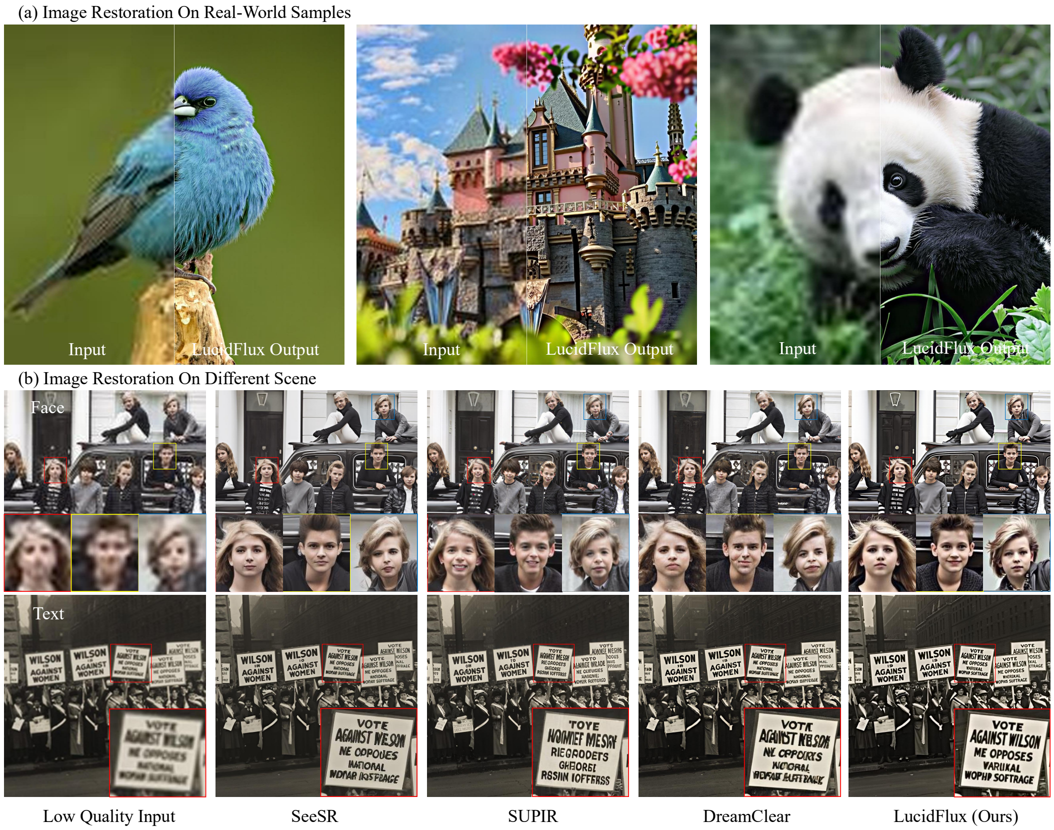
LucidFlux: Caption-Free Universal Image Restoration via a Large-Scale Diffusion Transformer

Song Fei[†]
The Hong Kong University of Science and Technology (Guangzhou)
sfei285@connect.hkust-gz.edu.cn

Tian Ye[†,‡]
The Hong Kong University of Science and Technology (Guangzhou)
tye610@connect.hkust-gz.edu.cn

Lujia Wang
The Hong Kong University of Science and Technology (Guangzhou)
eewanglj@hkust-gz.edu.cn

Lei Zhu[*]
The Hong Kong University of Science and Technology
The Hong Kong University of Science and Technology (Guangzhou)
leizhu@ust.hk


(a) Image Restoration On Real-World Samples

(b) Image Restoration On Different Scene

Figure 1: We present LucidFlux, a universal image restoration framework built on a large-scale diffusion transformer that delivers photorealistic restorations of real-world low-quality (LQ) images, outperforming state-of-the-art (SOTA) diffusion-based models across diverse degradations.

[†] Equal contribution [‡] Project Leader [*] Corresponding author.

Abstract

Universal image restoration (UIR) aims to recover images degraded by unknown mixtures while preserving semantics—conditions under which discriminative restorers and UNet-based diffusion priors often oversmooth, hallucinate, or drift. We present ***LucidFlux***, a caption-free UIR framework that adapts a large diffusion transformer (Flux.1) without image captions. Our LucidFlux introduces a lightweight *dual-branch conditioner* that injects signals from the degraded input and a lightly restored proxy to respectively anchor geometry and suppress artifacts. Then, a *timestep- and layer-adaptive modulation* schedule is designed to route these cues across the backbone's hierarchy, in order to yield coarse-to-fine and context-aware updates that protect the global structure while recovering texture. After that, to avoid the latency and instability of text prompts or MLLM captions, we enforce *caption-free semantic alignment* via SigLIP features extracted from the proxy. A scalable curation pipeline further filters large-scale data for structure-rich supervision. Across synthetic and in-the-wild benchmarks, our LucidFlux consistently outperforms strong open-source and commercial baselines, and ablation studies verify the necessity of each component. LucidFlux shows that, for large DiTs, *when, where, and what* to condition on—rather than adding parameters or relying on text prompts—is the governing lever for robust and caption-free universal image restoration in the wild.

1 Introduction

Images acquired in the wild exhibit mixed, unknown degradations—sensor noise, motion blur, lens aberrations, compression artifacts—that erode perceptual fidelity and induce semantic drift in recognition and analysis. *Universal image restoration (UIR)* seeks to reconstruct images with high perceptual fidelity while preserving semantic consistency under such uncertainty and without access to degradation labels or side information. Despite steady progress, this combination of unknown mixtures, realism, and semantic preservation remains stubbornly challenging.

Discriminative restorers based on CNNs and Transformers Dong et al. (2016); Liang et al. (2021); Zamir et al. (2022) perform well on synthetic distortions but falter on in-the-wild mixtures, often oversmoothing textures or leaving visible artifacts. This gap has motivated generative approaches that leverage diffusion-based text-to-image priors to synthesize plausible structure and detail beyond the reach of purely discriminative models Yu et al. (2024); Ai et al. (2024); Wu et al. (2024b); Wang et al. (2024a;b); Yue et al. (2023); Wu et al. (2024a); Lin et al. (2025). Yet most such systems rely on Stable Diffusion (SD) UNet backbones Rombach et al. (2022): their capacity and inductive bias saturate under complex degradations, making it difficult to recover fine detail while maintaining global structure—suggesting the need to look beyond UNet-based designs.

Recent advances in diffusion transformers (DiTs) open a promising avenue. In contrast to UNet architectures, DiTs employ attention-centric backbones that more effectively couple global context with local detail and carry richer generative priors. For instance, DreamClear Ai et al. (2024) builds on PixArt-α Chen et al. (2023), a relatively small (0.6B) DiT, illustrating the promise of transformer backbones for restoration. However, their limited scale constrains robustness to mixed, real-world degradations and impedes the concurrent recovery of global structure and fine detail. Large-scale diffusion transformers such as Flux.1 Labs (2024) deliver strong modeling capacity for universal restoration, yet ***direct transfer rarely works off-the-shelf***. Previous ControlNet-style conditioning methods Yu et al. (2024); Ai et al. (2024); Zhang et al. (2023) ***disrupt the parameter–structure balance*** and underutilize the backbone's temporal and hierarchical division of labor. Unconstrained injection of degraded observations amplifies artifacts; relying on VLM-generated captions further ***increases latency and risks semantic drift***[0]. Meanwhile, backbones at this scale are ***decisively data-limited***: gains follow data–compute scaling only when trained on ***curated, large-scale, high-quality*** sets. Public web corpora fall short for UIR—they skew toward aesthetic, compression-heavy

[0] Appendix Sec. A.1 quantifies the prevalence of degradation-related terms in MLLM captions, and Appendix Fig. 5 demonstrates how such bias can misguide restoration.

images, contain substantial near-duplicates and low-information frames, and rarely cover the long-tail mixtures of real degradations or provide usable pairs. Without rigorous filtering and structure-aware selection, large DiTs underutilize capacity and overfit spurious artifacts, underscoring the need for an explicit curation pipeline. Taken together, these tensions point to a more structured path, one that schedules conditioning across timesteps and layers, couples robust input handling with caption-free inference, and remains practical to assemble on available datasets.

To operationalize this path, we introduce ***LucidFlux***, a caption-free UIR framework that adapts the large-scale Flux.1 diffusion transformer to restoration. The core of our ***LucidFlux*** is a *lightweight dual-branch conditioner*—a two-block transformer module that injects signals from the degraded input without inflating the parameters. One branch ingests the low-quality image to anchor the geometry and layout, while the other consumes a lightly restored proxy to suppress hard artifacts; their outputs are scheduled through a *timestep- and layer-adaptive modulation* that aligns guidance with the backbone's hierarchical roles, yielding coarse-to-fine, context-aware updates that preserve texture while protecting global structure. To avoid the latency and drift introduced by text prompts, we enforce semantic consistency via *caption-free alignment with SigLIP*, extracting semantic cues directly from the proxy. We pair the model with an automated three-stage curation pipeline—blur detection, flat-region filtering, and perceptual quality scoring—to assemble diverse training sets at the *billion-parameter scale.*

Our contributions are as follows:

- **LucidFlux framework**. We adapt a large diffusion transformer (Flux.1) to UIR with a lightweight dual-branch conditioner and timestep- and layer-adaptive modulation, aligning conditioning with the backbone's hierarchical roles while keeping less trainable parameters.
- **Caption-free semantic alignment**. A SigLIP-based module preserves semantic consistency without prompts or captions, mitigating latency and semantic drift.
- **Scalable data curation pipeline.** A reproducible, three-stage filtering pipeline yields diverse, structure-rich datasets that scale to billion-parameter training.
- **State-of-the-art results**. LucidFlux sets new SOTA on a broad suite of benchmarks and metrics, surpassing competitive open- and closed-source baselines; ablation studies confirm the necessity of each module.

2 Related Work

Generative Priors for UIR. Large-scale pretrained generative models, particularly text-to-image diffusion transformers Rombach et al. (2022); Podell et al. (2023), have shown strong capability in synthesizing high-fidelity textures and structures for image restoration. Existing approaches build on different backbones, with SUPIR Yu et al. (2024) using SDXL, DreamClear Ai et al. (2024) relying on PixArt-α Chen et al. (2023), StableSR on SD, SeeSR on SD2, and Resshift Yue et al. (2023) and SinSR Wang et al. (2024b) trained from scratch. While these methods perform well on their respective backbones, they struggle to scale as text-to-image models continue to grow in size, limiting both their performance and expressive capacity. Addressing these challenges, we propose **LucidFlux**, a universal image restoration framework built on the large-scale Flux.1 backbone, which leverages richer generative priors and greater expressive capacity.

Semantic Alignment. Preserving semantic fidelity during image restoration remains a significant challenge. Existing methods often rely on generating captions from degraded images using vision–language models at inference time Yu et al. (2024); Ai et al. (2024); Kong et al. (2025), which introduces additional computational cost and may produce inconsistencies between training and inference. Alternative strategies employ coarse textual cues Wu et al. (2024b;a), but such signals are generally insufficient to capture fine-grained semantic content. In contrast, LucidFlux leverages a SigLIP-based semantic alignment module that extracts rich semantic representations directly from lightly restored images, facilitating caption-free guidance and ensuring that restored outputs maintain high semantic consistency without hallucinations.

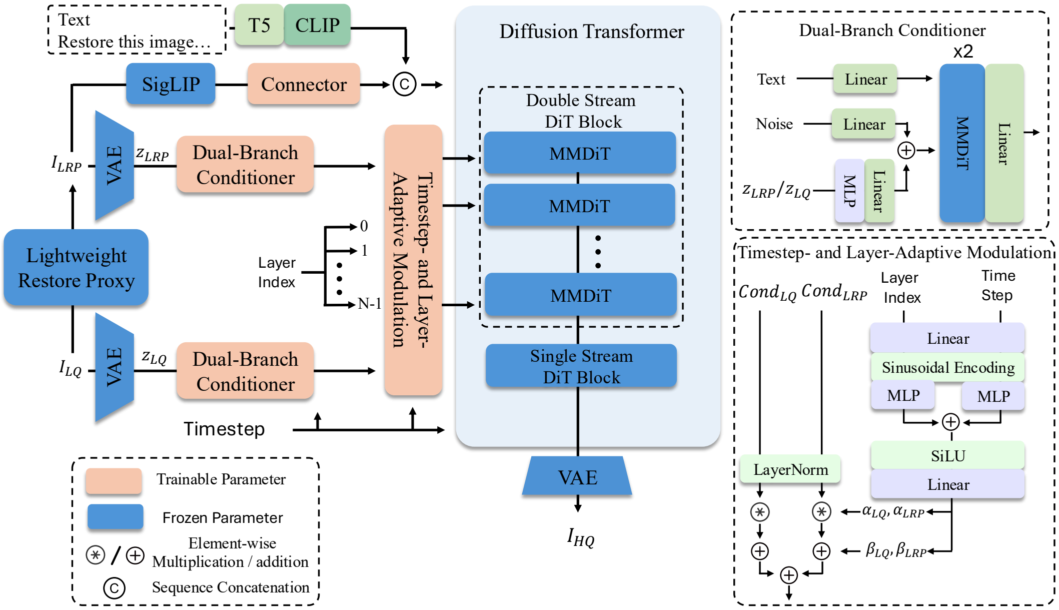


Figure 2: Overview of the proposed architecture for universal image restoration. Our method integrates dual condition streams (LQ and LRP) with timestep- and layer-adaptive modulation modules, and incorporates SigLIP semantic priors through a connector into a Flux-based DiT backbone to jointly enhance perceptual quality and semantic consistency.

3 Methodology

Our framework is built upon a Flux-based DiT backbone, augmented with two parallel ControlNet branches. The first branch processes the original low-quality image (LQ), while the second branch takes a lightly restored version of the input (LRP) generated by a lightweight restoration model. Both streams capture complementary information, which is subsequently modulated through timestep- and layer-adaptive modules to align with the DiT feature space. Moreover, we incorporate semantic priors extracted from SigLIP and enhanced with a connector, which are injected into the DiT layers to facilitate semantic consistency and fine-grained texture restoration.

Practically, we eschew inference-time captions: Appendix Sec. A.1 quantifies that 17–24% of MLLM captions introduce degradation-related terms, and Appendix Fig. 5 shows such bias misguides restoration, degrading perceptual quality.

3.1 Lightweight Dual-Branch Conditioner

Directly conditioning on the low-quality (LQ) image preserves high-frequency details but often leaks residual artifacts under mixed degradations; conditioning on a lightly restored proxy (LRP) suppresses artifacts but tends to oversmooth textures. Following the dual-branch paradigm proposed in Ai et al. (2024), we therefore decouple *structure anchoring* and *artifact suppression* into two signals and encode them with a *minimal-overhead* conditioner that interfaces with the Flux.1 DiT backbone without duplicating large blocks. As illustrated in the top-right of Fig. 2, only the core conditioning pathway is shown for clarity, while other components such as timestep embeddings remain consistent with Flux. Throughout, the Flux.1 backbone and VAE remain frozen for stability and efficiency. The output feature map I_{LRP} of the LRP is computed by:

$$I_{\mathrm{LRP}} = \mathrm{LRP}(I_{\mathrm{LQ}}). \tag{1}$$

Both I_{LQ} and I_{LRP} are mapped by the shared Flux.1 VAE encoder E into latents

$$z_{\mathrm{LQ}} = E(I_{\mathrm{LQ}}), \qquad z_{\mathrm{LRP}} = E(I_{\mathrm{LRP}}), \qquad z \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times d}, \tag{2}$$

where s is the VAE downsampling factor and d the latent width. The *Dual-Branch Conditioner (DBC)* then converts each latent into compact conditioning tokens via a two-block MMDiT applied at latent resolution,

$$\phi_{\text{LQ}} = \text{DBC}(z_{\text{LQ}}), \qquad \phi_{\text{LRP}} = \text{DBC}(z_{\text{LRP}}), \tag{3}$$

where DBC *patchifies* the VAE latent and projects patches into a 2D-positioned sequence before two stacked transformer blocks; *weights are not shared across branches*, as the LQ stream emphasizes detail-preserving, noise-tolerant cues while the LRP stream favors structure-first, artifact-suppressed representations. Using separate parameters avoids competing gradients and preserves branch complementarity, while the conditioner remains minimal (two blocks per branch, constant overhead w.r.t. layer depth and far smaller than ControlNet-style duplication of a large DiT). Intuitively, ϕ_{LQ} carries detail-preserving yet noisy cues, whereas ϕ_{LRP} provides artifact-robust structure; the subsequent *timestep- and layer-adaptive condition modulation* (Sec. 3.2) consumes these two complementary signals for coarse-to-fine, context-aware guidance without increasing the conditioner's footprint.

3.2 Timestep- and Layer-Adaptive Condition Modulation

Diffusion transformers exhibit a temporal–hierarchical division of labor: early timesteps reconstruct coarse structures while later ones refine high-frequency details; similarly, shallower layers capture low-level edges and deeper layers process semantics Park et al. (2023); Qian et al. (2024). Applying identical conditioning across all timesteps and layers risks redundancy or conflict.

We therefore modulate the outputs of the *dual-branch conditioner (DBC)* in a way that is *adaptive to both timestep t and layer index l* while keeping the heavy Flux.1 backbone frozen. Let ϕ_{LQ} and ϕ_{LRP} denote the features produced by DBC from the low-quality input (LQ) and the lightly restored proxy (LRP). A lightweight modulation head takes sinusoidally encoded $(t/T,\, l/L)$ and predicts *feature-wise* (per-channel) scale and bias for each branch independently:

$$\alpha_{\bullet}^{t,l},\, \beta_{\bullet}^{t,l} \;=\; \text{Modulation}_{\bullet}\big(\text{PE}(t/T,\, l/L)\big), \quad \bullet \in \{\text{LQ}, \text{LRP}\},\ \alpha_{\bullet}^{t,l}, \beta_{\bullet}^{t,l} \in \mathbb{R}^{d_c}. \tag{4}$$

These parameters effect an AdaptiveLN-style adjustment,

$$\tilde{\phi}_{\text{LQ}}^{t,l} = \alpha_{\text{LQ}}^{t,l} \odot \phi_{\text{LQ}} + \beta_{\text{LQ}}^{t,l}, \qquad \tilde{\phi}_{\text{LRP}}^{t,l} = \alpha_{\text{LRP}}^{t,l} \odot \phi_{\text{LRP}} + \beta_{\text{LRP}}^{t,l}, \tag{5}$$

We then fuse the branches without additional normalization via a feature-wise convex mixture:

$$\gamma^{t,l} = \sigma\big(\text{Gate}\big(\text{PE}(t/T,\, l/L)\big)\big), \quad \gamma^{t,l} \in (0,1)^{d_c}, \qquad \text{Cond}^{t,l} = \gamma^{t,l} \odot \tilde{\phi}_{\text{LQ}}^{t,l} + \big(1 - \gamma^{t,l}\big) \odot \tilde{\phi}_{\text{LRP}}^{t,l}. \tag{6}$$

Predicting α/β *per channel* supplies sufficient flexibility to track the backbone's roles across t and l without inflating capacity, and *independent* modulation for LQ vs. LRP preserves their complementary inductive biases (detail-preserving vs. artifact-robust). Keeping modulation inside the lightweight conditioner maintains negligible overhead while enabling coarse-to-fine, timestep- and layer-aware guidance; ablations (Sec. 4.3) show that removing either the temporal or the hierarchical dependency degrades fidelity under mixed degradations.

3.3 SigLIP for Caption-Free Semantic Alignment

T2I diffusion models are typically conditioned on text, and many restoration methods adopt captions as semantic guidance Ai et al. (2024); Yu et al. (2024). During training, such captions are often derived from clean ground truth, yielding idealized supervision. At inference, however, only degraded inputs are available; captions generated from low-quality images tend to inherit degradation-specific artifacts and, when produced by large VLMs, add substantial latency and exacerbate a train–test mismatch Sun et al. (2024).

We replace caption generation with a *caption-free* semantic pathway. Concretely, we extract image semantics from the lightly restored proxy I_{LRP} using a frozen SigLIP encoder and map them into the backbone's textual embedding space via a lightweight *Connector*:

$$z_s = \text{Connector}\big(\text{SigLIP}(I_{\text{LRP}})\big). \tag{7}$$

The projected semantics z_s are concatenated with a small set of prompt tokens c (default instruction) to form the multimodal context fed to the DiT backbone:

$$\textit{Context} = \text{Concat}(z_s,\, c). \tag{8}$$

Grounding semantics in I_{LRP} stabilizes content under mixed degradations, while the Connector furnishes a drop-in bridge to the text-conditioning interface, avoiding any duplication of heavy modules. This design eliminates external captions at both training and inference, reducing latency and removing a major source of caption-induced semantic variance (e.g., differences across captioners or paraphrases). Grounding semantics in z_s keeps outputs structurally faithful and semantically aligned to the input.

3.4 Scaling Up Real-world High-Quality Data for Universal Image Restoration

Although large-scale text-to-image (T2I) diffusion models are pretrained on hundreds of millions of image–text pairs, they are not tailored for the universal image restoration task of our work. Training large diffusion transformers for UIR requires *task-aligned data at scale with strong structure and perceptual quality*. However, publicly available restoration corpora remain modest and/or lack reproducible quality control: DIV2K Agustsson & Timofte (2017a) (800/100), Flickr2K Agustsson & Timofte (2017b) (2,650), LSDIR Li et al. (2023) ($\approx$ 85K with manual curation), and SUPIR Yu et al. (2024) (20M without disclosed filtering criteria), while DreamClear Ai et al. (2024) synthesizes 1M pairs at substantial computational cost. This leaves a practical gap between what large DiTs need and what current datasets provide.

To bridge this gap, we introduce, to our knowledge, the first publicly documented and extensively validated *UIR-specific data filtering pipeline*. It is *fully automatic* (parameters empirically set, pipeline automatic once fixed) and comprises three stages—blur screening, flat-region suppression, and perceptual-quality ranking—explicitly designed to retain structure-rich, high-quality images while discarding unsuitable samples.

Data source. Our initial dataset is collected from two sources. First, we collect 2.3M images from the Internet. In addition, we incorporate 557K images from the Photo-Concept-Bucket dataset bghira (2023), yielding a total of 2.9M candidate images. This combined pool serves as the raw data for subsequent filtering.

Blur detection. Images that are heavily blurred or contain excessive high-frequency noise provide unreliable structural cues and are thus unsuitable for training. Following LSDIR Li et al. (2023), we quantify the degree of blur using the variance of the Laplacian $S_{\text{blur}}(I) = \mathrm{Var}\big(\nabla^2 I\big)$, where I denotes an input image. Only images with $150 \leq S_{\text{blur}}(I) \leq 8000$ Li et al. (2023) are retained, effectively excluding both overly blurred and noisy samples. These bounds are empirically hand-tuned based on preliminary experiments and careful visual audits on a held-out subset to balance removal of extreme blur/noise while retaining legitimate shallow-depth-of-field and low-light scenes.

Flat-region detection. Images dominated by textureless regions may bias the model towards producing over-smoothed outputs. To mitigate this, each image is divided into non-overlapping 240×240 patches, and the edge richness of each patch is measured using the Sobel operator with $S_{\text{flat}} = \mathrm{Var}\Big(\sqrt{(\partial_x I)^2 + (\partial_y I)^2}\Big)$. Patches with $S_{\text{flat}} < 800$ are considered textureless, and images containing more than 50% such patches are discarded. Both the 800 patch-level threshold and the 50% image-level ratio are empirically set by manual inspection of edge-statistics distributions and visual audits; they provide a conservative balance that suppresses large flat backgrounds yet preserves natural sky/water regions. This ensures that retained images exhibit sufficient edge and texture diversity, essential for high-fidelity restoration. After applying blur and flat-region filtering, 1.28M candidate images remain.

IQA Filtering for High-quality Data. While LSDIR employs manual curation in its final stage, such human intervention is impractical for scaling to larger datasets. We apply CLIP-IQA to further ensure perceptual quality of our training data. The remaining images are ranked by their perceptual scores s_i, and only the top 20% are retained, i.e., $\{i \mid s_i \geq \text{quantile}_{0.8}(\{s_i\})\}$, resulting in 257K high-quality images. The 20% cutoff is empirically chosen after careful inspection at multiple percentiles (e.g., 10/20/30%), trading off perceptual quality against semantic/content diversity. By additionally incorporating 84K high-quality samples from LSDIR Li et al. (2023), the final curated dataset comprises 342K high-quality images. Once these cutoffs are fixed, the pipeline executes fully automatically at scale. For generating paired training data, degraded counterparts are synthesized using the Real-ESRGAN degradation pipeline Wang et al. (2021) as implemented in Ai et al. (2024),

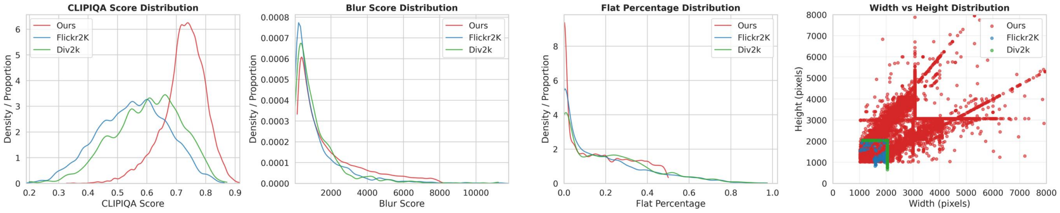


Figure 3: Comparison of dataset attributes. Our dataset exhibits higher CLIP-IQA scores, lower flatness, and more diverse resolutions than Flickr2K Agustsson & Timofte (2017b) and DIV2K Agustsson & Timofte (2017a).

across 4 epochs, producing a total of 1.36M image pairs. This procedure ensures both diversity and realism in the low-quality inputs, facilitating effective model training. To assess the effectiveness of our filtered data, we randomly select 10K samples and compare their attribute distributions with existing datasets. Figure 3 shows that our dataset achieves higher CLIP-IQA scores, comparable blur scores, lower flatness values that reflect richer textures, and more diverse resolutions than Flickr2K and DIV2K. In Appendix 6, we also analyze semantic diversity using t-SNE, and it shows that our dataset demonstrates substantially broader semantic coverage.

4 Experiment

4.1 Implementation Details

We train a large Flux-based generative model, LucidFlux, while *freezing all blocks of the Flux backbone* and training only the task-specific modules introduced by our method. Freezing the backbone stabilizes optimization and prevents catastrophic forgetting, while concentrating capacity on the new modules that realize our objective. Training runs on 8×NVIDIA A800 GPUs with DeepSpeed ZeRO-2. We choose ZeRO-2 because it shards optimizer states and gradients—dramatically reducing memory footprint—without partitioning model parameters, which preserves simple forward passes and yields higher throughput than ZeRO-3 in our setting. This enables larger activation budgets at 1024×1024 resolution and steady scaling with modest communication overhead. We use Adafactor Shazeer & Stern (2018) with a learning rate of 2×10^{-5} and weight decay 0.01. The per-GPU batch size is 2 with gradient accumulation of 2 steps, giving an effective batch size of 32 across 8 GPUs. We resume our Siglip connector based on Flex.1-alpha-Redux checkpoint. The full training completes in approximately 7 GPU-days. Following many existing works, we employ SwinIR Liang et al. (2021) as a lightweight restore proxy.

4.2 Comparison with Open-Source State-of-the-Art Methods

We evaluate our approach against several state-of-the-art diffusion-based methods, including ResShift(Yue et al. (2023)), StableSR(Wang et al. (2024a)), SinSR(Wang et al. (2024b)), SeeSR(Wu et al. (2024b)), SUPIR(Yu et al. (2024)), and DreamClear(Ai et al. (2024)). Following many existing works (Wang et al. (2024a); Wu et al. (2024b); Ai et al. (2024); Yu et al. (2024)), experiments are conducted on both synthetic and real-world benchmark datasets.

For the synthetic data, we randomly crop 2,124 patches from the validation sets of DIV2K Agustsson & Timofte (2017a) and LSDIR Li et al. (2023). For DIV2K, we use the five original degradation types: bicubic, unknown, mild, difficult, and wild. LSDIR-Val is generated by applying the same degradation pipeline used during training. For the real-world data, we adopt center-cropped images from RealSR Cai et al. (2019), DRealSR Wei et al. (2020) as used in Wu et al. (2024b) and RealLQ250 Ai et al. (2024). All evaluations are performed at a resolution of 1024×1024.

Metrics. We evaluate all methods using a set of no-reference image quality assessment metrics, including CLIP-IQA+ Wang et al. (2023), Q-Align Wu et al. (2023), MUSIQ Ke et al. (2021), MANIQA Yang et al. (2022), NIMA Talebi & Milanfar (2018), CLIP-IQA Wang et al. (2023), and NIQE Zhang et al. (2015), as well as reference-based metrics including PSNR, SSIM Wang et al.

Table 1: Quantitative comparison across different IQA metrics on RealSR Wu et al. (2024b), RealLQ250 Ai et al. (2024), DIV2K-Val, LSDIR-Val and DRealSR.

Benchmark		Metric	ResShift	StableSR	SinSR	SeeSR	DreamClear	SUPIR	LucidFlux(Ours)
Caption-Free			✓	✓	✓	✗	✗	✗	✓
Real-world	**DRealSR**	CLIP-IQA+ ↑	0.4655	0.3732	0.5402	0.6257	0.4461	0.5494	**0.6748**
		Q-Align ↑	2.6311	2.1245	3.1334	3.2745	2.4213	3.4720	**3.6919**
		MUSIQ ↑	40.9795	29.6691	53.9138	61.3222	35.1911	54.9279	**66.6833**
		MANIQA ↑	0.2687	0.2402	0.3455	0.4505	0.2675	0.3482	**0.4985**
		NIMA ↑	4.3178	3.9048	4.6226	4.6401	3.9368	4.5063	**4.9625**
		CLIP-IQA ↑	0.4964	0.3383	0.6631	0.6760	0.4360	0.5309	**0.6879**
		NIQE ↓	10.3005	8.6022	6.9800	6.4502	7.0163	5.9091	**4.7034**
	RealSR	CLIP-IQA+ ↑	0.5005	0.4408	0.5416	0.6731	0.5331	0.5640	**0.7074**
		Q-Align ↑	3.1045	2.5087	3.3615	3.6073	3.0044	3.4682	**3.7555**
		MUSIQ ↑	49.50	39.98	57.95	67.57	49.48	55.68	**70.20**
		MANIQA ↑	0.2976	0.2356	0.3753	0.5087	0.3092	0.3426	**0.5437**
		NIMA ↑	4.7026	4.3639	4.8282	4.8957	4.4948	4.6401	**5.1072**
		CLIP-IQA ↑	0.5283	0.3521	0.6601	**0.6993**	0.5390	0.4857	0.6783
		NIQE ↓	9.0674	6.8733	6.4682	5.4594	5.2873	5.2819	**4.2893**
	RealLQ250	CLIP-IQA+ ↑	0.5529	0.5804	0.6054	0.7034	0.6810	0.6532	**0.7406**
		Q-Align ↑	3.6318	3.5586	3.7451	4.1423	4.0640	4.1347	**4.3935**
		MUSIQ ↑	59.50	57.25	65.45	70.38	67.08	65.81	**73.01**
		MANIQA ↑	0.3397	0.2937	0.4230	0.4895	0.4400	0.3826	**0.5589**
		NIMA ↑	5.0624	5.0538	5.2397	5.3146	5.2200	5.0806	**5.4836**
		CLIP-IQA ↑	0.6129	0.5160	**0.7166**	0.7063	0.6950	0.5767	0.7122
		NIQE ↓	6.6326	4.6236	5.4425	4.4383	3.8700	**3.6591**	3.6742
Synthetic	**DIV2K-Val**	CLIP-IQA+ ↑	0.5583	0.5760	0.6128	0.7116	0.6585	0.6719	**0.7492**
		Q-Align ↑	3.5761	3.4226	3.7336	4.1167	3.9323	4.1659	**4.5311**
		MUSIQ ↑	60.5932	57.4246	66.0906	71.4947	65.8187	67.9074	**73.9045**
		MANIQA ↑	0.3421	0.2902	0.4341	0.5104	0.4369	0.4148	**0.5819**
		NIMA ↑	5.0430	5.0341	5.1810	5.2709	5.1663	5.1516	**5.4884**
		CLIP-IQA ↑	0.6017	0.5002	**0.7166**	0.7149	0.6663	0.5848	0.7034
		NIQE ↓	6.1976	4.9810	5.3679	4.2823	4.1634	3.7701	**3.7283**
		PSNR ↑	**18.3802**	18.3269	18.0956	18.2529	17.5701	17.7567	15.4393
		SSIM ↑	0.4394	**0.4819**	0.4259	0.4684	0.4291	0.4482	0.3837
		LPIPS ↓	0.3738	0.3933	0.3919	**0.3497**	0.3621	0.3785	0.4312
	LSDIR-Val	CLIP-IQA+ ↑	0.5248	0.5576	0.5582	0.7258	0.6995	0.7126	**0.7440**
		Q-Align ↑	3.5317	3.4878	3.7095	4.2997	4.2391	4.3468	**4.5959**
		MUSIQ ↑	57.6691	57.0838	63.9586	72.0142	70.7186	70.3340	**74.1923**
		MANIQA ↑	0.3408	0.2990	0.4131	0.5529	0.5059	0.4482	**0.5979**
		NIMA ↑	5.0916	5.0628	5.3353	5.4245	5.3773	5.3692	**5.6221**
		CLIP-IQA ↑	0.5691	0.4991	0.6766	**0.7314**	0.6941	0.6105	0.6836
		NIQE ↓	6.4447	4.2104	5.1771	3.9402	3.3318	**2.9610**	3.5571
		PSNR ↑	**17.3040**	17.1480	16.8241	17.0782	16.2114	16.1598	14.8688
		SSIM ↑	0.3935	0.4026	0.3710	**0.4113**	0.3823	0.3636	0.3697
		LPIPS ↓	0.4824	0.4655	0.4637	0.3969	**0.3720**	0.4408	0.4148

(2004), and LPIPS Zhang et al. (2018). Together, these metrics provide a comprehensive assessment of restoration performance across perceptual quality, semantic alignment, and structural fidelity.

Qualitative Comparisons. Figure 4 presents visual comparisons on representative samples from RealLQ250. SeeSR and DreamClear reduce some degradations but tend to leave residual artifacts or produce oversmoothed outputs with limited texture recovery. SUPIR generates cleaner results yet often loses fine details, leading to overly smooth surfaces. In contrast, our method achieves clearer edges, richer textures, and better semantic consistency with the degraded inputs, especially in challenging regions such as hair, text, and high-frequency patterns. These qualitative observations align with the quantitative results in Table 1, further confirming the effectiveness of our approach.

Quantitative Comparisons. Table 1 reports the IQA metric results on real-world and synthetic benchmarks. Our method consistently outperforms prior approaches on perceptual and semantic-oriented metrics, such as CLIP-IQA+, MUSIQ, MANIQA, Q-Align, and NIMA, highlighting its ability to generate visually faithful and semantically aligned restorations. On real-world datasets (e.g., DRealSR, RealSR, RealLQ250), LucidFlux achieves clear gains over existing caption- or tag-based methods. For distortion-focused measures like PSNR and SSIM on synthetic datasets, prior approaches report slightly higher values, yet these metrics are widely recognized as being less correlated with human perceptual quality. In contrast, our method delivers state-of-the-art performance on modern IQA benchmarks, supporting the view that advanced IR frameworks should be evaluated with perceptual and semantic quality measures rather than traditional distortion metrics.

Runtime and Model Scale Comparison. We compare LucidFlux with SeeSR, SUPIR, and DreamClear in terms of runtime and model size in Table 3. Despite using a substantially larger backbone (12B), our LucidFlux achieves a competitive total runtime by eliminating the caption preprocessing. In contrast, SeeSR, SUPIR, and DreamClear require additional preprocessing and rely on smaller backbones (1.29B, 3.5B, 0.6B), resulting in higher latency relative to their size. For trainable

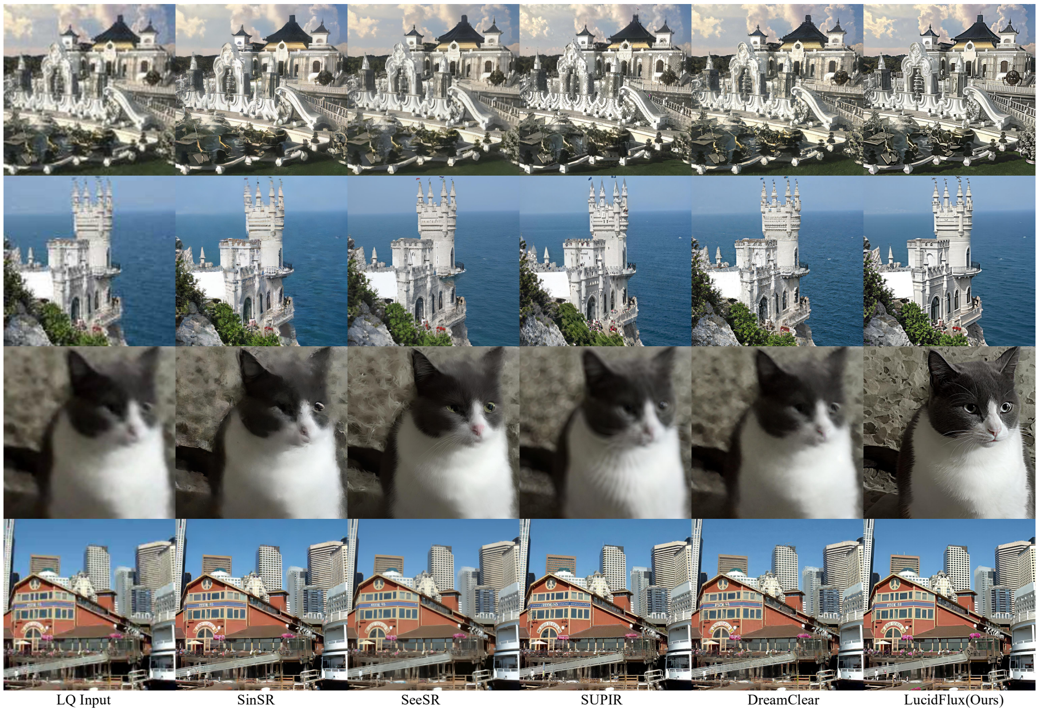


Figure 4: Qualitative comparisons on RealLQ250. Baseline methods either leave noticeable artifacts or yield over-smoothed textures, while our approach restores sharper details. See Figure 8 to Figure 13 in Appendix for more visual comparisons.

adapters, LucidFlux maintains a balanced design (1.6B), outperforming SUPIR (1.3B) in representational capacity while remaining more efficient than DreamClear (2.2B).

4.3 Ablation Study

We ablate our three contributions in testing RealLQ250 and report the quantitative results in Table 4. Starting from the Dual-Branch Conditioner (DBC) trained on LSDIR, our CLIP-IQA / CLIP-IQA+ / MUSIQ scores are 0.585/0.609/61.582, and three scores are enlarged after adding caption-free SigLIP semantic alignment. Our timestep- and layer-adaptive condition modulation (TLCM) further improves score performance, and scaling to our curated large-scale high-quality data provides the largest jump over TLCM on three metrics. The progression indicates that SigLIP alignment stabilizes semantics; TLCM exploits the DiT hierarchy; and data curation supplies structure-rich supervision, and thus all three modifications on DBC are required for the final outcome.

4.4 Comparison with Close-Source Commercial Methods

To assess the effectiveness of LucidFlux, we further compare it with several widely used commercial image restoration solutions, including HYPIR-FLUX Group (2025), Seedream 4.0 ByteDance Seed Vision Team (2025), Topaz Labs (2025), Gemini-NanoBanana DeepMind (2025), and MeiTu SR MeiTu (2025). All evaluations are conducted under the same experimental settings and the identical IQA metrics are used in the open-source comparisons. Table 2 reports the quantitative results of different methods. Our LucidFlux achieves the largest scores across all metrics and outperforms other commercial solutions. MeiTu SR shows the best performance among compared methods, but its restoration results generally have less details than our LucidFlux. In contrast, our method balances strong quantitative performance with reliable and consistent restoration, which makes it particularly suitable for real-world applications. See our Appendix Figure 7 for qualitative comparisons.

Table 2: Quantitative comparison across different IQA metrics with commercial models on RealLQ250.

Method	CLIP-IQA+ ↑	Q-Align ↑	MUSIQ ↑	MANIQA ↑	NIMA ↑	CLIP-IQA ↑	NIQE ↓
LQ Input	0.6218	2.1693	44.1541	0.3718	3.8664	0.6079	6.0790
Seedream 4.0	0.5002	3.6931	52.3771	0.2794	4.7024	0.4124	4.9393
Gemini-NanoBanana	0.3780	3.3114	44.6310	0.2548	4.6571	0.4434	6.0865
MeiTu SR	0.6653	4.1464	66.5936	0.4498	5.2103	0.6663	5.4125
LucidFlux (Ours)	**0.7406**	**4.3935**	**73.01**	**0.5589**	**5.4836**	**0.7122**	**3.6742**

Table 3: Runtime (s) and parameter scale (B).

	SeeSR	SUPIR	DreamClear	LucidFlux
Caption (s)	0.10	5.9	8.7	0
Inference (s)	22.38	16.6	28.9	23.6
Total (s)	22.48	22.5	37.6	23.6
Backbone (B)	1.29	2.6	0.6	12
Adapter (B, train.)	1.6	1.3	2.2	1.6
Total (B)	2.89	3.9	2.8	13.6

Table 4: Ablation study on RealLQ250. Evaluation metrics for three main contributions of our method.

Setting	CLIP-IQA	CLIP-IQA+	MUSIQ
Dual-Branch Conditioner Only	0.585	0.609	61.582
+ SigLIP Alignment	0.600	0.620	62.000
+ TLCM	0.622	0.635	65.500
+ Large HQ Data (Our method)	**0.7122**	**0.7406**	**73.0088**

5 CONCLUSION

LucidFlux demonstrates that caption-free universal image restoration is best achieved by *when, where, and what* to condition a large diffusion transformer, rather than by adding parameters or prompts. A lightweight dual-branch conditioner—grounded in the degraded input and a lightly restored proxy—and a timestep- and layer-adaptive modulation schedule recover high-frequency detail while preserving global structure and suppressing artifacts, all with a frozen Flux.1 backbone. SigLIP-based semantics provide training–inference consistency without captions. To make post-training practical, we introduce, to our knowledge, the first publicly documented and extensively validated UIR data-filtering pipeline. It is fully automatic once hyper-parameters are fixed and scales to 342K high-quality images and 1.36M paired samples, supplying structure-rich supervision at the capacity needed by large DiTs.Across real and synthetic benchmarks, LucidFlux delivers state-of-the-art perceptual quality and semantic fidelity with competitive runtime and minimal trainable overhead. We hope the pipeline, data recipe, and design insights provide a reliable foundation for restoration in the wild, and inspire future work on learned data selection, multi-frame/video extensions, and higher-resolution backbones—all while retaining caption-free inference.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 126–135, 2017a.

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 126–135, 2017b.

Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Xiaotian Han, Zhengyu Chen, Quanzeng You, and Hongxia Yang. Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation. *Advances in Neural Information Processing Systems*, 37:55443–55469, 2024.

bghira. Photo concept bucket, 2023. URL `https://huggingface.co/datasets/bghira/photo-concept-bucket`. Accessed: 2025-09-05.

ByteDance Seed Vision Team. Seedream 4.0. `https://www.doubao.com/chat/`, 2025. Accessed: 2025-09-24.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3086–3095, 2019.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

Google DeepMind. Gemini 2.5 flash image (nano banana), 2025. URL `https://aistudio.google.com/models/gemini-2-5-flash-image`. Accessed: 2025-09-24.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2): 295–307, 2016. doi: 10.1109/TPAMI.2015.2439281.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

XPixel Group. Hypir - ultra-hd ai image restoration tool, 2025. URL `https://www.hypir.org/`. Accessed: 2025-09-24.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Dehong Kong, Fan Li, Zhixin Wang, Jiaqi Xu, Renjing Pei, Wenbo Li, and WenQi Ren. Dual prompting image restoration with diffusion transformers, 2025. URL `https://arxiv.org/abs/2504.17825`.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Topaz Labs. Topaz enhance ai: Image upscaling and enhancement, 2025. URL `https://app.topazlabs.com/enhance/upscale`.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1775–1787, 2023.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021.

Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *arXiv preprint arXiv:2507.20590*, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

MeiTu. Meitu designkit – ai-powered high-fidelity image enhancement. `https://www.designkit.com/quality/`, 2025. Accessed: 2025-09-24.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36:24129–24142, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Yurui Qian, Qi Cai, Yingwei Pan, Yehao Li, Ting Yao, Qibin Sun, and Tao Mei. Boosting diffusion models with moving average sampling in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8911–8920, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pp. 4596–4604. PMLR, 2018.

Haoze Sun, Wenbo Li, Jiayue Liu, Kaiwen Zhou, Yongqiang Chen, Yong Guo, Yanwei Li, Renjing Pei, Long Peng, and Yujiu Yang. Text boosts generalization: A plug-and-play captioner for real-world image restoration, 2024. URL `https://openreview.net/forum?id=RjwWClPZtV`.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023.

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

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1905–1914, 2021.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25796–25805, 2024b.

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.

Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *European conference on computer vision*, pp. 101–117. Springer, 2020.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Project Lead by Wu, Haoning. Corresponding Authors: Zhai, Guangtai and Lin, Weisi.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *arXiv preprint arXiv:2406.08177*, 2024a.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25456–25467, 2024b.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1191–1200, 2022.

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, 2024.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36: 13294–13307, 2023.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022.

Lin Zhang, Lei Zhang, and Alan C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. doi: 10.1109/TIP.2015.2426416.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

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 conference on computer vision and pattern recognition*, pp. 586–595, 2018.

A Appendix

A.1 Likelihood of Degradation-Related Terms in Captions Generated by Different Multimodal Large Language Models

When using captions from multimodal large language models (MLLMs) as semantic guidance for restoration tasks, a potential risk is that these models may unintentionally introduce degradation-related terms (e.g., blur, noise, or low resolution). Such bias can mislead the restoration model by attributing degradations to input images even when they are not visually apparent. To quantify this effect, we evaluate the occurrence of degradation-related descriptions in captions generated by a set of representative MLLMs on RealLQ250, specifically LLaVA-v1.6-Vicuna-13B Liu et al. (2024) and Qwen2.5-VL-7B-Instruct Team (2025). Each caption is produced using the same prompt as DreamClear Ai et al. (2024), i.e., *"describe the key subjects and style"*, which is designed to neutrally guide the model toward content description without explicitly emphasizing or suppressing degradation cues. We then employ **Gemini-2.5-Flash-Image** as an external evaluator to analyze whether captions contain degradation-related mentions. Each caption is processed with a structured instruction to extract and categorize any degradation-related terms.

Prompt A.1 (Identifying Quality Degradations in Image Captions)

You are a professional image quality analysis expert. Carefully analyze the following image description text and identify any image quality issues that are either explicitly mentioned or implicitly implied.
Image description text: {`caption_content`}
Your task is to identify quality issues mentioned in the description. Focus on:

- Sharpness issues such as blur, unclear details, or defocus
- Noise, grain, or artifacts
- Low resolution, compression traces, or general quality problems
- Overexposure, underexposure, or color distortion
- Physical damage such as scratches, stains, or aging

Return results strictly in the following JSON format, without any additional explanation or text:

```
{
    "caption_content": "{caption_content}",
    "degradation_keywords": ["Extracted degradation-related terms"]
    ,
    "degradation_categories": {
        "Blur-related": ["blur", "unclear", "defocus"],
        "Noise-related": ["noise", "grain", "artifacts"],
        "Quality-related": ["resolution", "compression"],
        "Exposure-related": ["overexposure", "underexposure", "
    color issues"],
        "Damage-related": ["damage", "stains", "aging"]
    },
    "degradation_score": 0.0,
    "severity_level": "None/Minor/Moderate/Severe",
    "primary_issues": ["Main issue types"],
    "analysis_summary": "Brief analysis summary"
}
```

Scoring standard:

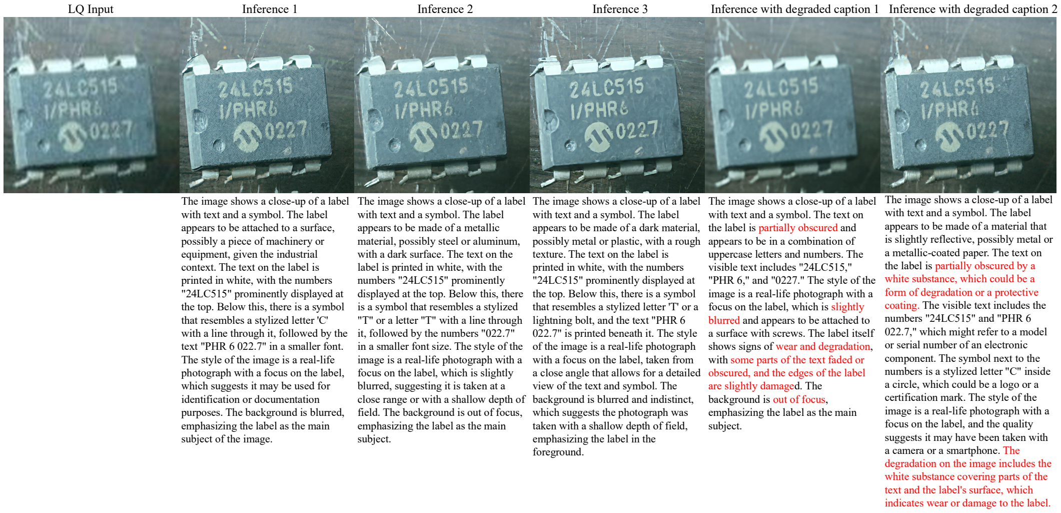
- `degradation_score`: 0 means no degradation, < 0.3 minor, 0.3–0.6 moderate, > 0.6 severe
- If no quality issues are mentioned in the text, set all arrays empty, score = 0, severity_level = "None"

Important: Only return pure JSON format results, without markdown code blocks or extra commentary.

If a caption contains terms that explicitly refer to image degradations, such as blur, noise, low resolution, or compression artifacts, we consider it a *degradation-related* caption. Table 5 presents the likelihood of captions containing such degradation-related terms. We observe that LLaVA-v1.6-vicuna-13b produces degradation-related captions in 17% of cases, whereas Qwen2.5-VL-7B-Instruct exhibits a higher occurrence of 24%. This indicates that different MLLMs vary in their tendency to

Table 5: Occurrence rates (%) of degradation-related terms in captions generated by different MLLMs on RealLQ250.

Model	LLaVA-v1.6-Vicuna-13B	Qwen2.5-VL-7B-Instruct
With Degradation (%)	17	24


LQ Input
Inference 1
Inference 2
Inference 3
Inference with degraded caption 1
Inference with degraded caption 2

The image shows a close-up of a label with text and a symbol. The label appears to be attached to a surface, possibly a piece of machinery or equipment, given the industrial context. The text on the label is printed in white, with the numbers "24LC515" prominently displayed at the top. Below this, there is a symbol that resembles a stylized letter 'C' with a line through it, followed by the text "PHR 6 022.7" in a smaller font. The style of the image is a real-life photograph with a focus on the label, which suggests it may be used for identification or documentation purposes. The background is blurred, emphasizing the label as the main subject of the image.

The image shows a close-up of a label with text and a symbol. The label appears to be made of a metallic material, possibly steel or aluminum, with a dark surface. The text on the label is printed in white, with the numbers "24LC515" prominently displayed at the top. Below this, there is a symbol that resembles a stylized "T" or a letter "T" with a line through it, followed by the numbers "022.7" in a smaller font size. The style of the image is a real-life photograph with a focus on the label, which is slightly blurred, suggesting it is taken at a close range or with a shallow depth of field. The background is out of focus, emphasizing the label as the main subject.

The image shows a close-up of a label with text and a symbol. The label appears to be made of a dark material, possibly metal or plastic, with a rough texture. The text on the label is printed in white, with the numbers "24LC515" prominently displayed at the top. Below this, there is a symbol that resembles a stylized letter 'T' or a lightning bolt, and the text "PHR 6 022.7" is printed beneath it. The style of the image is a real-life photograph with a focus on the label, taken from a close angle that allows for a detailed view of the text and symbol. The background is blurred and indistinct, which suggests the photograph was taken with a shallow depth of field, emphasizing the label in the foreground.

The image shows a close-up of a label with text and a symbol. The text on the label is partially obscured and appears to be in a combination of uppercase letters and numbers. The visible text includes "24LC515," "PHR 6," and "0227." The style of the image is a real-life photograph with a focus on the label, which is slightly blurred and appears to be attached to a surface with screws. The label itself shows signs of wear and degradation, with some parts of the text faded or obscured, and the edges of the label are slightly damaged. The background is out of focus, emphasizing the label as the main subject.

The image shows a close-up of a label with text and a symbol. The label appears to be made of a material that is slightly reflective, possibly metal or a metallic-coated paper. The text on the label is partially obscured by a white substance, which could be a form of degradation or a protective coating. The visible text includes the numbers "24LC515" and "PHR 6 022.7," which might refer to a model or serial number of an electronic component. The symbol next to the numbers is a stylized letter "C" inside a circle, which could be a logo or a certification mark. The style of the image is a real-life photograph with a focus on the label, and the quality suggests it may have been taken with a camera or a smartphone. The degradation on the image includes the white substance covering parts of the text and the label's surface, which indicates wear or damage to the label.

Figure 5: Impact of captions with and without degradation-related descriptions on restoration results. The second to fourth columns illustrate that inconsistent captions generated by the same MLLM across different runs lead to variations in the restoration outcomes. The fifth and sixth columns show that captions containing explicit degradation descriptions misguide the restoration model and result in inferior quality compared with captions focusing purely on content and style.

introduce degradation cues into captions, which may potentially bias downstream restoration tasks if these captions are directly used as supervision.

A.2 Impact of Degradation-Related Captions on Model Restoration

To further investigate the influence of degradation-related descriptions in MLLM-generated captions on restoration performance, we conducted experiments using two types of captions generated by LLaVA-v1.6-vicuna-13b. The first type uses the prompt *"Describe the key subjects and style"* to generate captions without emphasizing image degradations, while the second type uses the prompt *"Describe the key subjects and style, retain the descriptions of degradations on the image"* to produce captions that explicitly include degradation-related content.

As shown in Figure 5, two patterns emerge from the qualitative results. First, captions generated by the same MLLM using the neutral prompt exhibit variability across multiple runs, resulting in differences in the restoration outputs for the same input image. Second, when captions explicitly include degradation-related descriptions, the model's restoration performance is adversely affected, producing outputs of lower perceptual quality compared with captions that focus solely on key subjects and style. These findings indicate that both the consistency and content of MLLM-generated captions can significantly influence downstream restoration performance, underscoring the importance of controlling for degradation-related content when employing such captions as guidance.

These observations further highlight the practical limitations of relying on MLLM-generated captions during inference. The variability in captions leads to inconsistent restoration results, the presence of degradation-related descriptions can mislead the model and reduce output quality, and generating captions introduces additional computational overhead. Together, these factors underscore the advantages of a caption-free approach, which avoids reliance on potentially inconsistent or misleading textual guidance while reducing inference cost and maintaining robust restoration performance.

A.3 Additional Related Works

Large-Scale Image Restoration Datasets. The availability of large, high-quality datasets is critical for training generative restoration models. Existing datasets exhibit notable limitations: LSDIR Li et al. (2023) provides 85K images but depends on manual filtering, SUPIR Yu et al. (2024) collects 20M images without disclosing quality control procedures, and DreamClear Ai et al. (2024) generates 1M images via SDXL fine-tuning at a cost of 1280 V100 GPU days. To overcome these constraints, LucidFlux employs a fully automated three-stage filtering pipeline integrating blur detection, flat-region detection, and perceptual quality assessment. This approach produces diverse, structurally rich datasets that are reproducible, scalable, and suitable for training billion-parameter diffusion backbones efficiently.

Transformer-based T2I models (DiTs). Recent text-to-image systems increasingly adopt Transformer backbones—either diffusion transformers (DiTs) or rectified-flow transformers (RFTs)—which scale well and capture long-range dependencies in latent space Peebles & Xie (2022). Stable Diffusion 3 (SD3) introduces a *Multimodal Diffusion Transformer (MMDiT)* with *separate* weights for image and text tokens and bidirectional information flow; it is trained with *rectified flow* and improved noise sampling biased toward perceptually relevant scales, yielding stronger text comprehension and typography Esser et al. (2024). PixArt-α proposes an efficient DiT recipe—three-stage training (pixel dependency, text–image alignment, aesthetics), injecting cross-attention into DiT, and dense pseudo-captioning—achieving 1024px photorealistic quality at a fraction of typical compute Chen et al. (2023). FLUX Labs (2024) scales a *rectified-flow Transformer* (rather than a diffusion transformer), with open-weight variants (e.g., `dev/schnell`) built around cross-attention over text embeddings. Building on this line, **LucidFlux** leverages a large MM-DiT backbone (Flux.1) and specializes conditioning for caption-free restoration, improving detail fidelity while preserving semantics.

A.4 Extended Dataset Analysis

To further examine semantic diversity, we visualize the CLIP image–text embeddings using t-SNE. We randomly sample 10K images from our filtered data, while using all available images from Flickr2K and DIV2K. As shown in Figure 6, our dataset spans a substantially broader semantic range, reflecting richer and more diverse image–text concepts. This confirms the advantage of our dataset in supporting models that rely on wide semantic generalization.

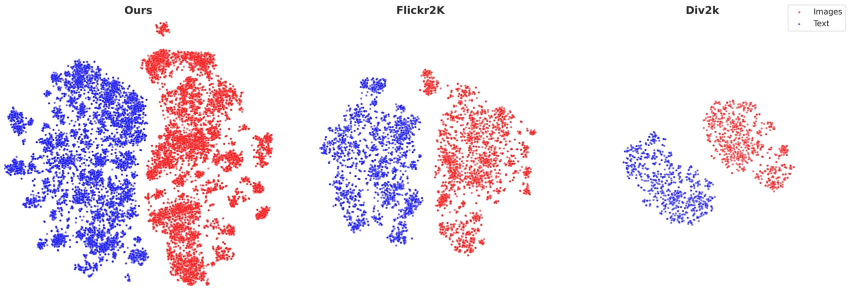


Figure 6: t-SNE visualization of CLIP image–text embeddings. Our dataset covers a broader semantic range than Flickr2K and DIV2K, indicating richer image–text diversity.

A.5 Extended Visual Comparisons

To provide a more comprehensive evaluation, we present extended qualitative results across all benchmark datasets. Figures 8–13 include representative examples from RealLQ250, DRealSR, RealSR, DIV2K-Val, and LSDIR-Val. These comparisons consistently demonstrate that our method produces sharper edges, more faithful textures, and better preservation of semantic structures compared with existing open-source state-of-the-art approaches. The additional results further corroborate the advantages of our approach observed in the main paper.

A.6 Visual Comparison with Close-Source Commercial Methods

Figure 7 illustrates representative visual results on RealLQ250. HYPIR-FLUX and Seedream 4.0 fail to fully remove degradations, leaving noticeable residual artifacts. Topaz suppresses degradations more effectively but generates flat and over-smoothed textures. Gemini-NanoBanana provides visually plausible outputs but often struggles to recover high-frequency details. MeiTu SR shows relatively strong restoration ability, producing sharper and more natural results compared with most commercial counterparts. Among the evaluated models, LucidFlux consistently delivers the sharpest structures and most faithful details, particularly in fine-grained regions, while maintaining high structural fidelity and reliability.

A.7 Inference Details

For all experiments, we use the FlowMatch Euler sampling introduced for SD3's rectified-flow formulation and implemented in Diffusers' FlowMatchEulerDiscreteScheduler Esser et al. (2024) to sample FLUX Labs (2024) and inherit the adaptive shift adjustments from the official Flux implementation, ensuring stable sampling dynamics and consistent step-wise updates. All inference is performed with 28 sampling steps in FP16 precision and utilizes the wavelet color alignment method from Ai et al. (2024) and the full default instruction is restore this image into high-quality, clean, high-resolution result.

A.8 Limitations

While LucidFlux attains strong perceptual quality and semantic fidelity, several practical limitations remain:

Large model scale. LucidFlux is built on a high-capacity DiT backbone (Flux.1). This provides rich generative priors but entails substantial parameter count and compute cost. Training generally requires multi-GPU setups; for inference, high-end GPUs are preferable to maintain reasonable throughput.

Inference GPU memory. VRAM usage during inference is sizable and grows with input resolution and batch size. Transformer-based diffusion exhibits quadratic attention complexity with respect to token count, so higher resolutions can quickly amplify memory pressure. This constrains deployment on memory-limited devices unless tiling or resolution reductions are used.

Sampling steps vs. quality. High-quality outputs typically require more than ˜15 denoising steps. Fewer steps may lead to over-smoothing and loss of fine textures. This introduces a latency–quality trade-off that can be restrictive for real-time or interactive applications.

Mitigations. Promising directions include model compression (distillation/pruning), low-precision inference, memory-efficient attention, and step reduction via progressive distillation. These optimizations are orthogonal to our method and could reduce compute and memory while preserving quality.

A.9 Acknowledgments

We thank the maintainers and communities behind Flux.1 Labs (2024), PixArt-α Chen et al. (2023), SD/SDXL Rombach et al. (2022); Podell et al. (2023), SigLIP, and the dataset providers for LSDIR, DIV2K, Flickr2K, RealSR, and DRealSR for making their resources available to the research community.

Intended application scope. LucidFlux is designed for benign restoration of non-sensitive, natural photographic images with mixed, unknown degradations (e.g., sensor noise, motion blur, compression). Appropriate uses include consumer photo enhancement, archival preservation, academic research, and benchmarking under unknown degradations. The model aims to improve perceptual quality while preserving semantics, but as a generative diffusion system it may synthesize plausible details not present in the input.

Out-of-scope scenarios. The method is *not* intended for domains that require pixel-accurate fidelity or expert supervision (e.g., medical imaging, scientific microscopy, satellite/remote sensing, or legally binding forensic evidence). It is also not tailored for document restoration or OCR-critical text recovery.

Prohibited or discouraged uses. LucidFlux should not be used to circumvent privacy, safety, or consent—such as deblurring or enhancing faces, license plates, or personally identifiable content for surveillance or re-identification; removing watermarks or intentional obfuscation; fabricating or altering imagery for deception; or generating identity-sensitive content (e.g., deepfakes). When restoration might affect downstream decisions about people, human oversight is required.

Operational caveats. Because performance depends on degradation type and sampling steps, outputs should be reviewed before downstream use, especially in safety-critical or regulatory contexts. If exact visual truth is required, classical reconstruction baselines or domain-specific methods with uncertainty quantification are preferable.

A.10 LLM Usage

We used large language models solely for editorial assistance—to refine grammar and phrasing, improve clarity and flow, and condense overly verbose passages. No ideas, methods, code, figures, citations, or results were generated by an LLM, and no unverifiable content was introduced. All technical content, study design, experiments, analyses, and conclusions were conceived, executed, and validated by the authors, who take full responsibility for the manuscript.

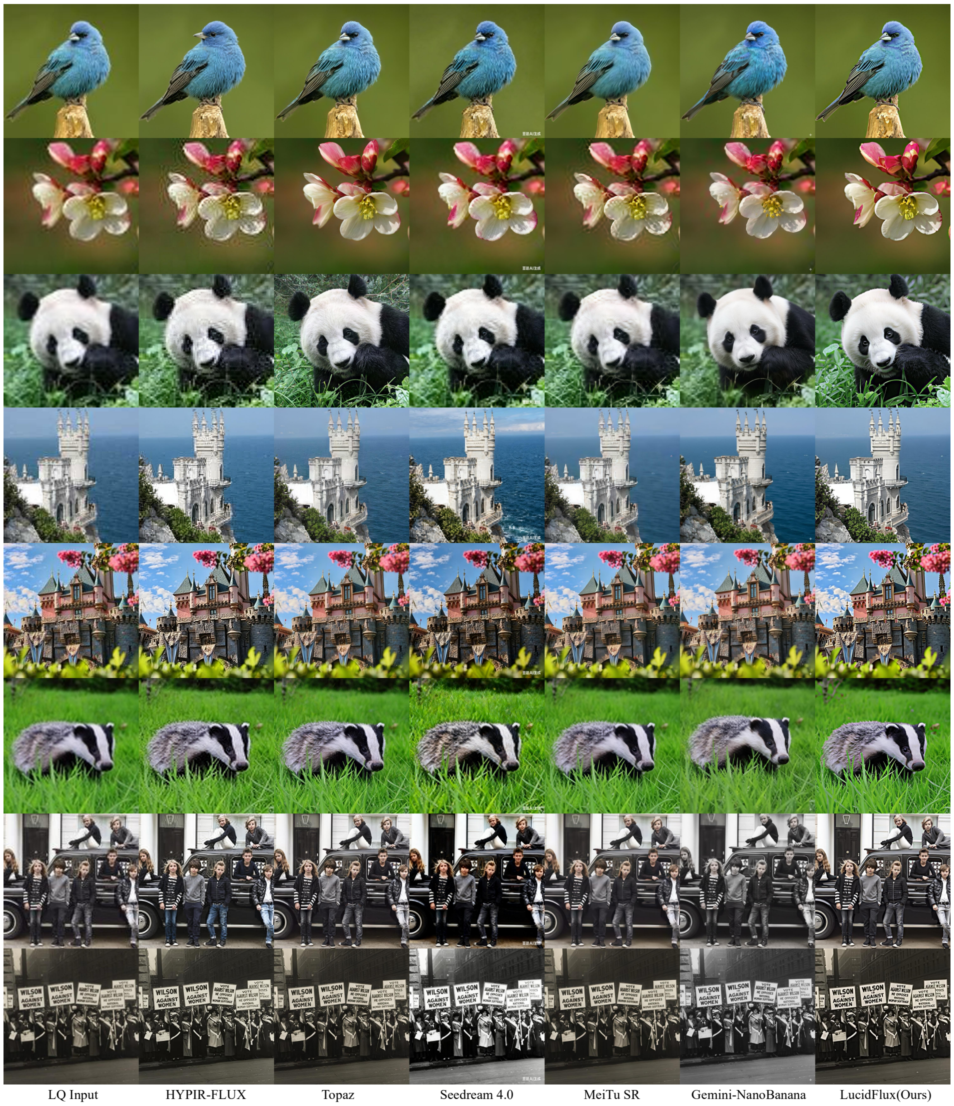


Figure 7: Qualitative comparison with commercial models on RealLQ250.

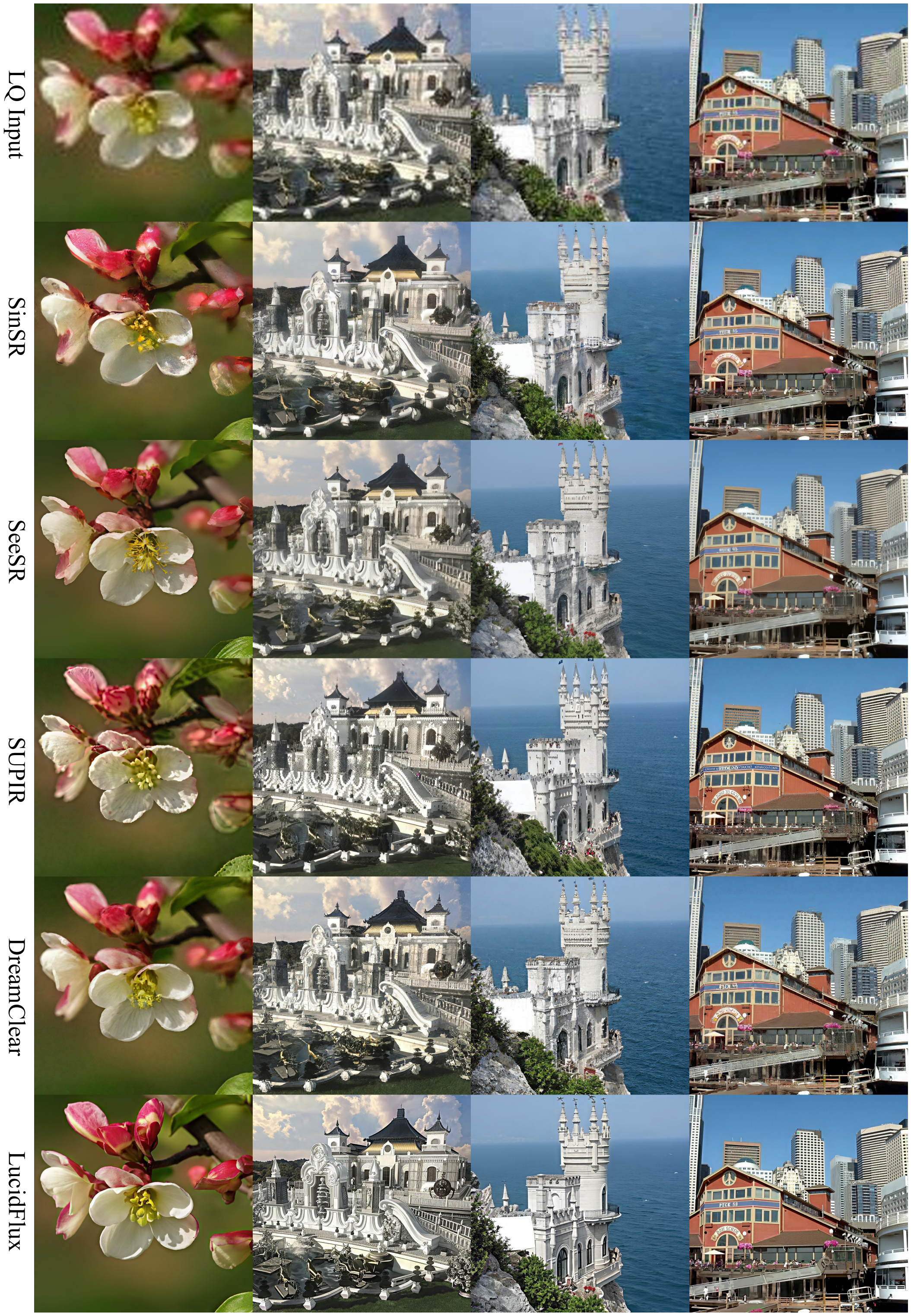


Figure 8: More examples of visual comparison with open-source state-of-the-art methods on RealLQ250.

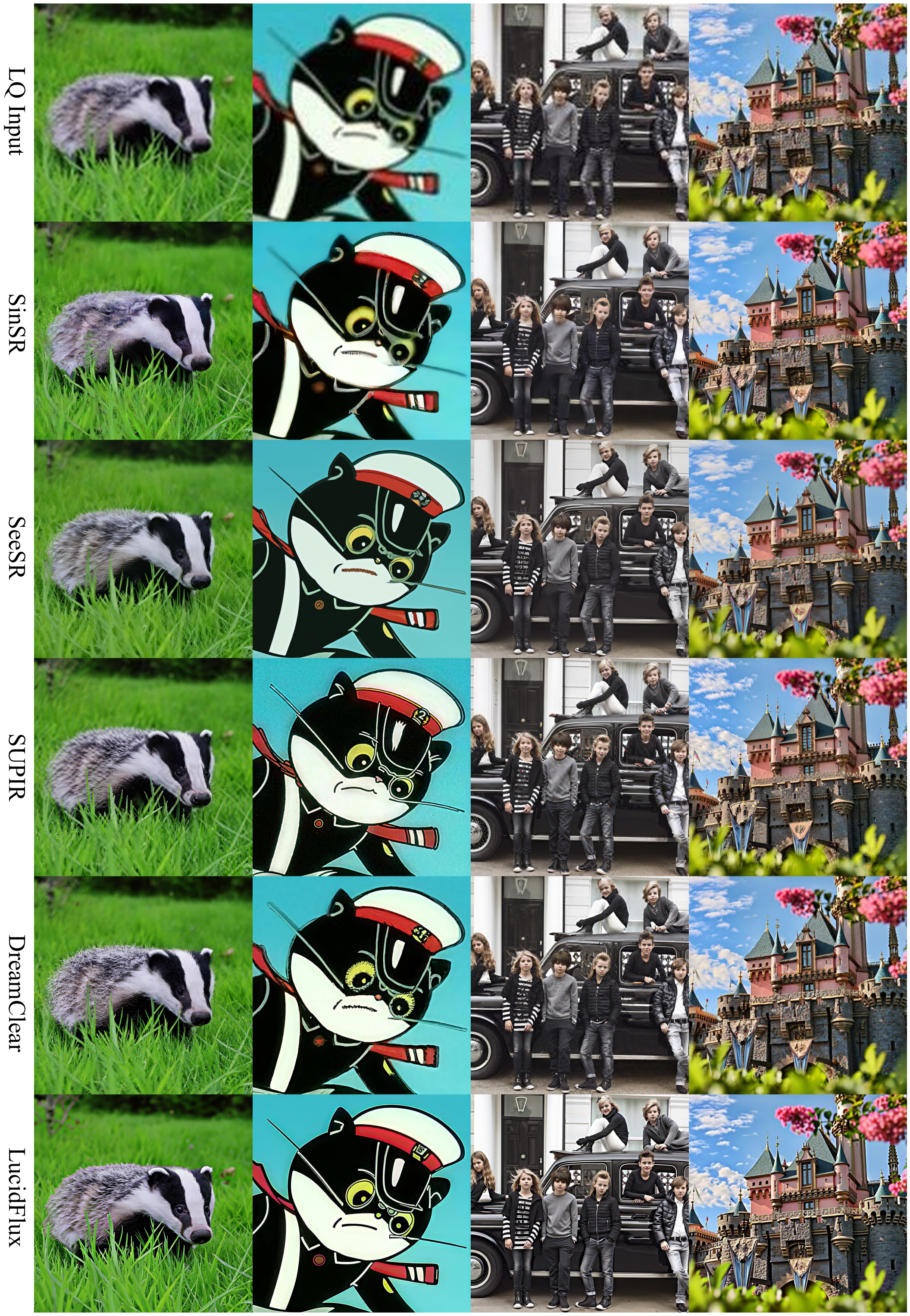


Figure 9: More examples of visual comparison with open-source state-of-the-art methods on RealLQ250.

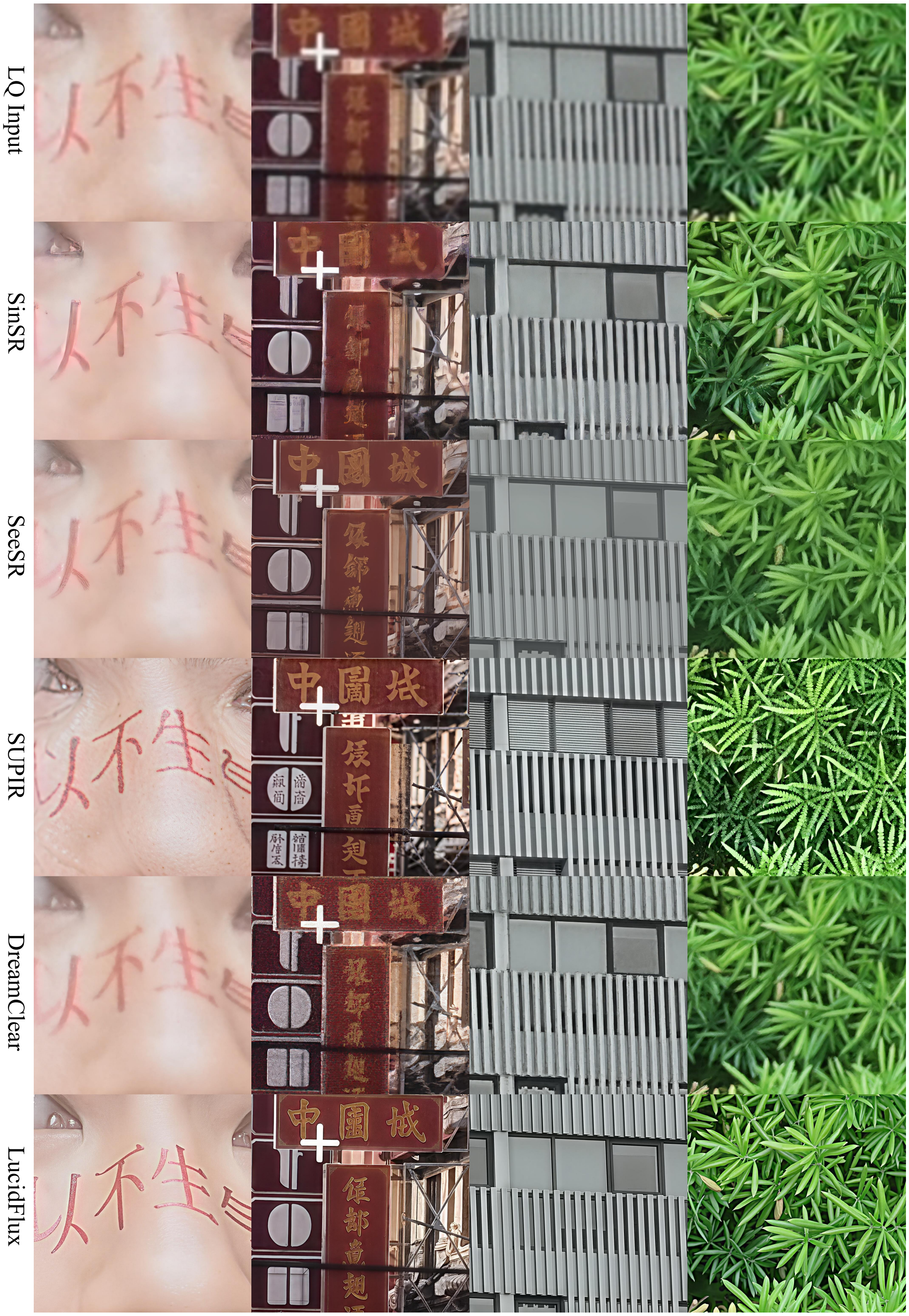


Figure 10: More examples of visual comparison with open-source state-of-the-art methods on DRealSR.

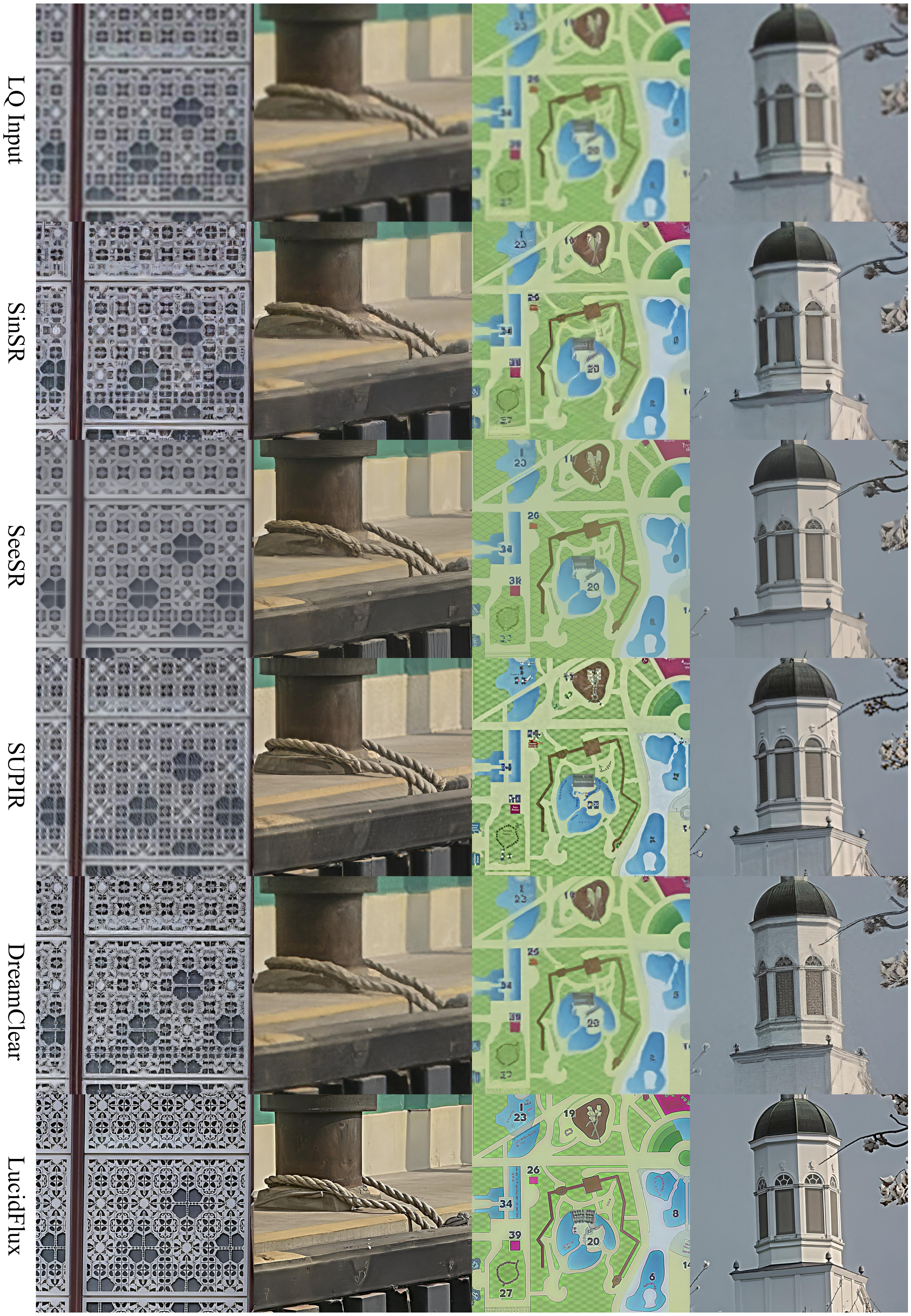


Figure 11: More examples of visual comparison with open-source state-of-the-art methods on RealSR.

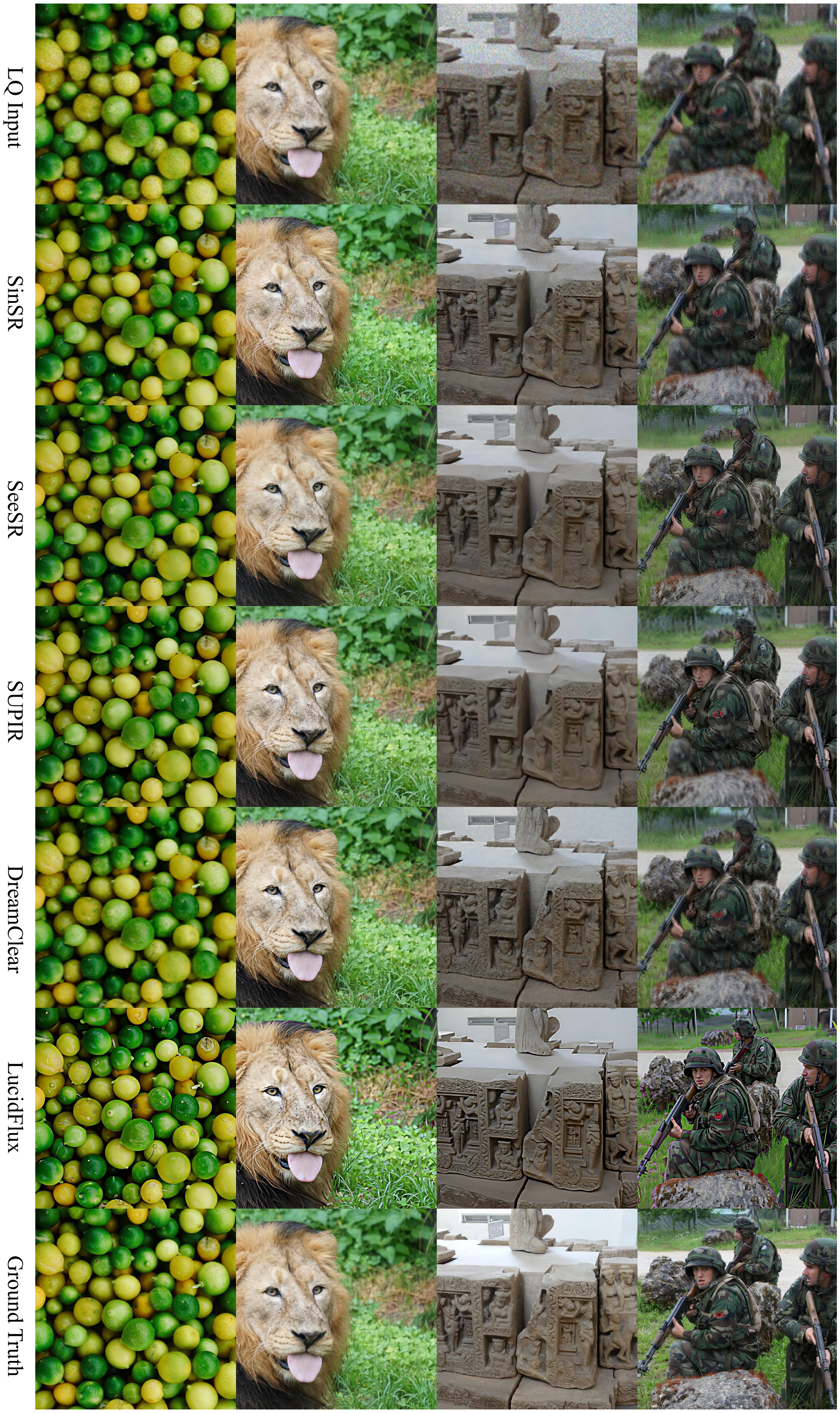


Figure 12: More examples of visual comparison with open-source state-of-the-art methods on Div2k-Val.

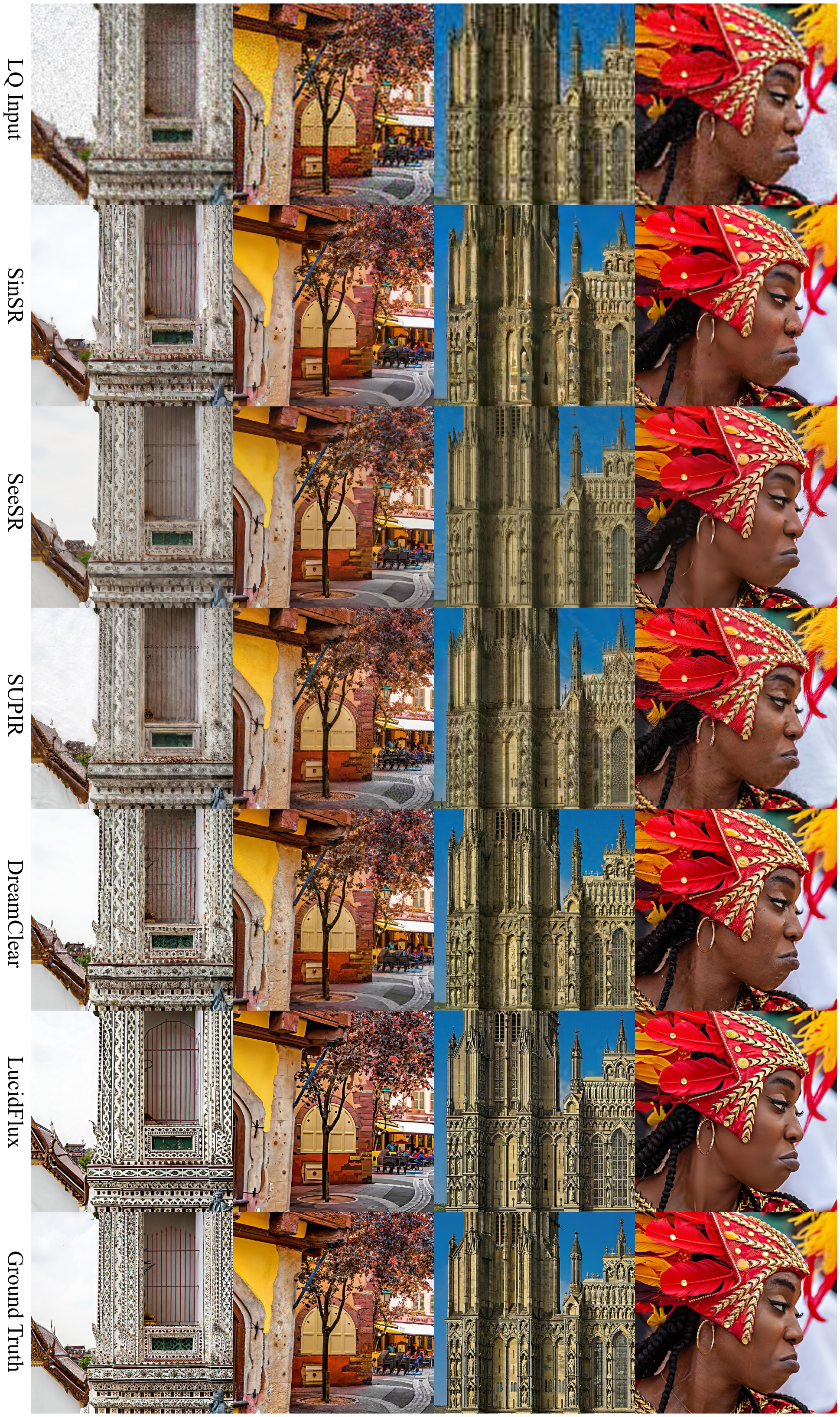


Figure 13: More examples of visual comparison with open-source state-of-the-art methods on LSDIR-Val.