA(F_{\text{text}}, F_{\text{remover}}))\|_2^2\right], \tag{4}$$

where $\mathrm{DCA}(\cdot)$ denotes the Decoupled Cross Attention mechanism.

Inference Pipeline. During inference, our framework processes an LQ face image through two parallel paths: the input is first passed through the frozen Degradation Mapper and Remover to obtain a degradation-purified embedding F_{remover}; simultaneously, the same LQ input is used to generate a text embedding F_{text} via LLaVA [19]. To ensure high-quality text generation, we preprocess the input with GFPGAN-v1.4 [36] prior to feeding it into LLaVA-v1.5-13B [19].

4 Experiments

4.1 Experimental Settings

Training Datasets. We train all models on the FFHQ dataset [11], which contains 70,000 high-quality face images resized to 512×512. We follow DiffBIR [18] to generate corresponding low-quality (LQ) images. To obtain semantic guidance, we use LLaVA [19] to generate face description prompts for each image.

Testing Datasets. We evaluate our method on one synthetic dataset and three real-world datasets. The synthetic dataset, CelebA-Test [10], comprises 3,000 images sampled from CelebA-HQ, with LQ versions generated using the same degradation pipeline as training. The real-world datasets include LFW-Test [9] (1,711 images, mild degradation), WIDER-Test [42] (970 images, heavy degradation), and SCface [8], which features extreme surveillance degradations such as aliasing, jagged artifacts, low-light noise, and defocus blur. SCface contains 130 subjects captured by five surveillance cameras; for evaluation, we use images taken from a distance of 2.6 meters.

Compared Methods. We compare our proposed SPIDER with seven state-of-the-art BFR methods across three categories: (1) StyleGAN-prior methods: GFPGAN [36]; (2) Codebook-prior methods: CodeFormer [52] and DAEFR [30]; (3) Diffusion-based methods: DR2 [37], PGDiff [41], DiffBIR [18] and FaithDiff [2]. All experiments are conducted using official code.

Evaluation Metrics. We adopt both reference-based and no-reference image quality metrics for comprehensive evaluation. For synthesized datasets with ground truth, we use PSNR [38], SSIM [38], LPIPS [50], FID [1], as well as no-reference perceptual metrics MANIQA [43] and CLIPIQA [33] to better capture perceptual quality. For real-world datasets without ground truth, we evaluate performance using MANIQA, CLIPIQA, and FID.

Implementation Details Both the Degradation Mapper and Remover are trained for one epoch on the FFHQ dataset during Stage I, with a batch size of 4, using the Adam optimizer and a learning rate of 1×10^{-6} on a single NVIDIA V100 GPU. Our SPIDER model is built upon the pretrained stable-diffusion-2.1 and trained for 15 epochs on two NVIDIA RTX 4090 GPUs, with a batch size of 192, using the Adam optimizer [12] and a learning rate of 5×10^{-5}.

4.2 Comparisons with State-of-the-Art Methods

Table 1: Quantitative comparison on synthetic dataset of CelebA-Test [10]. The best results are marked in red, and the second best in blue.

Metrics	Input	GFPGAN [36]	CodeFormer [52]	DAEFR [30]	DR2 [37]	PGDiff [41]	DiffBIR [18]	FaithDiff [2]	**SPIDER(Ours)**
FID ↓	152.64	21.33	22.58	15.55	27.75	19.82	28.48	20.92	22.57
MANIQA ↑	0.1683	0.4289	0.5062	0.5426	0.5160	0.4658	0.6534	0.5184	0.5834
CLIPIQA ↑	0.2403	0.5391	0.6828	0.6769	0.5972	0.5583	0.7648	0.6570	0.7013
LPIPS ↓	0.7292	0.4554	0.3312	0.4153	0.3354	0.3286	0.3882	0.3145	0.3177
PSNR ↑	22.56	17.86	22.67	21.89	22.26	21.52	22.90	22.76	22.94
SSIM ↑	0.5006	0.5407	0.5540	0.5966	0.5854	0.5678	0.5410	0.5782	0.6111

Figure 4: Qualitative comparison on synthetic dataset of CelebA-Test.

Evaluation on Synthetic Dataset. As shown in Table 1, our method achieves state-of-the-art or near-state-of-the-art performance across multiple evaluation metrics on CelebA-Test, ranking first in PSNR and SSIM, and second in LPIPS, MANIQA, and CLIPIQA. Figure 4 further demonstrates that our approach generates perceptually realistic and semantically consistent face restorations, with well-preserved identity features and effective suppression of visual artifacts. In the first-row example, clearer eye contours and finer skin details are recovered, while in the second-row case, the original gaze direction and eye semantics are faithfully maintained—corroborating the superior quantitative performance.

Figure 5: Qualitative comparisons on real-world datasets. Our method is able to restore high quality faces, showing robustness to the heavy degradation.

Evaluation on Real-World Dataset. As shown in Table 2, our method achieved the best FID scores on the LFW and the WIDER dataset. It also ranked second on both the MANIQA and CLIPIQA benchmarks. For the SCface dataset, which involves extreme surveillance degradations, our approach attained the highest MANIQA and CLIPIQA scores. Although the FID score was slightly inferior to that of PGDiff, our method outperformed it in all other evaluation metrics. Figure 5 presents qualitative results on real-world datasets. On LFW, our model effectively restores side faces with

clear facial features and minimal background-induced artifacts. Benefiting from the multi-step noise suppression of the DRM module, SPIDER also preserves both primary and background subjects with high fidelity. On WIDER, it maintains key facial features (eyes, nose, mouth) and overall visual consistency, whereas other models suffer from artifacts around facial regions. On SCface, the complex noise patterns inherent in surveillance imagery pose significant challenges for existing models, leading to suboptimal reconstructions. In contrast, our model delivers more faithful and coherent results, demonstrating stronger generalization to real-world degradations.

Table 2: Quantitative comparisons on real-world datasets of LFW [9], WIDER [42], SCface [8]. The best results are marked in red, and the second best in blue.

Datasets	LFW			WIDER			SCface		
Degradation	Mild real-world degradations			Heavy real-world degradations			Extreme surveillance degradations		
Methods	FID ↓	MANIQA ↑	CLIPIQA ↑	FID ↓	MANIQA ↑	CLIPIQA ↑	FID ↓	MANIQA ↑	CLIPIQA ↑
GFPGAN [36]	53.87	0.4558	0.6324	50.36	0.4352	0.5756	106.97	0.4358	0.6461
CodeFormer [52]	52.84	0.5266	0.6889	39.22	0.4959	0.6984	99.07	0.4327	0.6852
DAEFR [30]	47.69	0.5420	0.6965	36.72	0.5205	0.6975	103.64	0.4600	0.7217
DR2 [37]	50.42	0.5248	0.6532	52.78	0.4746	0.5948	96.49	0.4438	0.5823
PGDiff [41]	41.86	0.4763	0.6070	38.06	0.4391	0.5880	85.48	0.3610	0.5127
DiffBIR [18]	40.91	0.6735	0.7948	35.82	0.6624	0.8083	149.98	0.3965	0.4648
FaithDiff [2]	41.34	0.4949	0.6787	36.07	0.5106	0.7092	88.48	0.5127	0.6880
SPIDER(Ours)	39.74	0.5784	0.7320	34.58	0.5630	0.7342	87.57	0.5514	0.7446

4.3 Ablation studies

Effectiveness of the Degradation Removal Module. As shown in Table 3, integrating DRM results in a clear performance improvement across all metrics. A visual comparison on WIDER-Test is presented in Figure 6 (a). In the first and third columns, without DRM, the BFR model misinterprets noise as authentic detail, leading to severe distortions in the hair, facial region, clothing and background. In the second and fourth columns, unclear contours and artifacts are generated because the BFR model struggles to effectively distinguish between noise and informative content. These results suggest that the module effectively suppresses noise while preserving fine-grained structural and textural details, thereby enhancing overall image fidelity.

Table 3: Ablation results showing the effectiveness of the DRM on the CelebA-Test.

Metrics	Without DRM	With DRM
PSNR↑	20.76	22.94
SSIM↑	0.5671	0.6111
LPIPS↓	0.3910	0.3177
FID↓	24.94	22.57

Insights into SPIDER Design. Table 4 shows that the order of the semantic prior injection and the DRM is critical. We adopt a design where the semantic prior is injected before DRM, allowing the diffusion model to amplify both signal and noise. This amplification helps DRM better distinguish informative structures from degradation, leading to improved restoration quality. As shown in Figure 6(b), applying DRM first may suppress useful details, resulting in a generated image that lacks detail and exhibits unnatural textures.

Table 4: Ablation results comparing different module orders on the CelebA-Test.

Metrics	DRM→Semantic prior	Semantic prior→DRM
PSNR↑	21.50	22.94
SSIM↑	0.5988	0.6111
LPIPS↓	0.3550	0.3177
FID↓	27.54	22.57

Importance of Face-oriented Prompt Design. Our method customizes the prompts fed into LLaVA [19] based on CelebA-Test facial attribute definitions [10], enabling richer semantic descriptions of facial structures. According to Table 5, compared to general image descriptions, incorporating detailed facial descriptions can enhance the performance of restoration. Thus, guiding the vision-language model to attend to facial attributes is essential for effectively leveraging its prior knowledge. However, facial attributes prompts often omit spatial information and frequently include redundant features (e.g., "normal nose," "average eyes") that are shared across many images, offering limited benefit for recovering fine textures. More prompt examples are in the appendix.

Table 5: Ablation results comparing different prompt styles on the CelebA-Test.

Metrics	General Prompt	Facial Attributes Prompt	Face Description Prompt
PSNR↑	21.38	21.42	22.94
SSIM↑	0.5899	0.5854	0.6111
LPIPS↓	0.3616	0.3744	0.3177
FID↓	24.57	30.50	22.57

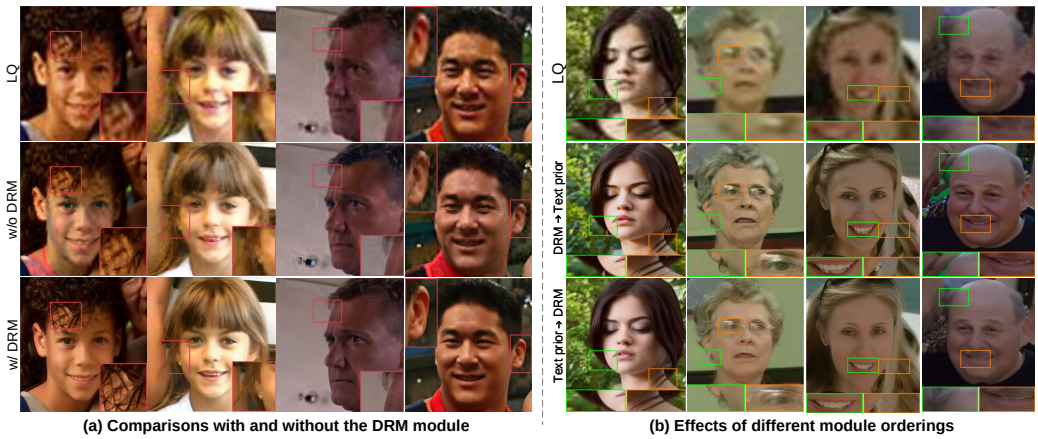

Figure 6: Ablation studies evaluating (a) the impact of the DRM module and (b) the ordering of modules within the model architecture.

4.4 Understanding the Degradation Removal Mechanism in SPIDER

We use t-SNE [32] to visualize feature changes after training the Degradation Mapper and Remover in SPIDER. We randomly select 100 CelebA-Test images (seed=42), synthesize their LQ versions at FID 150 and 250 using the same degradation pipeline as training, and apply the same process to training data (FFHQ) for reference. As shown in Figure 7, we observe that (1) features extracted by the Mapper are widely dispersed, with HQ and LQ clearly separated, indicating strong degradation sensitivity; (2) after applying the Remover, features across all quality levels converge into a degradation-invariant space, where even severely degraded images align closely with HQ ones. These consistent patterns across both training and testing datasets demonstrate that our DRM effectively suppresses degradation-related variations, enabling robust and faithful restoration under extreme degradations.

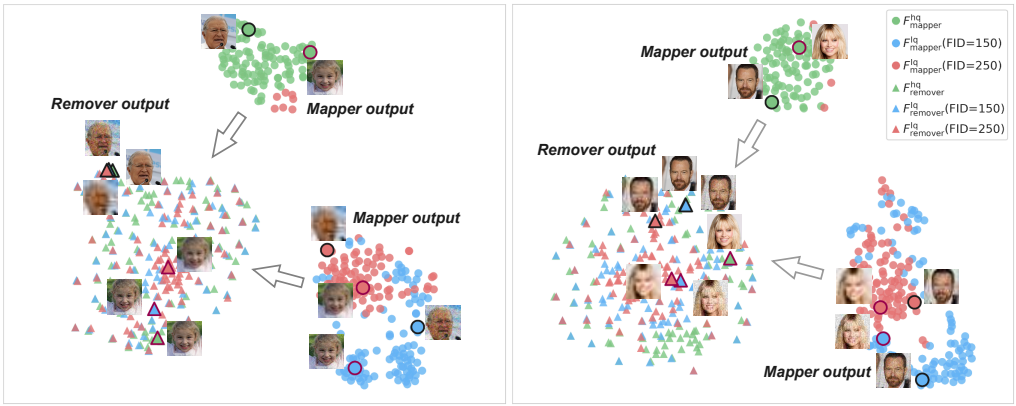

Figure 7: t-SNE visualization of feature distributions on FFHQ (left) and CelebA-HQ (right).

5 Conclusion

We propose SPIDER, a novel paradigm for blind face restoration that performs simultaneous multi-level prior injection and degradation removal. SPIDER adopts an interleaved architecture where prior injection precedes degradation removal at each level, ensuring that semantic and diffusion priors amplify both signal and noise in a way that helps the DRM better distinguish meaningful facial features from heavy degradations. Experiments on synthetic and real-world datasets confirm its superiority over state-of-the-art methods. Beyond its strong performance in BFR, SPIDER presents a novel learning paradigm with broad applicability to various image restoration tasks.

References

[1] N. Bynagari. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Asian Journal of Applied Science and Engineering*, 2019.

[2] Junyang Chen, Jinshan Pan, and Jiangxin Dong. Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2025.

[3] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 2492–2501, 2018.

[4] Dawei Dai, YuTang Li, YingGe Liu, Mingming Jia, Zhang YuanHui, and Guoyin Wang. 15m multimodal facial image-text dataset. *arXiv preprint arXiv:2407.08515*, 2024.

[5] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, pages 576–584, 2015.

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014.

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 12873–12883, 2021.

[8] Mislav Grgic, Kresimir Delac, and Sonja Grgic. Scface–surveillance cameras face database. *Multimedia Tools and Applications*, 51:863–879, 2011.

[9] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on Faces In'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.

[10] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Proceedings of the International Conference on Learning Representations(ICLR)*, 2018.

[11] 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)*, 2019.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations(ICLR)*, 2015.

[13] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018.

[14] Jingzhi Li, Changjiang Luo, Ruoyu Chen, Hua Zhang, Wenqi Ren, Jianhou Gan, and Xiaochun Cao. Faceinsight: A multimodal large language model for face perception. *arXiv preprint arXiv:2504.15624*, 2025.

[15] Litao Li, Rylen Sampson, Steven HH Ding, and Leo Song. Tasr: Adversarial learning of topic-agnostic stylometric representations for informed crisis response through social media. *Information Processing & Management*, 59(2):102857, 2022.

[16] Guoqiang Liang, Qingnan Fan, Bingtao Fu, Jinwei Chen, Hong Gu, and Lin Wang. Authface: Towards authentic blind face restoration with face-oriented generative diffusion prior. *arXiv preprint arXiv:2410.09864*, 2024.

[17] Jingbo Lin, Zhilu Zhang, Yuxiang Wei, Dongwei Ren, Dongsheng Jiang, Qi Tian, and Wangmeng Zuo. Improving image restoration through removing degradations in textual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 2866–2878, 2024.

[18] 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 *Proceedings of the European Conference on Computer Vision(ECCV)*, pages 430–448. Springer, 2024.

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances In Neural Information Processing Systems(NeurIPS)*, 36:34892–34916, 2023.

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances In Neural Information Processing Systems(NeurIPS)*, 36:34892–34916, 2023.

[21] Zheng Liu, Mengjie Liu, Jingzhou Chen, Jingwei Xu, Bin Cui, Conghui He, and Wentao Zhang. Fusion: Fully integration of vision-language representations for deep cross-modal understanding. *arXiv preprint arXiv:2504.09925*, 2025.

[22] Wanglong Lu, Jikai Wang, Tao Wang, Kaihao Zhang, Xianta Jiang, and Hanli Zhao. Visual style prompt learning using diffusion models for blind face restoration. *Pattern Recognition*, 161:111312, 2025.

[23] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for multi-task image restoration. In *Proceedings of the International Conference on Learning Representations(ICLR)*, 2024.

[24] Chenyang Qi, Zhengzhong Tu, Keren Ye, Mauricio Delbracio, Peyman Milanfar, Qifeng Chen, and Hossein Talebi. Spire: Semantic prompt-driven image restoration. In *Proceedings of the European Conference on Computer Vision(ECCV)*, 2025.

[25] Yunpeng Qu, Kun Yuan, Kai Zhao, Qizhi Xie, Jinhua Hao, Ming Sun, and Chao Zhou. Xpsr: Cross-modal priors for diffusion-based image super-resolution. In *Proceedings of the European Conference on Computer Vision(ECCV)*, pages 285–303. Springer, 2024.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning(ICML)*, pages 8748–8763. PmLR, 2021.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[28] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 8260–8269, 2018.

[29] Haomiao Sun, Mingjie He, Tianheng Lian, Hu Han, and Shiguang Shan. Face-mllm: A large face perception model. *arXiv preprint arXiv:2410.20717*, 2024.

[30] Yu-Ju Tsai, Yu-Lun Liu, Lu Qi, Kelvin CK Chan, and Ming-Hsuan Yang. Dual associated encoder for face restoration. In *Proceedings of the International Conference on Learning Representations(ICLR)*, 2024.

[31] Aaron Van Den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *Advances In Neural Information Processing Systems(NeurIPS)*, 30, 2017.

[32] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[33] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence(AAAI)*, volume 37, pages 2555–2563, 2023.

[34] 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(IJCV)*, 132(12):5929–5949, 2024.

[35] 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)*, pages 10581–10590, 2021.

[36] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 9168–9178, 2021.

[37] Zhixin Wang, Ziying Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 1704–1713, 2023.

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[39] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 17512–17521, 2022.

[40] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, pages 15943–15953, 2023.

[41] Peiqing Yang, Shangchen Zhou, Qingyi Tao, and Chen Change Loy. Pgdiff: Guiding diffusion models for versatile face restoration via partial guidance. *Advances In Neural Information Processing Systems(NeurIPS)*, 36:32194–32214, 2023.

[42] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 5525–5533, 2016.

[43] 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)*, pages 1191–1200, 2022.

[44] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 672–681, 2021.

[45] 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)*, pages 74–91. Springer, 2024.

[46] 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)*, pages 25669–25680, 2024.

[47] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, pages 4791–4800, 2021.

[48] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

[49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, pages 3836–3847, 2023.

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 586–595, 2018.

[51] Yuhong Zhang, Hengsheng Zhang, Zhengxue Cheng, Rong Xie, Li Song, and Wenjun Zhang. Ssp-ir: Semantic and structure priors for diffusion-based realistic image restoration. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[52] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances In Neural Information Processing Systems(NeurIPS)*, 35:30599–30611, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction(Section 1) accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our method in the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We introduce the method clearly in Section 3, and present the detailed experimental settings and implementation details in Section 4.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Codes are included in the supplementary zip file.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Detailed experimental settings can be found in Section 4.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Our model is trained on a large-scale dataset, and the inference process of diffusion-based super-resolution models is computationally intensive. Due to limited computational resources, we were not able to perform repeated trials or statistical significance tests. However, we ensured fair comparisons by using fixed seeds and consistent evaluation settings.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the computational resources in Section 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All the authors have reviewed the code of ethics and obey the code.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited related original papers and models in our paper. We ensure that all licenses and terms of use are respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release newly trained models and inference/training code, with documentation and usage instructions provided in the supplementary material.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve human subjects or crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our research does not involve human subjects and therefore does not require IRB or equivalent review.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Our method incorporates LLaVA, a vision-language large model (VLM) built upon LLMs, to extract semantic prior from degraded images. These prior are crucial for guiding both degradation removal and restoration, making LLM usage an integral part of our approach.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.