DreamClear: High-Capacity Real-World Image Restoration with Privacy-Safe Dataset Curation

Yuang Ai[♣,♡] **Xiaoqiang Zhou**[♣,◇] **Huaibo Huang**[♣,♡,✉]
Xiaotian Han[♠] **Zhengyu Chen**[♠] **Quanzeng You**[♠] **Hongxia Yang**[♠]
[♣]MAIS & NLPR, Institute of Automation, Chinese Academy of Sciences
[♡]School of Artificial Intelligence, University of Chinese Academy of Sciences
[♠]ByteDance, Inc [◇]University of Science and Technology of China
shallowdream555@gmail.com, huaibo.huang@cripac.ia.ac.cn
Code and models: https://github.com/shallowdream204/DreamClear

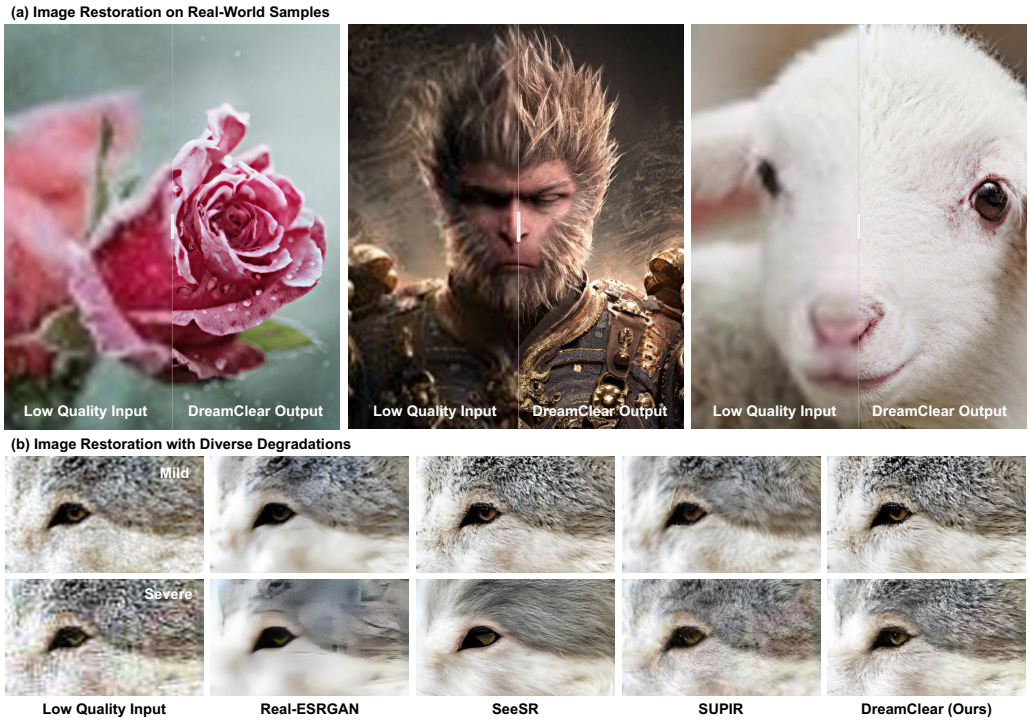

Figure 1: We present **DreamClear**, a high-capacity image restoration model that delivers photorealistic restoration of real-world LQ images, outperforming SOTA diffusion-based models in handling diverse degradations.

Abstract

Image restoration (IR) in real-world scenarios presents significant challenges due to the lack of high-capacity models and comprehensive datasets. To tackle these issues, we present a dual strategy: GenIR, an innovative data curation pipeline, and DreamClear, a cutting-edge Diffusion Transformer (DiT)-based image restoration model. **GenIR**, our pioneering contribution, is a dual-prompt learning pipeline that overcomes the limitations of existing datasets, which typically comprise only a few thousand images and thus offer limited generalizability for larger models. GenIR streamlines the process into three stages: image-text pair construction, dual-prompt based fine-tuning, and data generation & filtering. This approach circumvents the laborious data crawling process, ensuring copyright compliance and providing a cost-effective, privacy-safe solution for IR dataset construction. The result is a

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

large-scale dataset of one million high-quality images. Our second contribution, **DreamClear**, is a DiT-based image restoration model. It utilizes the generative priors of text-to-image (T2I) diffusion models and the robust perceptual capabilities of multi-modal large language models (MLLMs) to achieve photorealistic restoration. To boost the model's adaptability to diverse real-world degradations, we introduce the Mixture of Adaptive Modulator (MoAM). It employs token-wise degradation priors to dynamically integrate various restoration experts, thereby expanding the range of degradations the model can address. Our exhaustive experiments confirm DreamClear's superior performance, underlining the efficacy of our dual strategy for real-world image restoration.

1 Introduction

Image restoration (IR), a vital field in computer vision, targets transforming degraded low-quality (LQ) images into high-quality (HQ) counterparts. While IR has achieved significant advancements under predefined conditions, such as super-resolution [92, 10] and denoising [12, 40] tasks, real-world IR remains a formidable challenge due to the diversity and complexity of degradation types. The disconnect between training data and real-world scenarios is substantial, as existing datasets inadequately encapsulate the intricacies of real-world degradations. Efforts to bridge this gap include domain adaptation [5, 74, 23, 81], dataset collection [68, 8, 11, 91], and degradation simulation [64, 85, 49, 71]. However, in contrast to the leaps in Natural Language Processing (NLP) [1] and AI-Generated Content (AIGC) [59] enabled by large-scale models and extensive data, IR's progress is not as pronounced. Real-world challenges persist, and the potential of large-scale data and high-capacity models remains largely untapped. This leads us to two critical questions: *how can we obtain a large-scale dataset that accurately represents real-world IR, and based on this, how can we construct powerful models tailored for real-world IR scenarios?*

Addressing the first question, considerable efforts have been made to curate IR datasets. Given the challenge of collecting real-world paired IR data, these datasets are typically constructed by acquiring HQ images and then simulating degradations to generate corresponding LQ images. While many works [64, 85, 49, 71] have refined the degradation simulation process, this paper focuses on the acquisition of HQ images and the associated challenges of copyright and privacy protection. The predominant method for obtaining HQ images is web scraping. Current open-source IR datasets, such as DIV2K [44] and Flickr2K [2], contain only a few thousand images, insufficient for covering a broad spectrum of real-world scenarios. Larger collections like SUPIR [80], with 20 million images, highlight the labor-intensive nature of large-scale dataset curation. Moreover, images sourced from the internet often involve copyright issues and privacy concerns, particularly with identifiable human faces. To advance the IR field effectively, there is an urgent need for a dataset curation method that is privacy-safe and cost-effective.

In response, we present an under-explored approach in the image restoration (IR) field: creating high-quality, non-existent images to enhance dataset curation efficiency, while evading copyright and privacy issues. We unveil **GenIR**, a privacy-conscious, automated data curation pipeline that repurposes the generative prior in pretrained text-to-image (T2I) models for IR tasks, and uses multimodal large language models (MLLMs) to generate text prompts, thereby improving data synthesis quality. GenIR operates in three stages: (1) image-text pairs construction, (2) dual-prompt fine-tuning, and (3) data generation & filtering. Initially, GenIR utilizes existing IR datasets and the advanced MLLM, Gemini-1.5-Pro [62], to create image-text pairs, while generating negative samples via an image-to-image pipeline [50]. Subsequently, we apply a dual-prompt learning strategy to adapt pretrained T2I models to the IR task, generating suitable prompts for data synthesis. In the final stage, MLLMs create various scene descriptions and synthesize images using the adapted image prior, with a focus on ensuring no identifiable individuals are included. MLLMs also assess and filter the synthesized data based on quality, producing high-quality images that are privacy-safe and copyright-free. Through GenIR, we generate a dataset of one million high-quality images, proving its efficacy in training a robust real-world IR model.

Armed with a large-scale, high-quality image dataset, our focus shifts to the construction of a high-capacity IR model that can robustly generalize to real-world scenarios. Recent state-of-the-art approaches [77, 70, 80] employ the generative priors in pretrained Stable Diffusion [59] (SD) for realistic image restoration, underlining the power of rich generative prior in SD. As Fig.1

illustrates, SD-based methods outperform GAN-based ones. However, these strategies often neglect the degradation priors in input low-quality images, a critical element in blind IR [72]. This insight leads us to investigate the integration of degradation prior into diffusion-based IR models, and how to optimize its synergy with large models.

In this paper, we introduce **DreamClear**, a high-capacity real-world image restoration model, grounded on a large dataset. DreamClear is based on Diffusion Transformer (DiT) [53], the cornerstone of modern diffusion-based systems (*e.g.*, Sora [7], Stable Diffusion 3 [18]). Our model employs a dual-branch framework with textual guidance from multi-modal large language models (MLLMs) for photorealistic restoration. DreamClear first processes the low-quality image through a lightweight network to produce a reference image. We propose ControlFormer to enhance the control over DiT-based T2I models, thereby better utilizing the low-quality and reference images to guide the content of the generated image. To further improve the model's generalization across diverse and complex degradations, we incorporate implicit prior degradation information to refine the solution space. Specifically, we suggest a Mixture of Adaptive Modulator (MoAM), which extracts token-wise degradation representations and dynamically integrates various restoration experts for each token based on the Mixture-of-Experts (MoE) [61] structure, thereby enhancing the model's adaptability to different degradation severities (See Fig. 1).

The main contributions of this work can be summarized as follows:

- We propose GenIR, a pioneering automated data curation pipeline for image restoration. It addresses the urgent need for privacy-safe and cost-effective methods in image restoration, yielding a dataset of one million high-quality images.
- We present DreamClear, a robust, high-capacity IR model that incorporates degradation priors into diffusion-based frameworks. This model improves control over content generation, adapts to various degradations, and generalizes well across diverse real-world scenarios.
- Extensive experiments across both low-level (synthetic & real-world) and high-level (detection & segmentation) benchmarks have demonstrated DreamClear's state-of-the-art performance in handling intricate real-world scenarios.

2 Related Work

Image Restoration. Image Restoration aims at restoring a high-quality image from the low-quality input image. Over the past decade, different approaches have been proposed for image super-resolution [95, 31, 65, 96], denoising [87, 84, 67], deblurring [57, 69, 56], deraining [14, 34, 32, 33], inpainting [97, 3], etc. Recently, researchers have increasingly focused on enhancing the generalization ability to diverse degradations in real-world applications [74, 68, 63]. The degradation simulation is improved from simple degradations to complex degradation processes, such as BSRGAN [85], Real-ESRGAN [64] and AnimeSR [71]. With the improved degradation simulation process, many recent methods can deal with diverse degradation types and achieve promising performance in real-world scenarios [43, 8]. With the paired data, training a randomly initialized restoration model from scratch is one way to improve generalization ability [41]. The other way is to exploit the generative prior in the pre-trained generative model, such as GAN or diffusion models [52, 63, 4, 37, 66, 9]. In this work, we propose a data synthesis pipeline and introduce a real-world image restoration model with high generalizability.

Generative Prior. Generative models learn the image synthesis process and embed the image prior in the model weights. The image prior in a high-quality image generator, such as StyleGAN [36] and Stable Diffusion [59], can be adapted to other visual restoration tasks [43, 8, 38, 20]. To use the image prior in GANs, an additional encoder is often applied to convert the input image to the latent space [76, 48]. For the diffusion models, the forward process adds noise to the image gradually and finally converts the image to the latent noise space [15, 50]. By manipulating in the latent feature space, the input image is integrated into the generation process as a conditional input, and the synthesis process exploits the image prior in the pre-trained models. The generative prior in the pre-trained models can also serve as a good initialization for downstream synthesis tasks [63, 78, 35]. We exploit the generative prior in the pre-trained diffusion models to synthesize datasets for image restoration tasks and train a restoration model for real-world applications.

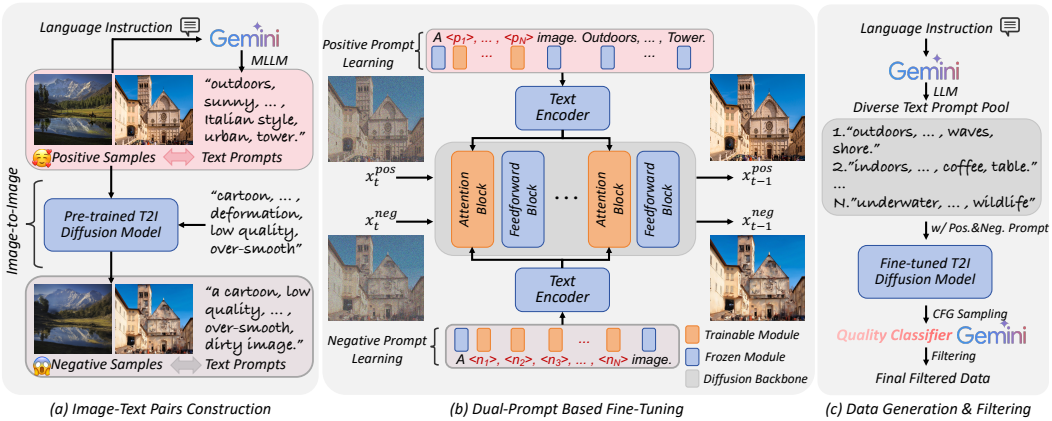

Figure 2: An overview of the three-stage **GenIR** pipeline, which includes (a) Image-Text Pairs Construction, (b) Dual-Prompt Based Fine-Tuning, and (c) Data Generation & Filtering.

Synthetic Dataset. Data size and data quality are widely recognized as essential for many vision tasks. A large-scale high-quality dataset can facilitate the large model training and improve the model ability greatly [83, 22, 21, 24, 26]. Existing large-scale datasets are often manually collected with laborious human efforts [16, 39]. More importantly, the data crawled from the internet may leak privacy information [58, 24], raising concerns related to AI security. The synthesized datasets can not only reduce the laborious human efforts, but also avoid the privacy information leakage. High-quality synthesized datasets are verified to be effective in many vision tasks [28, 25, 6]. Our work is the first to explore the dataset synthesis in the image restoration field.

3 Privacy-Safe Dataset Curation Pipeline

Traditionally, IR datasets are created by scraping web images and simulating degradations to generate low-quality (LQ) counterparts. This process is labor-intensive and rife with copyright and privacy issues, especially with identifiable human faces. To advance the IR field, a privacy-safe and cost-effective dataset curation method is needed. Drawing inspiration from the success of text-to-image (T2I) models in synthesizing high-quality images, we introduce the GenIR pipeline. This novel approach leverages the generative priors of pre-trained T2I models to construct extensive, privacy-safe datasets. However, the efficacy of T2I models is contingent upon carefully crafted prompts for generating high-quality images fitting for IR tasks.

To tackle this, GenIR, as illustrated in Fig. 2, employs a streamlined three-stage process. Initially, we construct positive and negative samples, each paired with corresponding text prompts. Subsequently, a dual-prompt based finetuning strategy concurrently learns both positive and negative prompts. Finally, we utilize LLMs to generate a diverse array of text prompts, leading to the creation and filtering of data. Throughout this process, we maintain stringent privacy standards, ensuring no specific personal information is embedded in the text prompts or the generated images.

Image-Text Pairs Construction. We use high-resolution, texture-rich images in existing IR datasets [44, 2, 22, 39] as positive samples. Given the unavailability of corresponding text prompts, we employ the sophisticated MLLM, Gemini-1.5-Pro [62], to generate necessary prompts via language instructions. Moreover, to identify and eliminate unwanted content during the T2I model's fine-tuning and enhance image quality, we generate negative samples representing undesirable outcomes using the T2I model. As depicted in Fig. 2 (a), we adopt the image-to-image pipeline proposed in [50], using the T2I model and manually designed prompts such as "*cartoon, painting, ... , over-smooth, dirty*", to directly generate negative samples.

Dual-Prompt based Fine-Tuning. Rather than relying on complex, labor-intensive prompts with limited applicability, we propose an innovative dual-prompt based fine-tuning approach to refine the T2I model for our data needs. As illustrated in Fig. 2 (b), we employ positive and negative samples to learn their corresponding prompts. Specifically, we use M positive tokens $\{\langle p_1 \rangle, \cdots, \langle p_M \rangle\}$ and N

negative tokens $\{\langle n_1 \rangle, \cdots, \langle n_N \rangle\}$ to represent desired and undesired attributes, respectively, and subsequently learn their embeddings. We initialize these new positive and negative tokens using frequently used positive (e.g., "*4k, highly detailed, professional ...*") and negative text prompts (e.g., "*deformation, low quality, over-smooth ...*"). As the text condition is integrated into the diffusion model via cross-attention, we also refine the attention block to better comprehend these new tokens. After fine-tuning the T2I model with our curated image-text pairs, we can efficiently employ the learned prompts and refined diffusion model to readily generate the needed data.

Data Generation & Filtering. In addition to the quality of images, the diversity of scenes within the IR dataset is of paramount importance. To address this, we leverage Gemini to generate one million text prompts, describing varied scenes under carefully curated language instructions. These instructions explicitly proscribe the inclusion of personal or sensitive information, thereby ensuring privacy. As depicted in Fig. 2 (c), we employ the fine-tuned T2I model in conjunction with the learned positive and negative prompts to generate HQ images.

Classifier-free guidance (CFG) [30] provides a mechanism to effectively utilize negative prompts, thereby mitigating the generation of undesired content. During the sampling phase, the denoising model ϵ_θ anticipates two outcomes: one associated with the positive prompt pos and the other with the negative prompt neg. The final CFG prediction is formulated as

$$\epsilon_\theta(z_t, t, pos, neg) = \omega \times \epsilon_\theta(z_t, t, pos) + (1 - \omega) \times \epsilon_\theta(z_t, t, neg), \tag{1}$$

where ω denotes the CFG guidance scale. Post-sampling, the generated images are evaluated by a quality classifier, which decides whether to retain the images based on the predicted probabilities. This binary classifier is trained on positive and negative samples. Gemini is subsequently used to ascertain whether the images exhibit blatant semantic errors or inappropriate content.

Contrasted with direct web crawling, our GenIR provides a more cost-effective and privacy-preserving approach to data acquisition. It circumvents the potential infringement of personal privacy information prevalent on the web, thereby ensuring our research is both ethical and secure - a crucial aspect in the current artificial intelligence landscape characterized by extensive data usage. Ultimately, we gather one million high-resolution (2040×1356) images, each of superior quality.

4 High-Capacity Image Restoration Model

The complex and varied degradation of real-world images presents a major challenge to the practical applicability of restoration models. We introduce DreamClear, a high-capacity image restoration model that dynamically integrates various restoration experts, guided by prior degradation information. DreamClear is built upon on PixArt-α [13], a pre-trained T2I diffusion model based on the Diffusion Transformer (DiT) [53] architecture, which has proven its powerful generative capabilities [93].

Architecture Overview. Fig. 3 illustrates that DreamClear features a dual-branch architecture, encompassing an LQ Branch and a Reference Branch. LQ images I_{lq} are processed by SwinIR [41], a lightweight degradation remover, resulting in smoother, albeit less detailed, reference images I_{ref}. Considering potential detail loss in I_{ref}, we employ both I_{lq} and I_{ref} to direct the diffusion model. Moreover, we utilize the open-source MLLM, LLaVA [47], to generate detailed captions for training images using the prompt "*Describe this image and its style in a very detailed manner*", supporting the T2I diffusion model in attaining more realistic restoration.

ControlFormer. ControlNet [88], a prevalent structure for managing diffusion models, is tailored for the U-Net [60] structure in SD. It is unsuitable for DiT, stemming from the architecture difference. To address this, we present ControlFormer, which inherits ControlNet's core features (trainable copy blocks and zero-initialized layers) but is adapted for the DiT-based T2I model, as shown in Fig. 3. ControlFormer, duplicating all DiT Blocks from PixArt-α, employs the MoAM block to combine LQ features x_{lq} and reference features x_{ref}. This DiT-optimized ControlFormer maintains ControlNet's essential components, providing effective spatial control within DiT.

Mixture-of-Adaptive-Modulator. To enhance our model's robustness to real-world degradations, we propose a degradation-aware Mixture-of-Adaptive-Modulator (MoAM) for effective LQ and reference feature fusion. As shown in Fig. 3, MoAM consists of adaptive modulators (AM), a

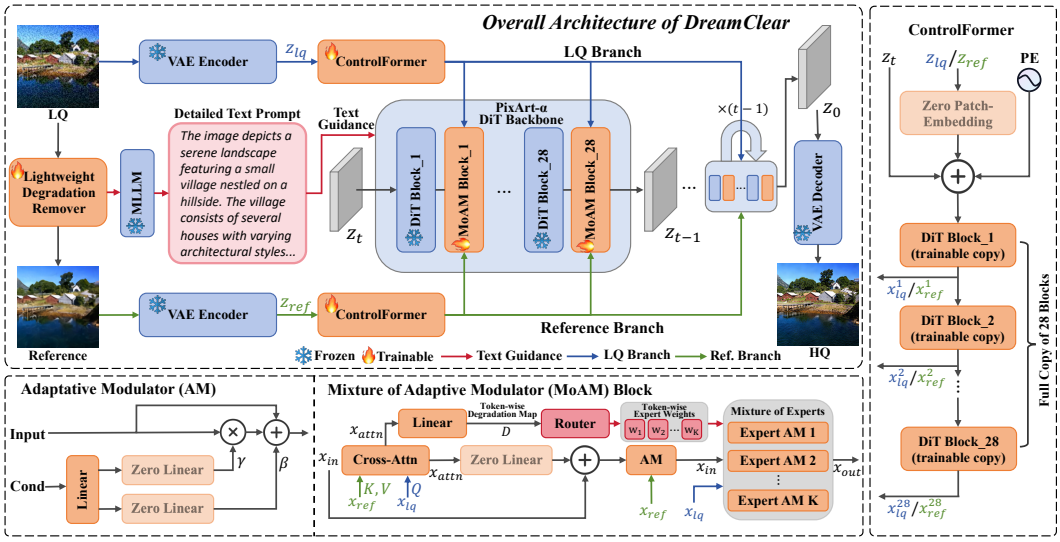

Figure 3: Architecture of the proposed **DreamClear**. DreamClear adopts a dual-branch structure, using Mixture of Adaptive Modulator to merge LQ features and Reference features. We utilize MLLM to generate detailed text prompt as the guidance for T2I model.

cross-attention layer, and a router block. AM employs AdaLN [54] to learn dimension-wise scale γ and shift β parameters, embedding conditional information into input features.

MoAM operates in three steps: 1) For DiT features x_{in}, we calculate the cross-attention output $x_{attn} \in \mathbb{R}^{N\times C}$ between $x_{lq} \in \mathbb{R}^{N\times C}$ and $x_{ref} \in \mathbb{R}^{N\times C}$, where N and C denote the number of visual tokens and hidden size. x_{in} is then modulated using x_{attn} followed by a zero linear layer. A token-wise degradation map $D \in \mathbb{R}^{N\times C}$ is generated through the linear mapping of x_{attn}. 2) Features are further modulated using AM, with x_{ref} as the AM condition to extract clean features. 3) We adopt a mixture of degradation-aware experts to adapt to diverse degradations, detailed below.

Given varying degradations in real-world images, our method dynamically processes tokens using degradation priors. Each MoAM block consists of K restoration experts (*i.e.*, AM) $\{E_1, \cdots, E_K\}$, each specialized for specific degradation scenarios. A routing network $R(\cdot)$ dynamically merges expert guidance for tokens, based on D. The routing network, a two-layer MLP followed by softmax, yields token-wise expert weights $w = R(D) \in \mathbb{R}^{N\times K}$. The dynamic mixture of restoration experts is formulated as

$$\gamma(i) = \sum_{k=1}^{K} w(i,k) \times Net_k^{\gamma}[x_{lq}(i)], \quad \beta(i) = \sum_{k=1}^{K} w(i,k) \times Net_k^{\beta}[x_{lq}(i)], \tag{2}$$

$$x_{out} = (1+\gamma) \otimes x_{in} + \beta, \tag{3}$$

where i and k index tokens and experts respectively, Net^{γ} and Net^{β} map within an expert to γ and β, and $\otimes$ denotes element-wise multiplication. MoAM dynamically fuses expert knowledge, leveraging degradation priors to tackle complex degradations.

5 Experiments

5.1 Experimental Setup

Datasets. We adopt a combination of DIV2K [44], Flickr2K [2], LSDIR [39], DIV8K [22], and our generated dataset to train DreamClear. We employ the Real-ESRGAN [64] degradation pipeline to generate LQ images, using the same degradation settings as SeeSR [70] to ensure a fair comparison. All experiments are conducted on scaling factor $\times 4$.

For testing datasets, following previous works [63, 80, 70], we evaluate our method on both synthetic and real-world benchmarks. For synthetic benchmarks, we randomly crop 3,000 patches from the

Table 1: Quantitative comparison with state-of-the-art real-world IR methods on both synthetic and real-world benchmarks. Best and second best performance are highlighted in **red** and **blue**, respectively.

Datasets	Metrics	BSRGAN [85]	Real-ESRGAN [64]	SwinIR-GAN [41]	DASR [43]	StableSR [63]	DiffBIR [46]	ResShift [82]	SinSR [65]	SeeSR [70]	SUPIR [80]	**DreamClear**
DIV2K-Val	PSNR ↑	19.88	**19.92**	19.66	19.73	19.73	**19.98**	19.80	19.37	19.59	18.68	18.69
	SSIM ↑	0.5137	**0.5334**	**0.5253**	0.5122	0.5039	0.4987	0.4985	0.4613	0.5045	0.4664	0.4766
	LPIPS ↓	0.4303	0.3981	0.3992	0.4350	0.4145	0.3866	0.4450	0.4383	**0.3662**	0.3976	**0.3657**
	DISTS ↓	0.2484	0.2304	0.2253	0.2606	0.2162	0.2396	0.2383	0.2175	0.1886	**0.1818**	**0.1637**
	FID ↓	54.42	48.44	49.17	59.62	29.64	37.00	46.12	37.84	**24.98**	28.11	**20.61**
	NIQE ↓	3.9322	3.8762	3.7468	3.9725	4.4255	4.5659	5.9852	5.7320	4.1320	**3.4014**	**3.2126**
	MANIQA ↑	0.3514	0.3854	0.3654	0.3110	0.2942	0.4268	0.3782	0.4206	**0.5251**	0.4291	**0.4320**
	MUSIQ ↑	63.93	64.50	64.54	59.66	58.60	64.77	62.67	65.27	**72.04**	**69.34**	68.44
	CLIPIQA ↑	0.5589	0.5804	0.5682	0.5565	0.5190	0.6527	0.6498	0.6961	**0.7181**	0.6035	**0.6963**
LSDIR-Val	PSNR ↑	**18.27**	18.13	17.98	18.15	18.11	**18.42**	18.24	17.94	18.03	16.95	17.01
	SSIM ↑	0.4673	**0.4867**	**0.4783**	0.4679	0.4508	0.4618	0.4579	0.4302	0.4564	0.4080	0.4236
	LPIPS ↓	0.4378	0.3986	0.4020	0.4503	0.4152	0.4049	0.4524	0.4523	**0.3759**	0.4119	**0.3836**
	DISTS ↓	0.2539	0.2278	0.2253	0.2615	0.2159	0.2439	0.2436	0.2265	0.1966	**0.1838**	**0.1656**
	FID ↓	53.25	46.46	45.31	60.60	31.26	35.91	43.25	36.01	**25.91**	30.03	**22.06**
	NIQE ↓	3.6885	3.4078	3.3715	3.6432	4.0218	4.3750	5.5635	5.4240	4.0590	**2.9820**	**3.0707**
	MANIQA ↑	0.3829	0.4381	0.3991	0.3315	0.3098	0.4551	0.3995	0.4309	**0.5700**	0.4683	**0.4811**
	MUSIQ ↑	65.98	68.25	67.10	60.96	59.37	65.94	63.25	65.35	**73.00**	**70.98**	70.40
	CLIPIQA ↑	0.5648	0.6218	0.5983	0.5681	0.5190	0.6592	0.6501	0.6900	**0.7261**	0.6174	**0.6914**
RealSR	PSNR ↑	**25.01**	24.22	24.89	**25.51**	24.60	24.77	24.94	24.47	24.66	22.67	22.56
	SSIM ↑	0.7422	0.7401	**0.7543**	**0.7526**	0.7387	0.6902	0.7178	0.6710	0.7209	0.6567	0.6548
	LPIPS ↓	0.2853	0.2901	**0.2680**	0.3201	**0.2736**	0.3436	0.3864	0.4208	0.2997	0.3545	0.3684
	DISTS ↓	0.1967	0.1892	**0.1734**	0.2056	**0.1761**	0.2195	0.2467	0.2432	0.2029	0.2185	0.2122
	FID ↓	84.49	90.10	80.07	91.16	88.89	**69.94**	88.91	70.83	71.92	71.63	**65.37**
	NIQE ↓	4.9261	5.0069	4.9475	5.9659	5.6124	6.1294	6.6044	6.4662	4.9102	**4.5368**	**4.4381**
	MANIQA ↑	0.3660	0.3656	0.3432	0.2819	0.3465	0.4182	0.3781	0.4009	**0.5189**	0.4296	**0.4337**
	MUSIQ ↑	64.67	62.06	60.97	50.94	61.07	61.74	60.28	60.36	**69.38**	**66.09**	65.33
	CLIPIQA ↑	0.5329	0.4872	0.4548	0.3819	0.5139	0.6202	0.5778	0.6587	**0.6839**	0.5371	**0.6895**
DrealSR	PSNR ↑	27.09	26.95	27.00	**28.19**	**27.39**	27.31	27.16	26.15	27.10	24.41	24.48
	SSIM ↑	0.7759	0.7812	0.7815	**0.8051**	**0.7830**	0.7140	0.7388	0.6564	0.7596	0.6696	0.6508
	LPIPS ↓	0.2950	0.2876	**0.2789**	0.3165	**0.2710**	0.3920	0.4101	0.4690	0.3117	0.3844	0.3972
	DISTS ↓	0.1956	0.1857	**0.1787**	0.2072	**0.1737**	0.2443	0.2553	0.2540	0.2103	0.2264	0.2145
	FID ↓	84.26	83.79	84.22	94.96	80.23	76.89	91.82	85.26	**75.07**	90.78	**74.78**
	NIQE ↓	5.5866	5.7422	5.5749	6.9663	6.1699	6.3433	7.5616	6.8770	5.7696	**5.1115**	**4.6295**
	MANIQA ↑	0.3420	0.3423	0.3269	0.2754	0.3171	0.3801	0.3350	0.3890	**0.4974**	**0.4174**	0.3676
	MUSIQ ↑	61.22	58.37	57.33	46.49	56.43	55.14	55.27	58.50	**67.42**	**64.53**	59.83
	CLIPIQA ↑	0.5385	0.4847	0.4819	0.3828	0.5344	0.6005	0.5788	0.6734	**0.7022**	0.5800	**0.6620**
RealLQ250	NIQE ↓	4.5229	4.1091	4.0912	4.7486	4.6349	4.8160	5.9727	5.7768	4.4126	**3.6336**	**3.5556**
	MANIQA ↑	0.3523	0.3592	0.3632	0.2782	0.2939	0.4017	0.3816	0.4229	**0.4992**	0.3926	**0.4351**
	MUSIQ ↑	63.66	62.74	63.63	53.39	57.11	62.18	61.55	64.09	**70.57**	66.03	**66.76**
	CLIPIQA ↑	0.5695	0.5465	0.5583	0.4671	0.5208	0.6420	0.6298	0.7044	**0.7104**	0.5800	**0.7116**

Table 2: Evaluation on COCO val2017 (object detection & instance segmentation) and ADE20K (semantic segmentation).

Metrics	GT	Zoomed LQ	BSRGAN	Real-ESRGAN	SwinIR-GAN	DASR	StableSR	DiffBIR	ResShift	SinSR	SeeSR	SUPIR	**DreamClear**
Object Detection (AP^b)	49.0	7.4	11.0	12.8	11.8	10.5	16.9	18.7	15.6	13.8	18.2	16.6	**19.3**
Object Detection (AP^b_{50})	70.6	12.0	17.6	20.7	18.9	17.0	26.7	29.9	25.0	22.3	29.1	27.2	**30.8**
Object Detection (AP^b_{75})	53.8	7.5	11.4	13.1	12.1	10.7	17.6	19.4	15.9	14.2	18.9	17.0	**19.8**
Instance Segmentation (AP^m)	43.9	6.4	9.6	11.3	10.2	9.3	14.6	16.2	13.6	12.0	15.9	14.1	**16.7**
Instance Segmentation (AP^m_{50})	67.7	11.2	16.4	19.3	17.5	15.9	24.6	27.5	23.3	20.6	26.6	24.5	**28.2**
Instance Segmentation (AP^m_{75})	47.3	6.3	9.7	11.5	10.2	9.4	14.9	16.6	13.7	12.1	16.1	14.0	**16.8**
Semantic Segmentation (mIoU)	50.4	11.5	18.6	17.3	14.3	30.4	19.6	23.6	29.7	19.6	26.9	27.7	**31.9**

validation sets of DIV2K and LSDIR, and degrade them using the same settings as training. We name these two benchmarks as *DIV2K-Val* and *LSDIR-Val* respectively. For real-world benchmarks, we conduct experiments on commonly used *RealSR* [8] and *DRealSR* [68] datasets. Besides, we establish another real-world benchmark, called *RealLQ250*, which includes a total of 250 LQ images of size 256×256 used in previous works [70, 42, 64, 86, 80] or sourced from the Internet, without corresponding GT images. For all testing datasets with GT images, the resolution of the HQ-LQ image pairs is 1024×1024 and 256×256, respectively.

Metrics. Following SeeSR [70], we adopt PSNR and SSIM (calculated on the Y channel of transformed YCbCr space) as reference-based distortion metrics, LPIPS [90] and DISTS [17] as reference-based perceptual metrics, NIQE [89], MANIQA [75], MUSIQ [75] and CLIPIQA [75] as no-reference metrics. FID [29] is used to evaluate the image quality. These metrics can achieve a comprehensive evaluation of the restoration effects.

Implementation Details For training GenIR and DreamClear, we both use the original latent diffusion loss [59]. The proposed GenIR framework is built on SDXL [55] and trained over 5 days using 16 NVIDIA A100 GPUs. The training is conducted on 1024×1024 resolution images with

Figure 4: Qualitative comparisons on both synthetic (the first row) and real-world (the last two rows) benchmarks. Please zoom in for a better view.

a batch-size of 256. For data generation, we use 256 NVIDIA V100 GPUs and spend 5 days to generate the large-scale dataset. Our DreamClear is built upon PixArt-α [13] and LLaVA [47]. The SwinIR model in DiffBIR [46] is used as the lightweight degradation remover. We use the AdamW optimizer with a learning rate of $5e^{-5}$ to train our model. The training is conducted on 1024×1024 resolution images, running for 7 days on 32 NVIDIA A100 GPUs with a batch-size of 128. The number of experts K in Eq. (2) is set to 3. For inference of DreamClear, we adopt iDDPM [51] with 50 sampling steps, CFG guidance scale $\omega = 4.5$.

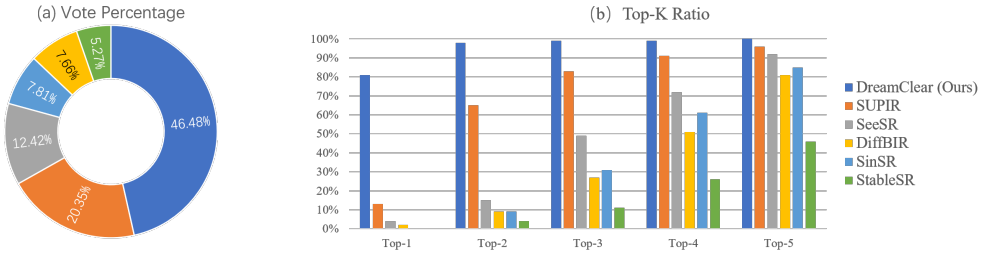

Figure 5: User study. Vote percentage denotes average user preference per model. The Top-K ratio indicates the proportion of images preferred by top K users. Our model is highly preferred, with most images being rated as top quality by the majority.

5.2 Comparison with State-of-the-Art Methods

We compare our method with state-of-the-art GAN-based methods (BSRGAN [85], Real-ESRGAN [64], SwinIR-GAN [41], and DASR [43]) and recent diffusion-based methods (StableSR [63], DiffBIR [46], ResShift [82], SinSR [65], SeeSR [70], and SUPIR [80]).

Quantitative Comparisons. Tab. 1 presents quantitative results on various benchmarks. Our method consistently excels in perceptual metrics (LPIPS, DISTS, FID) on synthetic datasets, signifying high perceptual quality. On real-world benchmarks, our method performs strongly across most no-reference metrics (NIQE, MANIQA, MUSIQ, CLIPIQA), attesting to the high quality of our restorations. Our diffusion-based method prioritizes photorealistic restoration. Despite lower PSNR/SSIM scores, recent works [79, 80] argue these metrics inadequately represent visual quality, and it is necessary to reconsider the reference values of existing metrics and propose more effective methods to evaluate modern image restoration methods. We believe that as the field of image quality assessment (IQA) evolves, more suitable metrics will emerge to adequately measure the performance of advanced IR models.

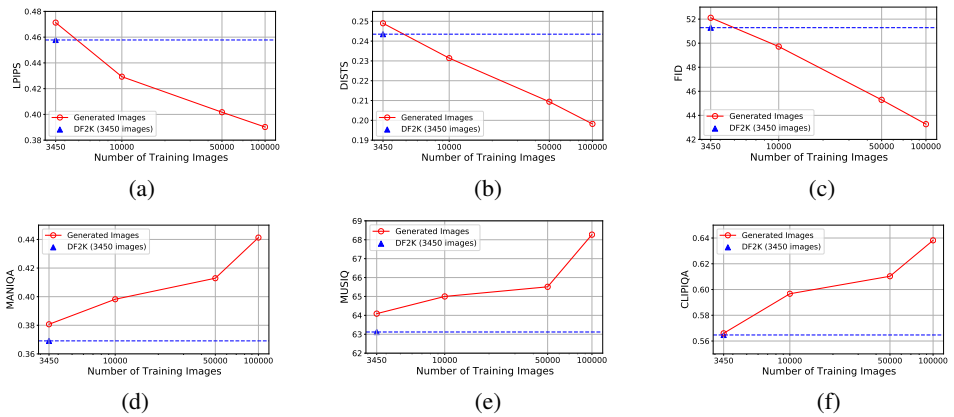

Figure 6: Impact of synthetic training data. As data size increases, performance improves on *LSDIR-Val*.

Table 3: Ablation results on *DIV2K-Val*, COCO val2017 and ADE20K for DreamClear.

	LPIPS ↓	DISTS ↓	FID ↓	MANIQA ↑	MUSIQ ↑	CLIPIQA ↑	AP^b	AP^m	mIoU
Mixture of AM	**0.3657**	**0.1637**	**20.61**	**0.4320**	**68.44**	**0.6963**	**19.3**	**16.7**	**31.9**
AM	0.3981	0.1843	25.75	0.4067	66.18	0.6646	18.0	15.6	28.6
Cross-Attention	0.4177	0.2016	29.74	0.3785	63.21	0.6497	17.2	15.1	26.3
Zero-Linear	0.4082	0.1976	29.89	0.4122	66.11	0.6673	17.6	15.3	27.2
Dual-Branch	**0.3657**	**0.1637**	**20.61**	**0.4320**	**68.44**	**0.6963**	**19.3**	**16.7**	**31.9**
w/o Reference Branch	0.4207	0.2033	30.91	0.3985	64.04	0.6582	15.9	14.0	24.7
Detailed Text Prompt	0.3657	0.1637	20.61	**0.4320**	**68.44**	**0.6963**	**19.3**	**16.7**	**31.9**
Null Prompt	**0.3521**	**0.1607**	**20.47**	0.4230	67.26	0.6812	18.8	16.2	29.8

Qualitative Comparisons. We provide qualitative comparisons in Fig. 4. When handling severe degradations (the first row), only our DreamClear can not only reason the correct structure but also generate clear details, while other methods may generate deformed structure and blurry results. When it comes to real-world images, our method can achieve results that are rich in detail and more natural (the third row). More real-world visual comparisons are in Apendix A.4.

User Study. We conducted a user study to evaluate our model's restoration quality using 100 low-quality images, restored by our method and five others. Users were asked to rank the restorations considering visual quality, naturalness, and accuracy, among others. The study, involving 256 evaluators, was designed for fairness and wide participation. Two metrics, vote percentage and Top-K ratio, were used to analyze the results. As shown in Fig. 5, our model led on both metrics, receiving over 45% of total votes and being the top choice for 80% of the images, demonstrating its consistent superiority in producing high-quality images. Refer to Appendix A.2 for more details.

5.3 Evaluation on Downstream Benchmarks

We assess the benefits of image restoration for downstream high-level vision tasks by conducting detection and segmentation experiments on the COCO 2017 [45] and ADE20K [94] datasets using various restoration models. Low-quality images are generated and restored under the same conditions as in training. We use the robust visual backbone RMT [19] (with Mask R-CNN [27] 1× for COCO, with UperNet [73] for ADE20K) for these tasks. Tab. 2 shows that our model obtains the best performance, implying its superiority in benefiting downstream tasks. Despite its superior performance in semantic restoration and fine-grained image recognition tasks, there's still substantial room for improvement.

5.4 Ablation Study

Analysis of Generated Dataset for Real-World IR. Due to the extensive time required to train diffusion-based models, we use SwinIR-GAN to investigate the impact of generated datasets on

real-world image restoration (IR). SwinIR is trained on varying quantities of generated data for comparison with the DF2K-trained model. Fig. 6 shows that the model trained on an equivalent number of generated images exhibits marginally lower perceptual but higher no-reference metrics than the DF2K model. As the dataset size increases, all metrics improve, reinforcing our belief that larger datasets enhance model generalizability and restoration performance. Notably, the model trained with 100,000 generated images outperforms the DF2K model, underscoring the advantages of utilizing large-scale synthetic datasets for real-world IR. More ablations of GenIR are provided in Appendix A.3.

Ablations of DreamClear. We conduct ablation studies to scrutinize the contribution of each component within DreamClear. Evaluating perceptual fidelity via LPIPS, DISTS, and FID metrics, and assessing the image quality of restoration results with MANIQ, MUSIQ, and CLIPIQA, we find that DreamClear outperforms its ablated versions across most metrics, substantiating the importance of these components (Tab. 3). Notably, null prompts slightly outperform detailed prompts on perceptual metrics. However, superior results on no-reference and high-level vision metrics suggest that the MLLM-provided detailed text prompts better preserve semantic information. Visual comparisons detailed in Appendix A.3 further reinforce the benefits of semantic guidance in image restoration provided by text prompts.

6 Limitations and Broader Impact

Our model leverages the generative prior of pre-trained diffusion models for image restoration, with diverse synthesized datasets used during training to enhance model performance. In situations of severe image degradation, while our method could predict reasonable and realistic results, the synthesized texture details may not exist in the ground-truth image. A high-quality reference image or explicit human instruction may compensate such a limitation in some degree.

Another limitation lies in the deployment in practical applications. Our model is a diffusion-based model and it needs multiple inference steps to restore the input low-quality image. While our model can predict more plausible results than existing methods, it can not meet the requirement of real-time inference speed in many practical applications. Model distillation and model quantization may compensate the limitation of inference speed.

This paper is a purely academic study of real-world image restoration (IR). However, considering image restoration's vital role in many practical applications, this work may not only bring some positive societal influence, *e.g.*, improving the quality of images captured by smartphones, but also lead to some potential risks like privacy information leakage from photos on social media. However, the positive societal effects of image restoration far exceed the potential negative impacts, and people may make use of some other techniques, such as inpainting and watermarking, to remove the private information in images.

7 Conclusion

To address the challenges in real-world image restoration (IR), we develop **GenIR**, a privacy-safe automated pipeline that generates a large-scale dataset of one million high-quality images, serving as a robust training resource for IR models. Additionally, we introduce **DreamClear**, a potent IR model that seamlessly integrates degradation priors into diffusion-based IR models. It introduces the novel Mixture of Adaptive Modulator (MoAM) to adapt to diverse real-world degradations. Our comprehensive experiments underline its outstanding performance in managing complex real-world situations, marking a substantial progression in IR.

Acknowledgements

This research is partially funded by Beijing Nova Program (20230484276), and Youth Innovation Promotion Association CAS (Grant No. 2022132).

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, pages 126–135, 2017. 2, 4, 6

[3] Y. Ai, H. Huang, and R. He. Lora-ir: Taming low-rank experts for efficient all-in-one image restoration. *arXiv preprint arXiv:2410.15385*, 2024. 3

[4] Y. Ai, H. Huang, X. Zhou, J. Wang, and R. He. Multimodal prompt perceiver: Empower adaptiveness generalizability and fidelity for all-in-one image restoration. In *CVPR*, pages 25432–25444, 2024. 3

[5] Y. Ai, X. Zhou, H. Huang, L. Zhang, and R. He. Uncertainty-aware source-free adaptive image super-resolution with wavelet augmentation transformer. In *CVPR*, pages 8142–8152, 2024. 2

[6] S. Azizi, S. Kornblith, C. Saharia, M. Norouzi, and D. J. Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 4

[7] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh. Video generation models as world simulators, 2024. 3

[8] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, pages 3086–3095, 2019. 2, 3, 7

[9] J. Cao, Y. Shi, K. Zhang, Y. Zhang, R. Timofte, and L. Van Gool. Deep equilibrium diffusion restoration with parallel sampling. In *CVPR*, pages 2824–2834, 2024. 3

[10] M. Cao, C. Mou, F. Yu, X. Wang, Y. Zheng, J. Zhang, C. Dong, G. Li, Y. Shan, R. Timofte, et al. Ntire 2023 challenge on 360deg omnidirectional image and video super-resolution: Datasets, methods and results. In *CVPRW*, pages 1731–1745, 2023. 2

[11] C. Chen, Z. Xiong, X. Tian, Z.-J. Zha, and F. Wu. Camera lens super-resolution. In *CVPR*, pages 1652–1660, 2019. 2

[12] H. Chen, J. Gu, Y. Liu, S. A. Magid, C. Dong, Q. Wang, H. Pfister, and L. Zhu. Masked image training for generalizable deep image denoising. In *CVPR*, pages 1692–1703, 2023. 2

[13] J. Chen, Y. Jincheng, G. Chongjian, L. Yao, E. Xie, Z. Wang, J. Kwok, P. Luo, H. Lu, and Z. Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024. 5, 8

[14] X. Chen, H. Li, M. Li, and J. Pan. Learning a sparse transformer network for effective image deraining. In *CVPR*, pages 5896–5905, 2023. 3

[15] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *ICCV*, pages 14367–14376, 2021. 3

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 4

[17] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *TPAMI*, 44(5):2567–2581, 2020. 7

[18] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024. 3

[19] Q. Fan, H. Huang, M. Chen, H. Liu, and R. He. Rmt: Retentive networks meet vision transformers. In *CVPR*, 2024. 9

[20] B. Fei, Z. Lyu, L. Pan, J. Zhang, W. Yang, T. Luo, B. Zhang, and B. Dai. Generative diffusion prior for unified image restoration and enhancement. In *CVPR*, pages 9935–9946, 2023. 3

[21] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He. Dvg-face: Dual variational generation for heterogeneous face recognition. *TPAMI*, 44(6):2938–2952, 2021. 4

[22] S. Gu, A. Lugmayr, M. Danelljan, M. Fritsche, J. Lamour, and R. Timofte. Div8k: Diverse 8k resolution image dataset. In *ICCVW*, pages 3512–3516, 2019. 4, 6

[23] Y. Guo, J. Chen, J. Wang, Q. Chen, J. Cao, Z. Deng, Y. Xu, and M. Tan. Closed-loop matters: Dual regression networks for single image super-resolution. In *CVPR*, pages 5407–5416, 2020. 2

[24] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, pages 87–102, 2016. 4

[25] H. A. A. K. Hammoud, H. Itani, F. Pizzati, P. Torr, A. Bibi, and B. Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*, 2024. 4

[26] X. Han, Y. Jian, X. Hu, H. Liu, Y. Wang, Q. Fan, Y. Ai, H. Huang, R. He, Z. Yang, et al. Infimm-webmath-40b: Advancing multimodal pre-training for enhanced mathematical reasoning. *arXiv*

preprint arXiv:2409.12568, 2024. 4
[27] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 9
[28] R. He, S. Sun, X. Yu, C. Xue, W. Zhang, P. Torr, S. Bai, and X. Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 4
[29] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, pages 6626–6637, 2017. 7
[30] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5
[31] H. Huang, R. He, Z. Sun, and T. Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *ICCV*, pages 1689–1697, 2017. 3
[32] H. Huang, M. Luo, and R. He. Memory uncertainty learning for real-world single image deraining. *TPAMI*, 45(3):3446–3460, 2022. 3
[33] H. Huang, A. Yu, Z. Chai, R. He, and T. Tan. Selective wavelet attention learning for single image deraining. *IJCV*, 129(4):1282–1300, 2021. 3
[34] H. Huang, A. Yu, and R. He. Memory oriented transfer learning for semi-supervised image deraining. In *CVPR*, pages 7732–7741, 2021. 3
[35] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *ICCV*, pages 22623–22633, 2023. 3
[36] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 3
[37] B. Kawar, M. Elad, S. Ermon, and J. Song. Denoising diffusion restoration models. In *NeurIPS*, pages 23593–23606, 2022. 3
[38] H. Lee, K. Kang, H. Lee, S.-H. Baek, and S. Cho. Ugpnet: Universal generative prior for image restoration. In *WACV*, pages 1598–1608, 2024. 3
[39] Y. Li, K. Zhang, J. Liang, J. Cao, C. Liu, R. Gong, Y. Zhang, H. Tang, Y. Liu, D. Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *CVPRW*, pages 1775–1787, 2023. 4, 6
[40] Y. Li, Y. Zhang, R. Timofte, L. Van Gool, Z. Tu, K. Du, H. Wang, H. Chen, W. Li, X. Wang, et al. Ntire 2023 challenge on image denoising: Methods and results. In *CVPR*, pages 1904–1920, 2023. 2
[41] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. Swinir: Image restoration using swin transformer. In *ICCVW*, pages 1833–1844, 2021. 3, 5, 7, 8
[42] J. Liang, H. Zeng, and L. Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *CVPR*, pages 5657–5666, 2022. 7
[43] J. Liang, H. Zeng, and L. Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *ECCV*, pages 574–591, 2022. 3, 7, 8
[44] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, pages 136–144, 2017. 2, 4, 6
[45] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 9
[46] X. Lin, J. He, Z. Chen, Z. Lyu, B. Fei, B. Dai, W. Ouyang, Y. Qiao, and C. Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 7, 8, 19, 20, 21
[47] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *NeurIPS*, 2023. 5, 8
[48] M. Liu, Y. Wei, X. Wu, W. Zuo, and L. Zhang. Survey on leveraging pre-trained generative adversarial networks for image editing and restoration. *Science China Information Sciences*, 66(5):151101, 2023. 3
[49] Z. Luo, Y. Huang, , S. Li, L. Wang, and T. Tan. Learning the degradation distribution for blind image super-resolution. In *CVPR*, 2022. 2
[50] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 2, 3, 4, 15
[51] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171, 2021. 8
[52] X. Pan, X. Zhan, B. Dai, D. Lin, C. C. Loy, and P. Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *TPAMI*, 44(11):7474–7489, 2021. 3
[53] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 3, 5
[54] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 6

[55] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2023. 7

[56] Y. Quan, Z. Wu, and H. Ji. Neumann network with recursive kernels for single image defocus deblurring. In *CVPR*, pages 5754–5763, 2023. 3

[57] J. Rim, H. Lee, J. Won, and S. Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, pages 184–201, 2020. 3

[58] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35, 2016. 4

[59] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3, 7

[60] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 5

[61] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017. 3

[62] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 4

[63] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 3, 6, 7, 8, 19, 20, 21

[64] X. Wang, L. Xie, C. Dong, and Y. Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCVW*, pages 1905–1914, 2021. 2, 3, 6, 7, 8

[65] Y. Wang, W. Yang, X. Chen, Y. Wang, L. Guo, L.-P. Chau, Z. Liu, Y. Qiao, A. C. Kot, and B. Wen. Sinsr: Diffusion-based image super-resolution in a single step. In *CVPR*, 2024. 3, 7, 8

[66] Y. Wang, J. Yu, and J. Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022. 3

[67] Z. Wang, Y. Fu, J. Liu, and Y. Zhang. Lg-bpn: Local and global blind-patch network for self-supervised real-world denoising. In *CVPR*, pages 18156–18165, 2023. 3

[68] P. Wei, Z. Xie, H. Lu, Z. Zhan, Q. Ye, W. Zuo, and L. Lin. Component divide-and-conquer for real-world image super-resolution. In *ECCV*, pages 101–117, 2020. 2, 3, 7

[69] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar. Deblurring via stochastic refinement. In *CVPR*, pages 16293–16303, 2022. 3

[70] R. Wu, T. Yang, L. Sun, Z. Zhang, S. Li, and L. Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024. 2, 6, 7, 8, 19, 20, 21

[71] Y. Wu, X. Wang, G. Li, and Y. Shan. Animesr: Learning real-world super-resolution models for animation videos. In *NeurIPS*, pages 11241–11252, 2022. 2, 3

[72] B. Xia, Y. Zhang, Y. Wang, Y. Tian, W. Yang, R. Timofte, and L. Van Gool. Knowledge distillation based degradation estimation for blind super-resolution. In *ICLR*, 2023. 3

[73] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018. 9

[74] X. Xu, P. Wei, W. Chen, Y. Liu, M. Mao, L. Lin, and G. Li. Dual adversarial adaptation for cross-device real-world image super-resolution. In *CVPR*, pages 5667–5676, 2022. 2, 3

[75] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, pages 1191–1200, 2022. 7

[76] T. Yang, P. Ren, X. Xie, and L. Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, pages 672–681, 2021. 3

[77] T. Yang, R. Wu, P. Ren, X. Xie, and L. Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023. 2

[78] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[79] Z. You, Z. Li, J. Gu, Z. Yin, T. Xue, and C. Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. *arXiv preprint arXiv:2312.08962*, 2023. 8

[80] F. Yu, J. Gu, Z. Li, J. Hu, X. Kong, X. Wang, J. He, Y. Qiao, and C. Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *CVPR*, 2024. 2, 6, 7, 8, 19, 20, 21

[81] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *CVPRW*, pages 701–710, 2018. 2

[82] Z. Yue, J. Wang, and C. C. Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *NeurIPS*, 2023. 7, 8

[83] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling vision transformers. In *CVPR*, pages 12104–12113, 2022. 4

[84] J. Zhang, Y. Zhang, J. Gu, J. Dong, L. Kong, and X. Yang. Xformer: Hybrid x-shaped transformer for image denoising. In *ICLR*, 2024. 3
[85] K. Zhang, J. Liang, L. Van Gool, and R. Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, pages 4791–4800, 2021. 2, 3, 7, 8
[86] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *TIP*, 26(7):3142–3155, 2017. 7
[87] K. Zhang, W. Zuo, and L. Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *TIP*, 27(9):4608–4622, 2018. 3
[88] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 5
[89] L. Zhang, L. Zhang, and A. C. Bovik. A feature-enriched completely blind image quality evaluator. *TIP*, 24(8):2579–2591, 2015. 7
[90] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 7
[91] X. Zhang, Q. Chen, R. Ng, and V. Koltun. Zoom to learn, learn to zoom. In *CVPR*, pages 3762–3770, 2019. 2
[92] Y. Zhang, K. Zhang, Z. Chen, Y. Li, R. Timofte, J. Zhang, K. Zhang, R. Peng, Y. Ma, L. Jia, et al. Ntire 2023 challenge on image super-resolution (x4): Methods and results. In *CVPRW*, pages 1864–1883, 2023. 2
[93] Z. Zheng, X. Peng, and Y. You. Open-sora: Democratizing efficient video production for all, 2024. 5
[94] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 9
[95] X. Zhou, H. Huang, R. He, Z. Wang, J. Hu, and T. Tan. Msra-sr: Image super-resolution transformer with multi-scale shared representation acquisition. In *ICCV*, pages 12665–12676, 2023. 3
[96] X. Zhou, H. Huang, Z. Wang, and R. He. Ristra: Recursive image super-resolution transformer with relativistic assessment. *TMM*, 26(8):6475–6487, 2024. 3
[97] X. Zhou, J. Li, Z. Wang, R. He, and T. Tan. Image inpainting with contrastive relation network. In *ICPR*, pages 4420–4427, 2021. 3

A Appendix

A.1 More Implementation Details

For generating negative samples, the strength in SDEdit [50] is set to 0.6. To minimize the risk of generating images that contain private information in GenIR, we employ Gemini for both text prompt filtering and generated image filtering. The prompts for Gemini are set as "*You are an AI language assistant, and you are analyzing a series of text prompts. Your task is to identify whether these text prompts contain any inappropriate content such as personal privacy violations or NSFW material. Delete any inappropriate text prompts and return the remaining ones in their original format.*" and "*You are an AI visual assistant, and you are analyzing a single image. Your task is to check the image for any anomalies, irregularities, or content that does not align with common sense or normal expectations. Additionally, identify any inappropriate content such as personal privacy violations or NSFW material. If the image does not contain any of the aforementioned issues, it has passed the inspection. Please determine whether this image has passed the inspection (answer yes/no) and provide your reasoning.*", respectively.

A.2 More Details of User Study

To evaluate our approach, we conducted a user study emphasizing restoration quality. We randomly selected 100 low-quality images from the test sets (Tab. 1), applying our model and five other leading methods to produce restored images, generating 100 groups of seven images each. Users, guided by the low-quality image, were asked to select the best-restored image from each group, considering factors such as visual quality, naturalness, detail accuracy, and absence of distortions or artifacts. Fairness was ensured by presenting each user with 10 randomly selected groups, randomizing image sequence within each group, and masking the methods. We widely disseminated the online questionnaire without restrictions, amassing feedback from 256 evaluators.

We employed two metrics to quantify our study's results: vote percentage and Top-K ratio. The former represents the proportion of total votes each method received, while the latter measures the frequency a method was among the top-k selections, indicating the proportion of most preferred images and a model's consistency in producing high-quality images. The Top-K ratio is defined as: $R_i^k = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}(i \in F_{topk}(s_j, k))$, where $s_j = \{s_{ij} | i = 0, ..., 5\}$ are the selection scores of the j-th group from N groups (with $N = 100$), and F_{topk} is the top-k operation.

As shown in Fig. 5, our model outperformed others on both metrics. Our model received over 45% of the total votes, indicating a strong user preference. Additionally, our method was the top choice for 80% of the images, and it was ranked first or second for 98% of the images, highlighting our method's reliable ability to generate superior quality images compared to other methods.

A.3 More Ablations

More analysis of DreamClear Ablation. We give a more detailed analysis in the following.

(a) Mixture of Adaptive Modulator (MoAM). MoAM acts as an interaction module between the LQ branch and the reference branch, aiming to enhance the model's robustness to intricate real-world degradations by explicitly guiding it through the introduction of token-wise degradation priors. It obtains the degradation map through cross-attention between the LQ features and reference features, guiding the dynamic fusion of expert knowledge to modulate LQ features.

Tab. 3 presents the ablation studies of MoAM. Notably, when we substitute the Mixture of AM design with AM, all metrics undergo a substantial decrease, underscoring the importance of the degradation prior guidance in MoAM for steering restoration experts. Moreover, we conducted experiments replacing AM with cross-attention and a zero-linear layer. The use of cross-attention, in comparison to AM, leads to a general reduction in model performance. While the zero-linear layer provides minor improvements in MANIQA and MUSIQ scores, it results in a significant drop in perceptual fidelity.

(b) Dual-branch framework. The incorporation of a reference branch allows the model to focus less on degradation removal and more on enhancing image details through generative priors, ultimately producing more photorealistic images. The results in Tab. 3 indicate that our dual-branch structure significantly outperforms using only the LQ branch across all metrics.

Figure 7: Visual comparisons for ablation study on DreamClear.

Table 4: Ablations for GenIR on *LSDIR-Val* using SwinIR-GAN.

Training Data	LPIPS ↓	DISTS ↓	FID ↓	MANIQA ↑	MUSIQ ↑	CLIPIQA ↑
Pre-trained T2I Model (3450images)	0.4819	0.2790	60.12	0.3271	61.94	0.5423
Ours GenIR (3450images)	**0.4578**	**0.2435**	**51.29**	**0.3691**	**63.12**	**0.5647**

(c) Text prompt guidance. We evaluate the impact of using detailed text prompts generated by MLLMs versus null prompts. Despite null prompts slightly outperforming detailed prompts on perceptual metrics, the superior performance on no-reference and high-level vision metrics indicates that text prompts more effectively retain semantic information. Visual comparisons in Fig. 7 further underscore the advantages of text prompts in providing semantic guidance for image restoration.

Visual results for DreamClear Ablation. Following the setting in Tab. 3, we provide more visual comparison results in Fig. 7. We find that when using a null prompt instead of a text prompt generated by MLLM, there are significant semantic errors in the eyes of the bear in the restoration results. This demonstrates that the semantic information provided by MLLM-generated detailed text prompts helps the model achieve more ideal restoration results. When using AM, zero-linear, and cross-attention instead of MoAM, the model tends to produce results that are either too smooth or contain semantic errors, proving the effectiveness of MoAM. Removing the reference branch results in a significant deterioration of the restoration outcomes. Overall, our full model, DreamClear, achieves the best results in terms of fidelity and perception.

Ablations for GenIR. We use the exact same prompts and sampling parameters to compare images generated by GenIR with those generated by the originally pre-trained T2I model. As shown in Fig. 10, the images generated by our proposed GenIR are more realistic and contain more texture details, while those generated by the pre-trained T2I model tend to have issues like being overly smooth and blurry. Intuitively, images generated by GenIR are likely to be more helpful for real-world IR. To verify this, following the setting of Fig. 6, we use manually designed prompts with the pre-trained T2I model to generate 3450 images for training SwinIR-GAN. As shown in Tab. 4, the model trained using images generated by our GenIR shows significant improvements across all metrics, quantitatively demonstrating the effectiveness of our approach.

In addition, we provide visual comparisons in Fig. 8 to verify the effectiveness of dual-prompt learning in GenIR. It shows that the dual-prompt learning strategy can effectively enhance image texture details, making the generated images more suitable for image restoration training. Fig. 9 also demonstrates the effectiveness of our generated data in enhancing the visual effect of IR models.

Figure 8: Visual comparisons for ablation study on GenIR.

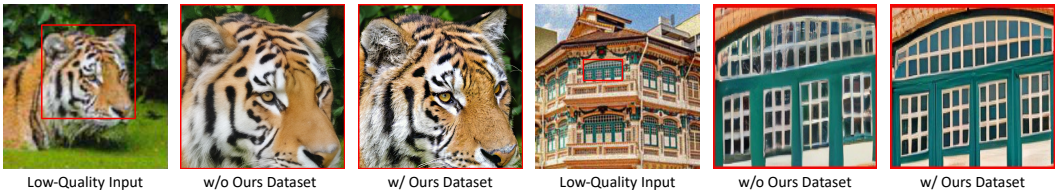

Figure 9: Visual comparisons for ablation study on training datasets.

A.4 More Real-World Visual Comparisons

We provide more real-world visual comparisons with state-of-the-art diffusion-based real-world image restoration methods in Fig. 11, Fig. 12 and Fig. 13.

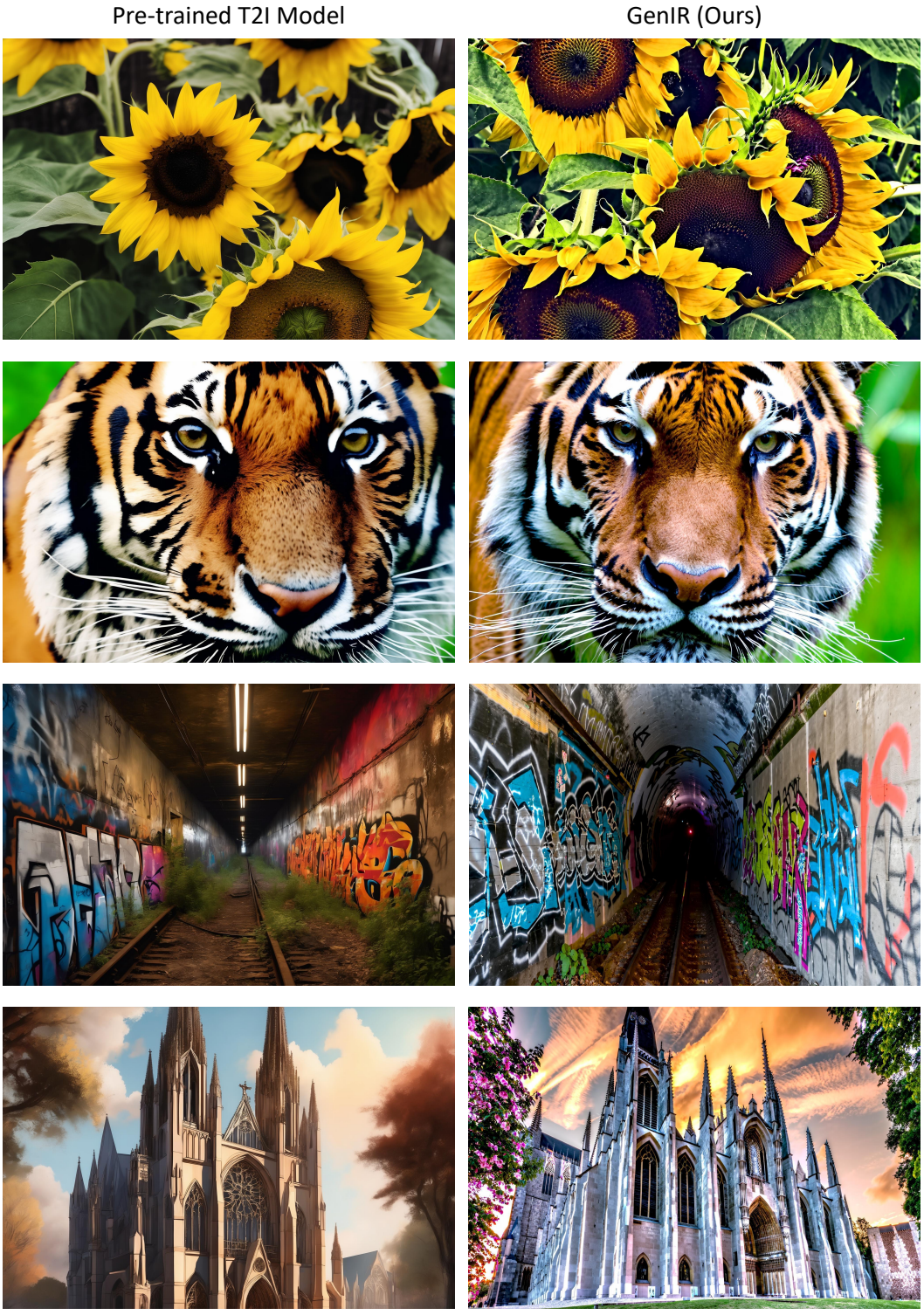

Figure 10: Visual comparison of images generated using the pre-trained T2I model and GenIR. Our proposed GenIR produces images with enhanced texture and more realistic details, exhibiting less blurring and distortion. This makes it better suited for training real-world IR models.

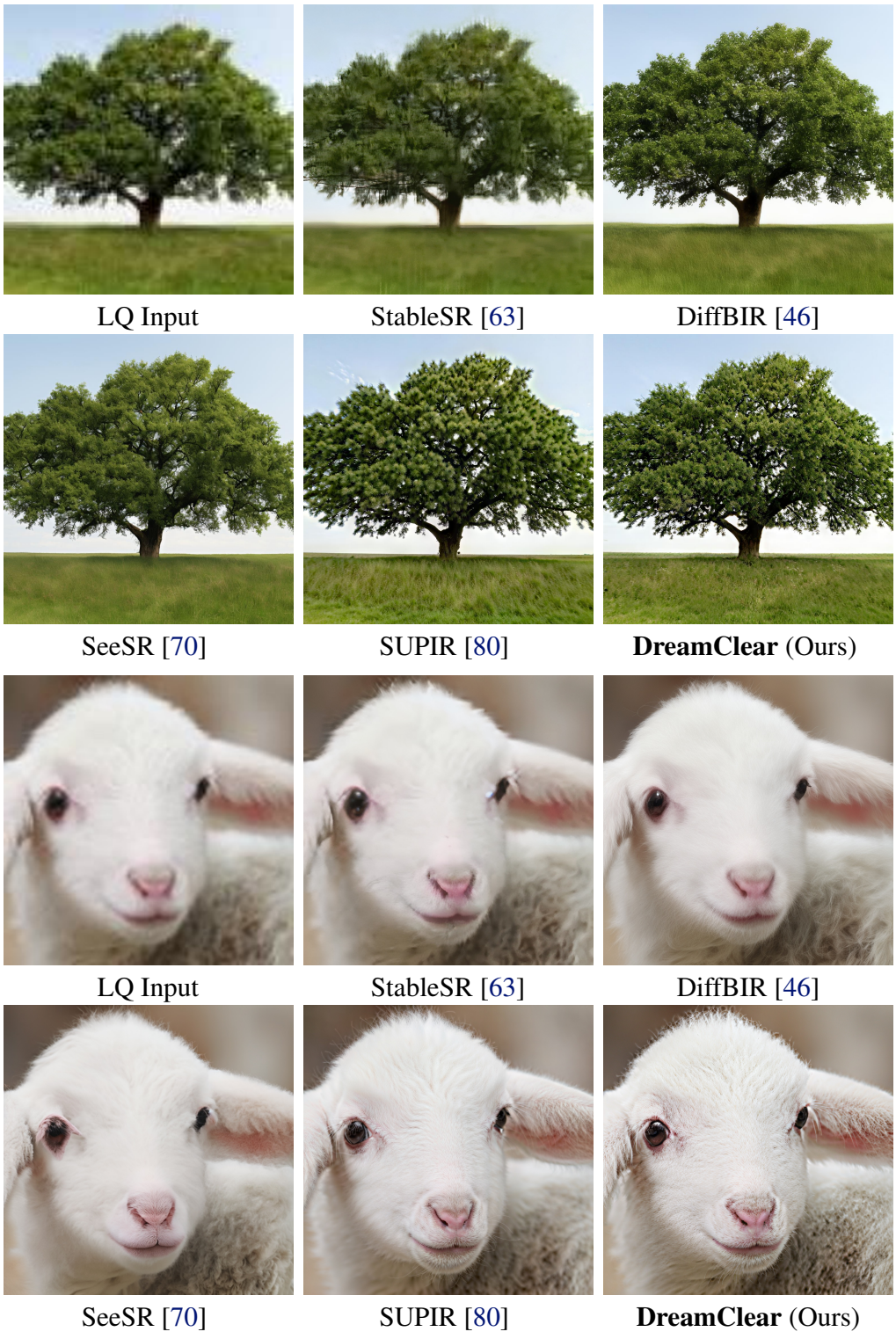

Figure 11: Visual comparisons on real-world benchmarks (1/3). Please zoom in for a better view.

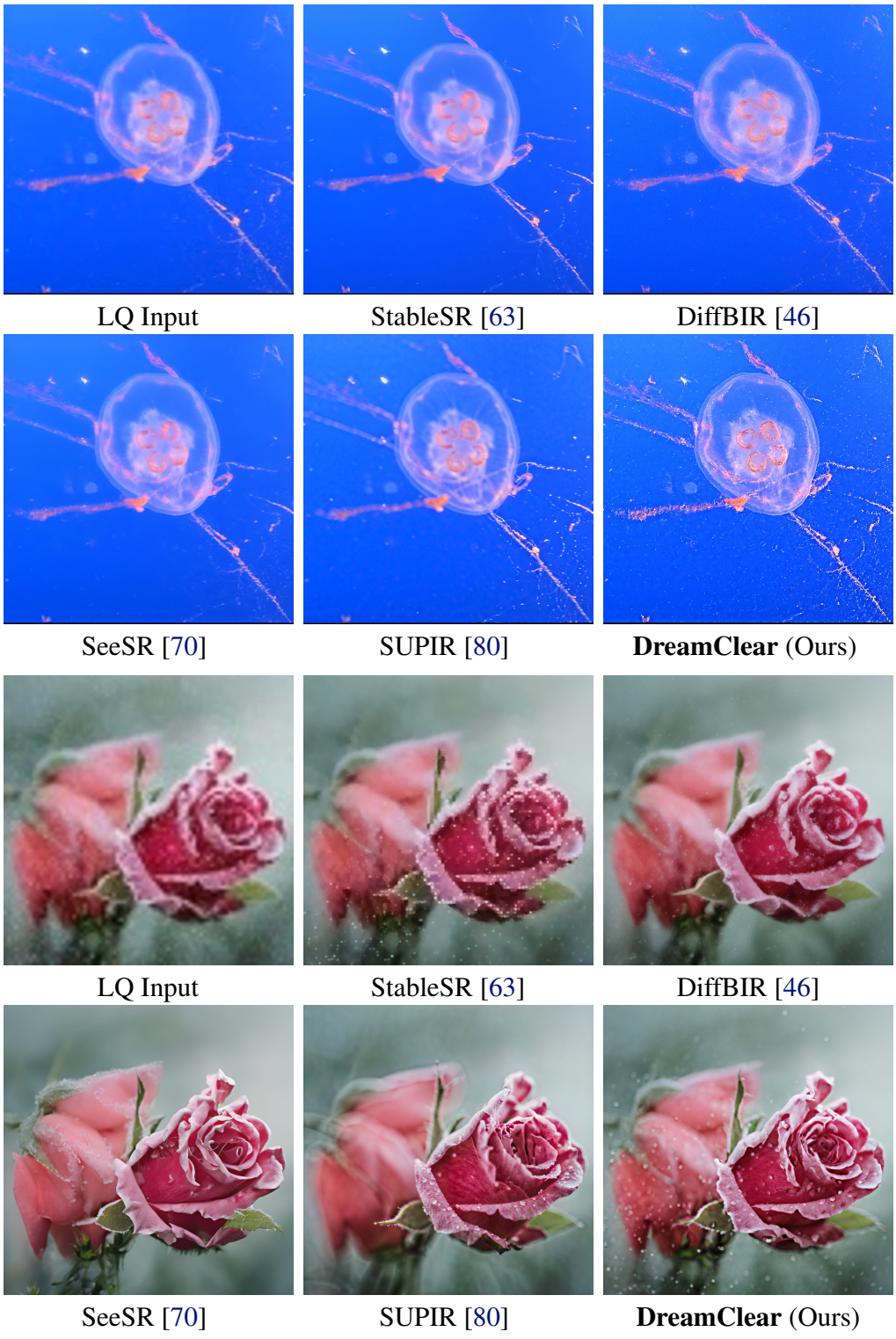

Figure 12: Visual comparisons on real-world benchmarks (2/3). Please zoom in for a better view.

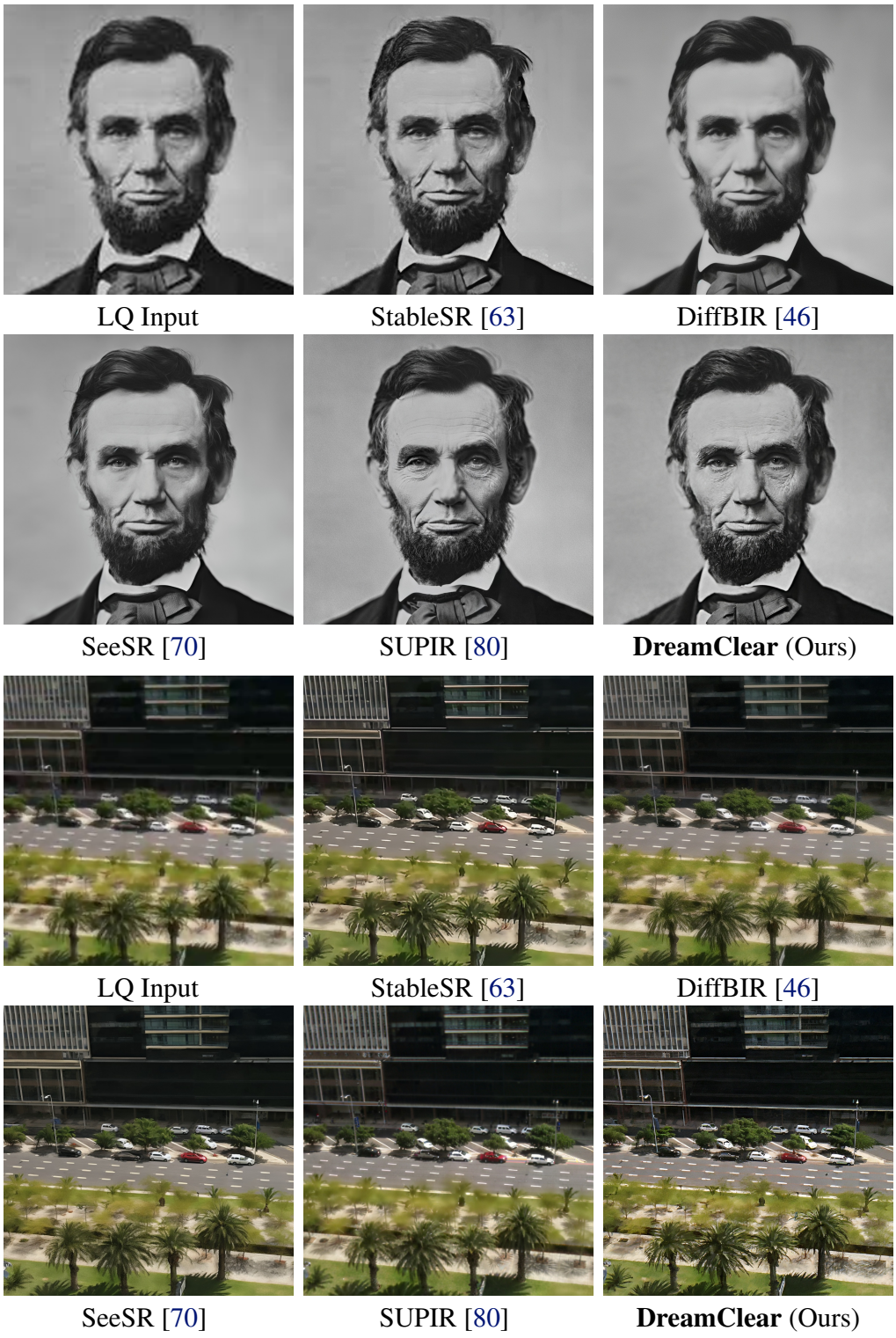

Figure 13: Visual comparisons on real-world benchmarks (3/3). Please zoom in for a better view.

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 accurately reflect the paper's contributions and scope, and we summarize the contributions at the end of Section 1.

   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 have discussed the limitations of the work in Section 6.

   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: There isn't theoretical derivation in our work.

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 Section 4, and present the detailed experimental settings and implementation details in Section 5.1.

   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: The code and pre-trained models are available at https://github.com/shallowdream204/DreamClear.

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 5.1.

   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 dataset and the inference speed of diffusion models is relatively slow. Due to limited computational resources, we can't afford the statistical significance experiments.

   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 5.1.

   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: [Yes]

    Justification: We have discussed both potential positive societal impacts and negative societal impacts of the work in Section 6.

    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 in our paper.

    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: [NA]

    Justification: Our paper does not release new assets.

    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: [Yes]

    Justification: We have included the full text of instructions given to participants in the paper.

    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: [Yes]

    Justification: We adhere to the NeurIPS Code of Ethics and the guidelines for our institution.

    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.