RESTOREVAR: VISUAL AUTOREGRESSIVE GENERATION FOR ALL-IN-ONE IMAGE RESTORATION

Anonymous authors

000

001

002003004

006

008 009

010 011

012

013

014

015

016

017

018

019

021

025

026

027

028

031

033

034

035

037

040

041

042

043

044

046

047

048

051

052

Paper under double-blind review

ABSTRACT

The use of latent diffusion models (LDMs) such as Stable Diffusion has significantly improved the perceptual quality of All-in-One image Restoration (AiOR) methods, while also enhancing their generalization capabilities. However, these LDM-based frameworks suffer from slow inference due to their iterative denoising process, rendering them impractical for time-sensitive applications. Visual autoregressive modeling (VAR), a recently introduced approach for image generation, performs scale-space autoregression and achieves comparable performance to that of state-of-the-art diffusion transformers with drastically reduced computational costs. Moreover, our analysis reveals that coarse scales in VAR primarily capture degradations while finer scales encode scene detail, simplifying the restoration process. Motivated by this, we propose RestoreVAR, a novel VAR-based generative approach for AiOR that significantly outperforms LDM-based models in restoration performance while achieving over $10 \times$ faster inference. To optimally exploit the advantages of VAR for AiOR, we propose architectural modifications and improvements, including intricately designed cross-attention mechanisms and a latent-space refinement module, tailored for the AiOR task. Extensive experiments show that RestoreVAR achieves state-of-the-art performance among generative AiOR methods, while also exhibiting strong generalization capabilities. The code will be made publicly available after the review process.

1 Introduction

Image restoration is a complex inverse problem that aims to recover clean images from degradations, such as haze, rain, snow, blur, and low-light conditions. Recently, the paradigm of All-in-One image Restoration (AiOR) has emerged, where a single network is trained to handle multiple degradation types. Existing AiOR methods can be broadly categorized into non-generative and generative approaches. Non-generative models such as AirNet (Li et al., 2022), PromptIR (Potlapalli et al., 2024), InstructIR (Conde et al., 2025), AWRaCLe (Rajagopalan & Patel, 2024), and AdaIR (Cui et al., 2024), deterministically map degraded images to their clean counterparts. While these methods offer fast inference and reliable pixel-level restoration performance, they often fail to generalize to diverse degradations encountered in real-world scenarios. To overcome this challenge, recent works have adopted generative models that aim to capture the distribution of clean images and produce more perceptually realistic outputs. Early works (Chen et al., 2022; Kupyn et al., 2018) based on GANs (Goodfellow et al., 2020) attempted this through adversarial learning, but suffered from mode collapse and unstable training. To improve fidelity and training stability, DiffUIR (Zheng et al., 2024) and DA-CLIP (Luo et al., 2023) employed pixel-space diffusion models (Ho et al., 2020). However, their high computational cost makes large-scale pretraining infeasible, limiting their ability to learn strong generative priors. In contrast, recent methods such as Diff-Plugin(Liu et al., 2024), AutoDIR(Jiang et al., 2023), and PixWizard (Lin et al., 2024) leverage latent diffusion models (LDMs), such as Stable Diffusion (Rombach et al., 2022). By operating in a latent space, LDMs significantly reduce computational costs, enabling large-scale pretraining which equips them with strong generative priors of natural images. These priors allow LDM-based AiOR methods to deliver perceptually realistic restoration and improved generalization to real-world degradations.

Despite their advantages, LDM-based AiOR methods have some shortcomings. (1) LDMs require multiple denoising steps during inference, resulting in significantly longer runtimes compared to non-generative models. Their slow inference speeds pose challenges for applications that demand

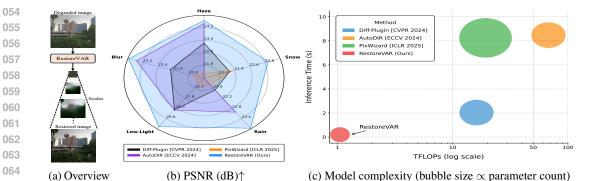


Figure 1: RestoreVAR, our proposed VAR-based (Tian et al., 2024) scale-space generative AiOR model (a), significantly outperforms LDM-based methods as shown in (b). RestoreVAR also offers drastic reductions in computational complexity as shown in (c).

real-time processing, such as video surveillance or autonomous navigation. (2) LDMs rely on variational autoencoders (VAEs) (Kingma, 2013) which are primarily trained for generative diversity, rather than accurate pixel-level reconstruction. Consequently, the restored images obtained from LDM-based AiOR methods exhibit loss of fine structural details, hindering their performance.

Autoregressive models have driven rapid advances in natural language processing through large language models (LLMs) such as GPT-3 (Radford et al., 2019b) and LLaMA (Touvron et al., 2023). These models generate outputs by predicting the next token, conditioned on previously generated tokens. Recently, Visual AutoRegressive (VAR) Modeling (Tian et al., 2024) introduced scale-space autoregression for image generation, performing next-scale prediction in the latent space of a multiscale vector-quantized VAE (VQVAE). VAR achieves performance comparable to state-of-the-art diffusion models such as DiT-XL/2 (Peebles & Xie, 2023), while operating 45× faster. Despite its success in generative tasks, the application of VAR to low-level vision tasks such as image restoration remains largely unexplored. To the best of our knowledge, only two prior works—VarSR (Qu et al., 2025) and Varformer (Wang & Zhao, 2024)—have used VAR for image restoration. VarSR focused exclusively on the super-resolution task, while Varformer utilized intermediate VAR features to guide a separate non-generative network for AiOR. In contrast, our approach is generative and fully exploits the strong priors of the pretrained VAR model by training it directly for the AiOR task. Our analysis in Sec. 3.2 also reveals that the scale-space decomposition of VAR captures degradations predominantly in coarse scales and scene-level details in fine scales, making it well-suited for AiOR.

To this end, we introduce RestoreVAR, a novel generative approach for AiOR that addresses some of the key limitations of LDM-based approaches. Firstly, RestoreVAR adopts the autoregressive structure of VAR, achieving state-of-the-art generative AiOR performance with over $10\times$ faster inference than LDM-based methods (see Fig. 1). Secondly, RestoreVAR employs cross-attention mechanisms conditioned on the degraded image latents, enabling the model to maintain spatial consistency and minimize hallucinations. Thirdly, to mitigate the loss of fine details by the vector quantization and VAE decoding processes, we propose a lightweight (only $\sim 3\%$ overhead) nongenerative latent refinement transformer which predicts de-quantized latents from the outputs of VAR. Additionally, we fine-tune the VAE decoder to operate on these continuous latents, further enhancing reconstruction quality. Finally, through extensive experiments, we demonstrate that RestoreVAR achieves state-of-the-art performance among generative restoration models, while also exhibiting strong generalization to real-world degradations. To summarize, our key contributions are:

- 1. We propose RestoreVAR, the first VAR-based generative AiOR framework that achieves superior performance and a 10× faster inference than LDM-based methods.
- 2. To achieve semantically coherent restoration, we introduce degraded image conditioning through cross-attention at each block of the VAR transformer.
- 3. To mitigate the loss of fine details in the vector quantization and VAE decoding processes, we introduce a non-generative latent refiner transformer which converts discretized latents into continuous ones, and fine-tune the VAE decoder to operate on continuous latents.
- 4. Extensive experiments show that RestoreVAR attains state-of-the-art performance among generative AiOR approaches, with perceptually preferable results and strong generalization.

2 RELATED WORKS

108

109110

111 112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129 130

131 132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148 149

150 151

152

153 154

155 156

157

158

159 160

161

2.1 IMAGE RESTORATION

Early restoration models primarily addressed specific degradations (He et al., 2009; Zhang et al., 2020; Wang et al., 2019; Yasarla & Patel, 2019; Zhang et al., 2021a; Nah et al., 2017). Other methods such as Restormer (Zamir et al., 2022), MPRNet (Zamir et al., 2021) and SwinIR (Liang et al., 2021) introduced architectures for any single restoration task. However, they are restricted to handle one degradation at a time, making them ineffective for multiple degradations. All-in-One image Restoration (AiOR) methods aim to tackle multiple corruptions with a single model. Early approaches include non-generative models such as All-in-one (Li et al., 2020) and Transweather (Valanarasu et al., 2022). PromptIR (Potlapalli et al., 2024) used learnable prompts while AWRaCLe (Rajagopalan & Patel, 2024) utilized visual in-context learning to extract degradation characteristics. Other approaches such as InstructIR (Conde et al., 2025) adopted textual guidance, and DCPT (Hu et al., 2025) proposed a novel pre-training strategy for AiOR. DFPIR (Tian et al., 2025) proposed a feature perturbation strategy for AiOR. Recent AiOR methods have adopted diffusion models. Pixel-space diffusion models (PSDMs) such as DA-CLIP (Luo et al., 2023) and DiffUIR (Zheng et al., 2024) demonstrated improved AiOR performance but lacked robust generative priors. Recent methods have utilized the strong priors of LDMs for AiOR. Diff-Plugin (Liu et al., 2024) adopts task plugins to guide an LDM for AiOR. AutoDIR (Jiang et al., 2023) automatically detects and restores degradations using an LDM. PixWizard (Lin et al., 2024) is a multi-task SD-XL (Podell et al., 2023) based model capable of performing AiOR among other tasks. However, LDM-based approaches are slow at inference time—a limitation we aim to overcome using visual autoregressive modeling (VAR).

2.2 Autoregressive models in vision

Recent works (Van Den Oord et al., 2016; Tian et al., 2024) have extended autoregressive (AR) models to vision and can be categorized as pixel-space AR (Van Den Oord et al., 2016; Van den Oord et al., 2016; Chen et al., 2018), token-based AR (Van Den Oord et al., 2017; Yu et al., 2023; Ramesh et al., 2021) and scale-space AR (Tian et al., 2024; Ren et al., 2024; Guo et al., 2025). Pixel-space AR predicts raw pixels one by one in raster order, as in PixelRNN (Van Den Oord et al., 2016) and PixelCNN++ (Salimans et al., 2017), but is very slow at high resolutions. Token-based AR compresses images into discrete latent codes via vector quantization (e.g., VQ-VAE (Van Den Oord et al., 2017), VQGAN (Esser et al., 2021)) and then models code sequences with transformers (e.g. ImageGPT (Chen et al., 2020)). This trades-off codebook size and transformer capacity against tractability for high-resolution generation. Scale-space AR, as introduced in VAR (Tian et al., 2024), generates latents from coarse to fine scales and matches the quality of Diffusion Transformers (Peebles & Xie, 2023) at a fraction of the inference cost. HART (Tang et al., 2024) scales VAR to higher resolution and uses a MLP-based diffusion refiner to convert discrete VAR latents into continuous representations. Despite VAR's success in generative tasks, it remains underexplored for image restoration with only two prior works-VarSR (Qu et al., 2025) and Varformer (Wang & Zhao, 2024). VarSR addressed super-resolution, while Varformer used VAR's features to guide a non-generative AiOR model. In contrast, RestoreVAR is a generative model which directly trains VAR for AiOR.

3 Proposed Method

We first explain the working principles behind VAR for image generation. We then describe our scale-space analysis of VAR and detail RestoreVAR, our proposed VAR-based approach for AiOR.

3.1 Preliminaries: Visual Autoregressive Modelling

Visual Autoregressive Modelling, or VAR, is a novel autoregressive class-conditioned image generation method which uses a GPT-2 (Radford et al., 2019a) style decoder-only transformer architecture for next-scale prediction. The VAR transformer operates in the latent space of a multi-scale VQ-VAE which uses K scales. Given an image $I \in \mathbb{R}^{H \times W \times 3}$, the VQVAE encoder outputs a latent representation $f_{\text{cont}} \in \mathbb{R}^{H_K \times W_K \times C}$. Hereafter, we will refer to f_{cont} as the *continuous latent*, and the latent obtained after quantization as *discrete latent*. Instead of directly quantizing f_{cont} , a multi-scale residual quantization using a shared codebook across K spatial scales is performed. First, the residual

166 167

168

169

170

171 172 173

174

175

176

177 178

179

181

182

183

185

186

187

188

189 190

191

192

193

196

197

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

and accumulated quantized (or discrete) reconstruction of f_{cont} are initialized as $f_{\text{res}}^{(0)} := f_{\text{cont}}$ and $f_{\text{quant}}^{(0)} := 0$, respectively. At each scale $k = 1, \dots, K$, an index map $r_k \in \mathbb{Z}^{H_k \times W_k}$ is obtained by quantizing the downsampled residual feature:

 $r_k := \text{quantize}\left(\text{downsample}\left(f_{\text{res}}^{(k-1)}\right)\right).$ The indices r_k are then decoded using the codebook embeddings $e(\cdot)$, upsampled to match the full resolution, and refined using a convolutional module $\phi_k(\cdot)$ to obtain

 $h_k := \phi_k \text{ (upsample } (e(r_k))), \in \mathbb{R}^{H_K \times W_K \times C}.$

This is done to approximate the information captured at the current scale which is used to update the

residual continuous features to be modelled by subsequent scales as $f_{\rm quant}^{(k)} := f_{\rm quant}^{(k-1)} + h_k, \quad f_{\rm res}^{(k)} := f_{\rm cont} - f_{\rm quant}^{(k)}.$ This process is repeated for all scales and yields a set of index maps $\{r_1, r_2, \ldots, r_K\}$, each consisting of the code-book indices of residual information at an increasingly finer scale.

For training, VAR uses teacher-forcing, where the ground-truth index maps $\{r_1, r_2, \dots, r_K\}$ are used to autoregressively predict the next scale. For each scale k, the accumulated reconstruction $f_{ ext{quant}}^{(k-1)} = \sum_{i=1}^{k-1} \phi_i \left(ext{upsample}(e(r_i)) \right)$ is interpolated to the resolution of scale k to obtain $\hat{f}_{ ext{quant}}^{(k)}$, which is then flattened into tokens, and concatenated with the remaining tokens to form the input sequence. A start-of-sequence (SOS) token, derived from the class label embedding, is then prepended to this input sequence. A block-wise causal attention mask is used to ensure that predictions for scale k attend only to the previous scales. VAR is trained to minimize the cross-entropy loss between predicted logits and the ground-truth index maps, modeling the likelihood

$$p(r_1, r_2, \dots, r_K) = \prod_{k=1}^{K} p(r_k \mid r_1, r_2, \dots, r_{k-1})$$

 $p(r_1,r_2,\ldots,r_K) = \prod_{k=1}^K p(r_k \mid r_1,r_2,\ldots,r_{k-1}).$ During inference, the SOS token is created from the target class label. VAR then autoregressively predicts each index map r_k , one scale at a time. After predicting r_k , its embedding is upsampled, refined and accumulated to form the input for the next scale, mimicking the same procedure used during training. The VAR model uses only K = 10 latent scales with key-value (KV) caching, enabling significantly faster inference compared to latent diffusion models.

SCALE-SPACE ANALYSIS OF VAR 3.2

In addition to VAR's competitive performance to LDMs with far superior inference speed, we found that its residual scalespace decomposition focuses on degradations and scene-level details across different scales. To demonstrate this, we consider clean (GT)-degraded image pairs and compute their scale-wise residual indices $\{r_k^{\rm GT}\}_{k=1}^K$ and $\{r_k^{\text{Deg}}\}_{k=1}^K$, respectively, where K=10. We define coarse scales as k = 1, ..., 5(low-resolution index maps) and fine scales as $k = 6, \dots, 10$ (higher-resolution index maps). The first and last columns of Fig. 2 show reconstructions from r_k^{Deg} and r_k^{GT} , respectively. In column 2, we replace the coarse scales of $r_k^{\rm GT}$ with those from $r_k^{\rm Deg}$. This introduces the degradation, although fine scales remain unchanged. Next, in column 3 we replace the fine scales of $r_k^{\rm GT}$ with those from r_k^{Deg} , which yields a clean image with some loss of fine details. These observations indicate that coarse scales in VAR capture degradations, while finer scales encode scene-level

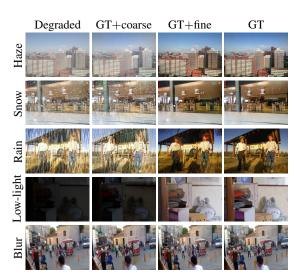


Figure 2: VAR captures degradations in early scales (coarse) and scene-level details in later scales (fine). Degraded and GT are VQVAE reconstructions of the degraded and ground truth images. GT+coarse replaces early GT scales with degraded ones, while GT+fine replaces the late GT scales.

detail. Notably, this observation holds across multiple degradations and simplifies restoration for VAR as removing degradations requires correctly predicting only the early scales which contain a small number of tokens, while scene details can be reconstructed in subsequent scales.

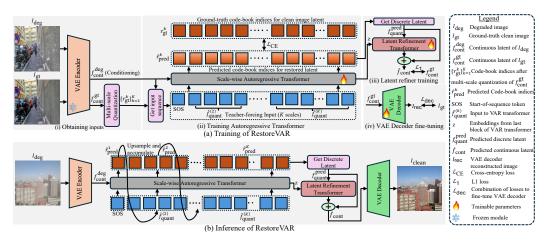


Figure 3: Illustration of RestoreVAR for training and inference. (a) Shows the training procedure for each component of RestoreVAR, and (b) shows the overall pipeline during inference.

3.3 RESTOREVAR

We now describe RestoreVAR, our proposed approach that effectively adapts VAR for AiOR, leveraging its substantial inference speed advantage over LDMs. Given a degraded image $I_{\text{deg}} \in \mathbb{R}^{H \times W \times 3}$, the goal is to predict a clean output I_{clean} , close to the ground-truth I_{gt} . Adapting VAR to AiOR is non-trivial due to the need for high-quality pixel-level reconstruction, which is compromised by two factors: (1) VAR's strong generative priors can cause hallucinations in the restored images without proper conditioning. (2) Vector quantization and VAE decoding introduce artifacts that hinder pixel-level restoration. RestoreVAR addresses these challenges through architectural enhancements, including cross-attention to incorporate semantic guidance from the degraded image, and a novel non-generative transformer that refines discrete latents into their continuous form to preserve fine details in the restored image. We describe these components below.

3.3.1 AUTOREGRESSIVE TRANSFORMER ARCHITECTURE

For training, the multi-scale teacher-forcing input is constructed from the ground-truth image $I_{\rm gt}$ (see Sec. 3.1). The start-of-sequence (SOS) token is computed from a fixed label index and augmented with a global context vector derived from the degraded image (see supplementary for details). These features are flattened and concatenated into a token sequence $\hat{f}_{\rm quant} \in \mathbb{R}^{L \times C}$, where L is the total number of tokens across all scales (see Fig. 3(a)(i)). The VAR transformer is then trained to autoregressively predict the next-scale indices $\{r_{\rm gt}^k\}_{k=1}^K \in \mathbb{R}^L$ of the clean image.

To enable semantically consistent restoration, we inject information from the degraded image through cross-attention at each transformer block. At block i, the queries are given by the output of the feed-forward network $(x_{\text{block}_i} \in \mathbb{R}^{L \times D})$, where D is the embedding dimension), while keys and values are derived from the continuous latent of the degraded image, $f_{\text{cont}}^{\text{deg}} \in \mathbb{R}^{H_K \times W_K \times C}$. This latent is reshaped into a sequence of conditioning tokens and is appropriately projected to the embedding dimension of the transformer. As shown in Sec. 4.4, conditioning on continuous latents significantly outperforms conditioning on discrete ones. To summarize, cross-attention $(CA(\cdot, \cdot))$ is applied as

$$x_{\text{block}_{CA}} = x_{\text{block}_i} + g_i \times CA(x_{\text{block}_i}, f_{\text{cont}}^{\text{deg}}).$$

We initialize $g_i=0$ to retain VAR's pretrained behavior and gradually introduce conditioning. Furthermore, we replace absolute positional embeddings in VAR with 2D Rotary Positional Embeddings (RoPE) for scaling resolution from 256×256 to 512×512 , as RoPE is well-suited for handling varying sequence lengths (Su et al., 2024). We also remove AdaLN layers, reducing ~ 100 M parameters with negligible impact on performance. Inference closely follows that of VAR (see Sec. 3.1), except that each scale prediction is now guided by the degraded latent. The output is a sequence of predicted indices $\{\hat{r}^k_{\text{pred}}\}_{k=1}^K$, which is then used to construct the discrete restored latent $f_{\text{quant}}^{\text{pred}} \in \mathbb{R}^{H_K \times W_K \times C}$. The above steps are shown in Fig. 3(a)(ii). More architectural details are given in the supplementary.

DETAIL-PRESERVING RESTORATION

270

271

272 273

274

275

276

278

279

280

281

282

283

284

285

287

288

289

290

291

292

293

295

296

297

298

299

300 301

302

303

304

305

306

307

308

309 310

311 312

313

314 315

316 317

318

319

320

321

322 323 The discrete latent $(f_{\text{quant}}^{\text{pred}})$ predicted by the RestoreVAR transformer is decoded by the VQVAE to produce the restored image. However, vector-quantization and VAE decoding cause a noticeable loss of fine details in the pixel-space, leading to distorted reconstructions. This presents a major challenge for using VAR in AiOR, as the scene semantics may not be accurately preserved. To address this, we introduce VAE decoder fine-tuning on continuous latents, and a lightweight latent refinement transformer (LRT) that converts discrete latents to continuous latents for decoding.

VAE Decoder Fine-Tuning. HART (Tang et al., 2024) addressed VAE-induced distortions by finetuning the VAE decoder on both discrete and continuous latents. While effective for generative tasks, the VAE decoder of HART produces overly textured outputs, compromising accurate reconstruction (see supplementary). Instead, we fine-tune the decoder only on continuous latents, bypassing the quantizer. The encoder and quantizer are kept frozen, and the decoder is trained on $(f_{\text{cont}}^{\text{gt}}, I_{\text{gt}})$ pairs. To avoid overly smooth outputs, we use a PatchGAN (Isola et al., 2017) discriminator (see Sec. 4.4) and optimize the decoder using pixel-wise, perceptual, and adversarial losses as

 $\mathcal{L}_{dec} = \lambda_1 \mathcal{L}_{L1} + \lambda_2 \mathcal{L}_{SSIM} + \lambda_3 \mathcal{L}_{percep} + \lambda_4 \mathcal{L}_{adv},$ where \mathcal{L}_{L1} is the L1 loss, \mathcal{L}_{SSIM} is the SSIM loss, \mathcal{L}_{percep} is the perceptual loss, \mathcal{L}_{adv} is the adversarial loss and λ_i are their respective weights (see Fig. 3(a)(iv)). Our fine-tuning approach yields a decoder that is well-aligned with the objectives of AiOR, achieving mean (over 1000 samples) reconstruction PSNR/SSIM scores of 28.14dB/0.842, outperforming both the VAR VQVAE (22.59dB/0.679) and HART decoders (26.48dB/0.804). Qualitative comparisons are given in the supplementary.

Refining Discrete Latents. Since the VAE decoder is fine-tuned for continuous latents, the predicted discrete latent, $f_{\text{quant}}^{\text{pred}}$, must be converted into a continuous form for decoding.

While HART uses a 37M parameter diffusionbased MLP for this, it incurs a $\sim 20\%$ inference overhead due to iterative denoising. Instead, we propose a lightweight, non-generative latent refinement transformer (LRT) that predicts a residual, which when added to $f_{\mathrm{quant}}^{\mathrm{pred}}$, produces a con-

tinuous latent, $\hat{f}_{\text{cont}} \in \mathbb{R}^{H_K \times W_K \times C}$ as $\hat{f}_{\text{cont}} = f_{\text{quant}}^{\text{pred}} + \text{LRM}(f_{\text{quant}}^{\text{pred}}, z),$ where $z \in \mathbb{R}^{L \times D}$ is the output from the final





Input Discrete Refiner Continuous Figure 4: Illustration of images decoded from discrete and continuous latents, along with the refiner's predicted residuals.

RestoreVAR transformer block. z is passed through cross-attention and provides pseudo-continuous guidance to the LRT which is critical for performance (see Sec. 4.4). The LRT is trained using \mathcal{L}_1 loss between the predicted and ground-truth continuous latents ($f_{\text{cont}}^{\text{gt}}$) as $\mathcal{L}_{\text{LRT}} = \mathcal{L}_1(\hat{f}_{\text{cont}}, f_{\text{cont}}^{\text{gt}})$. Our LRT introduces only 3% additional overhead and significantly outperforms HART's refiner in PSNR and SSIM scores (see Sec. 4.4). The training procedure of the LRT is shown in Fig. 3(a)(iii) and Fig. 4 provides a visual example of its predictions.

Thus, RestoreVAR combines the VAR transformer, LRT, and fine-tuned decoder to deliver fast, perceptually realistic, and structurally faithful results. Fig. 3(b) depicts inference of RestoreVAR.

EXPERIMENTS

In this section, we provide implementation details, comparisons with existing All-in-One image Restoration (AiOR) approaches, and present ablations on key components of our framework.

4.1 IMPLEMENTATION DETAILS

Each component of Restore VAR was trained independently to disentangle learning objectives. We used the VAR model of depth 16 as the transformer backbone and trained it with the AdamW optimizer (Loshchilov & Hutter, 2017), a learning rate (LR) of 10^{-4} , batch size of 48, for 100 epochs. The latent refiner was trained for 100 epochs with the AdamW optimizer, LR= 10^{-4} and a batch size of 96. The VAE decoder was fine-tuned using a weighted loss combination (see Sec. 3.3.2) with empirically chosen weights: $\lambda_1 = 2.0$, $\lambda_2 = 0.4$, $\lambda_3 = 0.2$, and $\lambda_4 = 0.01$. Fine-tuning was performed for 5 epochs with a learning rate of 3×10^{-4} and a batch size of 12, using AdamW. Training was conducted on 8 RTX A6000 GPUs, while inference was done on an RTX 4090 GPU.

Table 1: Quantitative comparisons of RestoreVAR with the state-of-the-art LDM-based generative AiOR approaches, and non-generative methods. RestoreVAR significantly outperforms generative methods on PSNR, SSIM and LPIPS scores. The best generative approach is indicated in **bold**.

Method	Venue	Venue		RESIDI	Ε	S	now100	k		Rain13I	ζ.	LOLv1				GoPro	
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
Non-generative methods																	
PromptIR	NeurIPS'23	32.02	0.952	0.013	31.98	0.924	0.115	29.56	0.888	0.087	22.89	0.847	0.296	27.21	0.817	0.250	
InstructIR	ECCV'24	26.90	0.952	0.017	-	_	-	29.56	0.885	0.088	22.81	0.836	0.132	28.26	0.870	0.146	
AWRaCLe	AAAI'25	30.81	0.979	0.013	30.56	0.904	0.088	31.26	0.908	0.068	21.04	0.818	0.146	26.78	0.820	0.248	
DCPT	ICLR'25	29.10	0.968	0.017	_	_	_	24.11	0.766	0.203	23.67	0.863	0.106	27.92	0.877	0.169	
DFPIR	CVPR'25	31.39	0.979	0.012	-	-	-	24.87	0.794	0.171	23.12	0.853	0.123	28.66	0.884	0.158	
						Gener	rative me	thods									
Diff-Plugin	CVPR'24	23.23	0.765	0.091	21.02	0.611	0.196	21.71	0.617	0.169	19.38	0.713	0.195	21.76	0.633	0.217	
AutoDIR	ECCV'24	24.48	0.780	0.081	19.00	0.515	0.347	23.02	0.642	0.162	19.43	0.766	0.135	23.55	0.700	0.168	
PixWizard	ICLR'25	21.28	0.738	0.142	21.24	0.594	0.206	21.38	0.596	0.180	15.84	0.629	0.305	20.49	0.602	0.223	
RestoreVAR (Ours))	24.67	0.821	0.074	24.05	0.713	0.156	23.97	0.700	0.153	21.72	0.782	0.126	23.96	0.737	0.167	

Input	Diff-Plugin	AutoDIR	PixWizard	RestoreVAR	GT
RESIDE					
Rain 13K					
GoPro					

Figure 5: Qualitative comparisons of RestoreVAR with LDM-based generative AiOR approaches. RestoreVAR achieves consistent restoration with enhanced preservation of fine-details.

4.2 Datasets

We trained RestoreVAR for five tasks: dehazing, desnowing, deraining, low-light enhancement and deblurring. For dehazing, we used the RESIDE (Li et al., 2019) dataset comprising 72135 training and 500 test images. The Snow100k dataset (Liu et al., 2018) was used for desnowing, with 50000 training and 16801 test images (heavy subset). For deraining, we used Rain13K (Zamir et al., 2021) consisting of 13711 training and 4298 test images. The LOLv1 (Wei et al., 2018) dataset was used for low-light enhancement, consisting of 485 training and 15 test images. For deblurring, we used the GoPro (Nah et al., 2017) dataset comprising 2103 training and 1111 test images. We also assess generalization performance on real-world, unseen and mixed degradation datasets, namely, LHP (Guo et al., 2023) (1000 images), REVIDE (Zhang et al., 2021b) (284 images), TOLED (Zhou et al., 2021) (30 images), POLED (Zhou et al., 2021) (30 images), CDD (Guo et al., 2024) (200 images, mix of haze and rain), and LOLBlur (Zhou et al., 2022) (482 images, mix of low-light and blur). TOLED and POLED datasets contain unseen degradation of under-display camera restoration.

4.3 Comparisons

We compare RestoreVAR with state-of-the-art generative and non-generative methods for AiOR. For non-generative approaches, we include PromptIR (Potlapalli et al., 2024), InstructIR (Conde et al., 2025), AWRaCLe (Rajagopalan & Patel, 2024), DCPT (Hu et al., 2025) and DFPIR (Tian et al., 2025). Among generative methods, we compare with the LDM-based approaches Diff-Plugin (Liu et al., 2024), AutoDIR (Jiang et al., 2023) and PixWizard (Lin et al., 2024). To ensure a fair comparison, we retrained PromptIR and AWRaCLe, as their official checkpoints were not trained for most of our AiOR tasks. All other methods were evaluated using their publicly released checkpoints. For AutoDIR, we report results without the structure correction module, as this module functions as an independent, non-generative restoration network (more details in supplementary). The results reported for PixWizard were obtained using its publicly released checkpoint. We do not compare with task-specific restoration models, as RestoreVAR is proposed for the AiOR setting.

Table 1 presents PSNR, SSIM and LPIPS scores on the RESIDE, Snow100k, Rain13K, LOLv1 and GoPro datasets. RestoreVAR surpasses LDM-based AiOR methods at a fraction of their computa-

Table 2: Quantitative comparisons of RestoreVAR against state-of-the-art non-generative approaches on real-world, unseen and mixed degradations. The best result is indicated in bold.

Method	LHP		REVIDE		TOLED		POLED		LOLBlur (L + B)		CDD (H + R)		Average	
	MUSIQ↑	CLIPIQA↑	MUSIQ↑	CLIPIQA↑	MUSIQ↑	CLIPIQA↑	MUSIQ↑	CLIPIQA↑	MUSIQ↑	CLIPIQA↑	MUSIQ↑	CLIPIQA↑	MUSIQ↑	CLIPIQA↑
PromptIR	56.780	0.366	61.191	0.459	43.218	0.281	34.536	0.303	33.693	0.166	65.895	0.483	49.219	0.343
InstructIR	58.269	0.359	63.116	0.416	44.985	0.298	23.317	0.241	40.221	0.202	65.491	0.482	49.900	0.333
AWRaCLe	57.889	0.333	59.287	0.368	44.670	0.285	40.533	0.332	38.186	0.171	66.253	0.484	51.470	0.329
DCPT	58.044	0.372	60.011	0.446	44.062	0.314	38.138	0.345	37.393	0.175	68.440	0.544	51.681	0.366
DFPIR	56.483	0.330	61.009	0.450	43.820	0.276	35.668	0.289	36.277	0.163	54.408	0.349	47.611	0.310
RestoreVAR	57.662	0.414	63.562	0.483	52.374	0.338	48.118	0.276	46.644	0.214	68.941	0.572	56.217	0.383

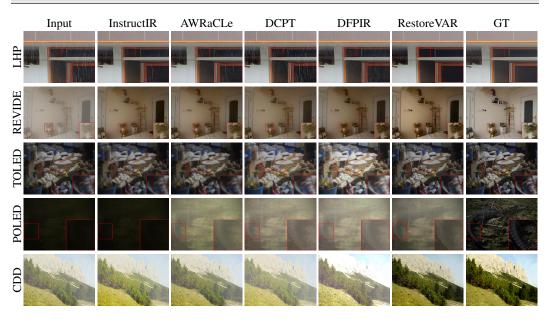


Figure 6: Qualitative comparisons of RestoreVAR with non-generative methods on real, unseen and mixed degradations. RestoreVAR consistently achieves better results.

tional cost (inference time (s) per image)—Diff-Plugin: 2.04s, AutoDIR: 8.477s, PixWizard: 8.247s and RestoreVAR: 0.201s, highlighting the efficacy of our framework. More detailed complexity comparisons are given in the supplementary along with a derivation showing that the time complexity of VAR with maximum latent resolution $n \times n$ is $\mathcal{O}(\log n)$ lower than an LDM operating at the same latent resolution. Qualitative comparisons with LDM-based methods in Fig. 5 further illustrate that RestoreVAR produces restored images of high quality while better preserving fine details. Visual results for the Snow100k and LOLv1 datasets are provided in the supplementary. While non-generative methods achieve better scores, it is important to recognize that the performance of RestoreVAR is inherently influenced by the quality of the VAE decoder; a limitation shared by all latent generative approaches. Despite this constraint, RestoreVAR narrows the gap with non-generative methods while maintaining the benefits of a generative framework, i.e., perceptually realistic results and strong generalization capabilities. To demonstrate these strengths, we evaluate generalization using no-reference image quality metrics (following prior works (Liu et al., 2024; Jiang et al., 2023; Rajagopalan & Patel, 2024)), and assess perceptual realism through a user study.

For testing generalization, we report MUSIQ (Ke et al., 2021) and CLIPIQA (Wang et al., 2023) scores in Table 2 on the real-world, unseen and mixed degradation datasets discussed in Sec. 4.2.

RestoreVAR achieves higher scores than non-generative models (on average), indicating better robustness under these degradations. Qualitative results for this experiment are shown in Fig. 6, where RestoreVAR consistently outperforms non-generative approaches. Due to space constraints, qualitative comparisons with PromptIR and visual results for LOLBlur are given in the supplementary. To further evaluate perceptual quality, we conducted a user study in which participants rated outputs from non-generative models, AutoDIR (LDM-based) and RestoreVAR, for 50 real-world scenes. We received 36 responses with each participant scoring outputs based on scene consistency, restoration quality, and overall appeal on a 5-point scale. Table 3

Table 3: Mean scores from user study.

Method	Score ↑
PromptIR	2.11
InstructIR	2.93
AWRaCLe	2.33
DCPT	2.42
DFPIR	2.35
AutoDIR	3.68
RestoreVAR	4.36

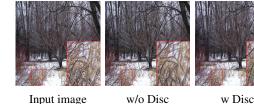


Figure 7: Validation accuracy of RestoreVAR under discrete vs. continuous conditioning.

Figure 8: Image reconstructed by VAE decoders fine-tuned on continuous latents with (w) and without (w/o) a discriminator (Disc).

shows that RestoreVAR received the highest average ratings (across all three criteria), highlighting its ability to produce images that align closely with human preferences.

These results highlight the effectiveness of RestoreVAR for AiOR. We discuss limitations of Restore-VAR in the supplementary.

4.4 ABLATIONS

432 433

438 439

440

441

442 443 444

445

446

447

448 449 450

451 452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473 474

475 476

477 478 479

480

481

482

483

484

485

Continuous vs. Discrete Conditioning. The RestoreVAR transformer conditions on the continuous latent of the degraded image ($f_{\text{cont}}^{\text{deg}}$). While conditioning with discrete multi-scale latents appears more aligned with VAR's multi-scale prediction objective, it results in significantly worse performance. To demonstrate this, we train RestoreVAR with discrete and continuous conditioning for 15 epochs each. As shown in Fig. 7, RestoreVAR with discrete conditioning exhibits much lower validation accuracy.

Discriminator for VAE fine-tuning. As described in Sec. 3.3.2, we fine-tune the VAE decoder on continuous latents using a combination of pixel-level loss and an adversarial loss. To analyze the impact of the discriminator, we compare the reconstructions of VAE decoders fine-tuned with and without the adversarial loss. As shown in Fig. 8, removing the discriminator leads to blurrier reconstruction while including it yields sharper and perceptually better looking outputs.

Latent Refiner Transformer. The Latent Refiner Table 4: Ablations on the types of latent refin-Transformer (LRT) is critical for preserving pixellevel detail in restored images. To analyze its impact, we compare four RestoreVAR variants: (i) No refiner, (ii) HART's diffusion refiner, (iii) LRT without final block outputs, and (iv) our proposed LRT. As shown in Table 4, our LRT achieves the best PSNR and SSIM, while maintaining low inference time and a low parameter count. Using no refiner yields poor

ers. Our proposed latent refiner transformer (LRT) performs best, with minimal overhead.

Refiner Type	Time (s)	Params (M)	PSNR / SSIM
No Refiner	_	_	21.71 / 0.690
HART Refiner	0.0455	36.06	23.48 / 0.777
LRT w/o Last-Block	0.0036	14.61	21.23 / 0.660
Proposed LRT	0.0061	22.97	24.67 / 0.821

PSNR/SSIM scores. Removing the last block outputs significantly reduces performance, indicating its importance as pseudo-continuous guidance for refinement. HART's MLP diffusion-based refiner performs worse than our LRT while having a much higher parameter count and runs $\sim 7 \times$ slower.

More ablations are provided in the supplementary.

5 **CONCLUSIONS**

We proposed RestoreVAR, a fast and effective generative approach for AiOR. Built on the VAR backbone, RestoreVAR benefits from VAR's strong generative priors and significantly faster inference compared to LDMs. To tailor VAR for AiOR, we introduced cross-attention mechanisms that inject semantic information from the degraded image into the generation process. Additionally, we proposed a non-generative latent refiner transformer to convert discrete latents to continuous ones, along with a VAE decoder fine-tuned on continuous latents, which together improve reconstruction fidelity. RestoreVAR achieves state-of-the-art performance among generative AiOR models, outperforming LDM-based methods while delivering over $10 \times$ faster inference and strong generalization.

ETHICS STATEMENT

We acknowledge that we have read and adhered to the ICLR Code of Ethics. For all our experiments, we used publicly available open-source datasets.

REPRODUCIBILITY STATEMENT

Our code will be made publicly available after the review process. Details to reproduce the work have been provided in Secs. 3 and 4.1 of the main paper and Sec. E of the supplementary.

REFERENCES

Mark Chen, Al

- Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pp. 1691–1703. PMLR, 2020.
- Xi Chen, Nikhil Mishra, Mostafa Rohaninejad, and Pieter Abbeel. Pixelsnail: An improved autoregressive generative model. In *International conference on machine learning*, pp. 864–872. PMLR, 2018.
 - Xiang Chen, Jinshan Pan, Kui Jiang, Yufeng Li, Yufeng Huang, Caihua Kong, Longgang Dai, and Zhentao Fan. Unpaired deep image deraining using dual contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2017–2026, 2022.
 - Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *European Conference on Computer Vision*, pp. 1–21. Springer, 2025.
 - Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Shahbaz Khan. Adair: Adaptive all-in-one image restoration via frequency mining and modulation. *arXiv* preprint arXiv:2403.14614, 2024.
 - Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.
 - Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
 - Hang Guo, Yawei Li, Taolin Zhang, Jiangshan Wang, Tao Dai, Shu-Tao Xia, and Luca Benini. Fastvar: Linear visual autoregressive modeling via cached token pruning. *arXiv preprint arXiv:2503.23367*, 2025.
 - Yu Guo, Yuan Gao, Yuxu Lu, Huilin Zhu, Ryan Wen Liu, and Shengfeng He. Onerestore: A universal restoration framework for composite degradation. In *European conference on computer vision*, pp. 255–272. Springer, 2024.
 - Yun Guo, Xueyao Xiao, Yi Chang, Shumin Deng, and Luxin Yan. From sky to the ground: A large-scale benchmark and simple baseline towards real rain removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12097–12107, October 2023.
 - Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1956–1963, 2009. doi: 10.1109/CVPR.2009.5206515.
 - Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
 - JiaKui Hu, Lujia Jin, Zhengjian Yao, and Yanye Lu. Universal image restoration pre-training via degradation classification. *arXiv preprint arXiv:2501.15510*, 2025.

- Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.
 - Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu. Autodir: Automatic all-in-one image restoration with latent diffusion. *arXiv* preprint arXiv:2310.10123, 2023.
 - 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.
 - Diederik P Kingma. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.
 - Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8183–8192, 2018.
 - Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28 (1):492–505, 2019. doi: 10.1109/TIP.2018.2867951.
 - Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17431–17441, 2022. doi: 10.1109/CVPR52688.2022.01693.
 - Ruoteng Li, Robby T. Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3172–3182, 2020. doi: 10.1109/CVPR42600.2020.00324.
 - Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 1833–1844, 2021. doi: 10.1109/ICCVW54120.2021. 00210.
 - Weifeng Lin, Xinyu Wei, Renrui Zhang, Le Zhuo, Shitian Zhao, Siyuan Huang, Huan Teng, Junlin Xie, Yu Qiao, Peng Gao, et al. Pixwizard: Versatile image-to-image visual assistant with openlanguage instructions. *arXiv* preprint arXiv:2409.15278, 2024.
 - Yuhao Liu, Zhanghan Ke, Fang Liu, Nanxuan Zhao, and Rynson WH Lau. Diff-plugin: Revitalizing details for diffusion-based low-level tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4197–4208, 2024.
 - Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing*, 27(6):3064–3073, 2018. doi: 10.1109/TIP.2018.2806202.
 - Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint* arXiv:1711.05101, 2017.
 - Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for universal image restoration. *arXiv preprint arXiv:2310.01018*, 3(8), 2023.
 - Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3883–3891, 2017.
 - William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.
 - 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.

- Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36, 2024.
 - Yunpeng Qu, Kun Yuan, Jinhua Hao, Kai Zhao, Qizhi Xie, Ming Sun, and Chao Zhou. Visual autoregressive modeling for image super-resolution. *arXiv preprint arXiv:2501.18993*, 2025.
 - Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019a.
 - Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019b.
 - Sudarshan Rajagopalan and Vishal M Patel. Awracle: All-weather image restoration using visual in-context learning. *arXiv preprint arXiv:2409.00263*, 2024.
 - Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.
 - Sucheng Ren, Yaodong Yu, Nataniel Ruiz, Feng Wang, Alan Yuille, and Cihang Xie. M-var: Decoupled scale-wise autoregressive modeling for high-quality image generation. *arXiv* preprint *arXiv*:2411.10433, 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.
 - Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv* preprint *arXiv*:1701.05517, 2017.
 - Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv* preprint arXiv:2010.02502, 2020.
 - Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
 - Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv* preprint arXiv:2410.10812, 2024.
 - Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.
 - Xiangpeng Tian, Xiangyu Liao, Xiao Liu, Meng Li, and Chao Ren. Degradation-aware feature perturbation for all-in-one image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28165–28175, 2025.
 - Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
 - J. Jose Valanarasu, R. Yasarla, and V. M. Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2343–2353, 2022.
- Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.
 - Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pp. 1747–1756. PMLR, 2016.

- Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
 - Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.
 - Siyang Wang and Feng Zhao. Varformer: Adapting var's generative prior for image restoration. *arXiv* preprint arXiv:2412.21063, 2024.
 - Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson W.H. Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12262–12271, 2019. doi: 10.1109/CVPR. 2019.01255.
 - Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.
 - Rajeev Yasarla and Vishal M Patel. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8405–8414, 2019.
 - Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18456–18466, 2023.
 - Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021.
 - 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.
 - He Zhang, Vishwanath Sindagi, and Vishal M. Patel. Joint transmission map estimation and dehazing using deep networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7): 1975–1986, 2020. doi: 10.1109/TCSVT.2019.2912145.
 - Kaihao Zhang, Rongqing Li, Yanjiang Yu, Wenhan Luo, and Changsheng Li. Deep dense multiscale network for snow removal using semantic and depth priors. *IEEE Transactions on Image Processing*, 30:7419–7431, 2021a. doi: 10.1109/TIP.2021.3104166.
 - Xinyi Zhang, Hang Dong, Jinshan Pan, Chao Zhu, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Fei Wang. Learning to restore hazy video: A new real-world dataset and a new method. In *CVPR*, pp. 9239–9248, 2021b.
 - Dian Zheng, Xiao-Ming Wu, Shuzhou Yang, Jian Zhang, Jian-Fang Hu, and Wei-shi Zheng. Selective hourglass mapping for universal image restoration based on diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
 - Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In *European conference on computer vision*, pp. 573–589. Springer, 2022.
 - Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, and Timothy Large. Image restoration for under-display camera. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 9179–9188, 2021.

A OVERVIEW OF SUPPLEMENTARY

In this supplementary, we first present detailed computational complexity comparisons of RestoreVAR and LDM-based AiOR approaches. We then provide a theoretical analysis comparing the time complexities of VAR and LDMs. Subsequently, we analyze the effect of using Absolute Positional Embeddings (APE) versus Rotary Positional Embeddings (RoPE) (Su et al., 2024) when scaling the resolution from 256×256 to 512×512 . Next, we present more architectural details of RestoreVAR, followed by a breakdown of the runtime and parameter count for each component of RestoreVAR. We then provide additional visual results, which include qualitative results for the continuous vs. discrete conditioning ablation, visual comparisons of VAE decoders, refiner ablation, and more qualitative comparisons with other methods. Subsequently, we provide experiments to show that the structure correction module of AutoDIR (Jiang et al., 2023) behaves as an independent non-generative restoration network. We then discuss the limitations of our approach and scope for future work. Finally, we mention the usage of LLMs in the paper. To summarize, the supplementary discusses the following:

- 1. Detailed computation complexity comparisons (Sec. B)
- 2. Theoretical complexity comparison with LDM (Sec. C)
- 3. Performance analysis: APE vs. RoPE (Sec. D)
- 4. Additional Architectural Details (Sec. E)
- 5. Runtime and Parameter Breakdown (Sec. F)
- 6. Additional Visual Results (Sec. G)
 - (a) Continuous vs. Discrete Conditioning (Sec. G.1)
 - (b) Qualitative comparisons of VAE decoders (Sec. G.2)
 - (c) Visual Comparison of Refiner Variants (Sec. G.3)
 - (d) Additional Qualitative Comparisons (Sec. G.4)
- 7. More details about AutoDIR comparison (Sec. H)
- 8. Limitations and scope for future work (Sec. I
- 9. LLM Usage (Sec. J)

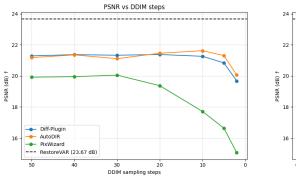
B DETAILED COMPUTATION COMPLEXITY COMPARISONS

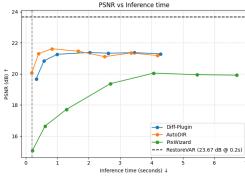
RestoreVAR achieves substantial performance improvements over LDM-based AiOR approaches at a fraction of their computational cost. To show this, we compared RestoreVAR with Diff-Plugin (Liu et al., 2024), AutoDIR (Jiang et al., 2023), and PixWizard (Lin et al., 2024) in terms of inference steps, runtime, TeraFLOPs, and total parameter count. As shown in Table 5, RestoreVAR achieves a $10\times$ speed-up over Diff-Plugin and a \sim $16\times$ reduction in TFLOPs. Compared to AutoDIR and PixWizard, RestoreVAR is over $40\times$ faster in inference.

Table 5: Comparison of the computational complexity of RestoreVAR with LDM-based AiOR approaches.

Method	Steps	Time (s)	TFLOPs	Params (M)
Diff-Plugin	20	2.04	16.08	859.50
AutoDIR	100	8.477	67.80	859.50
PixWizard	60	8.247	19.27	2011.40
RestoreVAR	10	0.201	1.05	296.95

Additionally, we conducted an experiment to speed up LDM sampling using DDIM Song et al. (2020) sampling and compared the performance with RestoreVAR. Specifically, we varied the number of DDIM sampling steps as 50, 40, 30, 20, 10, 5 and 2. Figs. 9(a) and (b) shows the variation of mean (across the datasets from Table 1) PSNR (dB) scores with inference time (seconds) and DDIM sampling steps. Diffusion models match the inference time of RestoreVAR only when using around 2 sampling steps, but show over a 3dB decrease in PSNR compared to RestoreVAR. Even at higher step counts with DDIM sampling, diffusion models lag by over 1dB, highlighting RestoreVAR's clear advantages in both speed and performance over LDM-based methods.





(a) PSNR vs DDIM steps

(b) PSNR vs inference time

Figure 9: Comparison of LDM-based methods accelerated using DDIM sampling and RestoreVAR. RestoreVAR achieves best results.

THEORETICAL COMPLEXITY COMPARISON WITH LDM

We now provide a theoretical comparison of the run-time complexities of VAR and diffusion transformers (DiT), offering fundamental insights into their efficiency differences. The VAR time complexity derivation closely follows that in the original VAR paper.

Let a > 1 be the geometric factor of the vector quantized (VQ) scale pyramid and let the largest scale have dimensions h = w = n. Let the number of scales be $K = \log_a n + 1$ so that the side length at scale i is $n_i = a^{i-1}$ and the largest scale is $n_K = n = a^{K-1}$. Assume a standard self-attention transformer as in VAR with time complexity $\mathcal{O}(T^2)$ for T tokens.

VAR. At the generation of the k-th scale the total number of tokens across the current and previous scales (r_1, \ldots, r_k) is

$$\sum_{i=1}^{k} n_i^2 = \sum_{i=1}^{k} a^{2(i-1)} = \frac{a^{2k} - 1}{a^2 - 1}.$$

Hence the cost of generation of the k-th scale is

$$C_k^{\text{VAR}} = \left(\frac{a^{2k} - 1}{a^2 - 1}\right)^2.$$

Summing over all K scales gives

$$C^{\text{VAR}} = \sum_{k=1}^{K} \left(\frac{a^{2k} - 1}{(a^2 - 1)} \right)^2 = \frac{1}{(a^2 - 1)^2} \sum_{k=1}^{K} \left(a^{4k} - 2a^{2k} + 1 \right).$$

Substituting
$$K = \log_a n + 1$$
 (so that $a^{2K} = a^2 n^2$ and $a^{4K} = a^4 n^4$) yields
$$C^{\text{VAR}} = \frac{1}{(a^2-1)^2} \left[\frac{a^4(n^4-1)}{a^4-1} - \frac{2a^2(a^2n^2-1)}{a^2-1} + (\log_a n + 1) \right].$$

The asymptotic time complexity is governed by the dominant term:

$$C^{\text{VAR}} \sim \frac{a^8}{(a^4 - 1)(a^2 - 1)^2} n^4 = \mathcal{O}(n^4).$$

Diffusion. Assume a self-attention DiT where each diffusion step uses the fixed largest resolution $n \times n$, i.e., n^2 tokens. So, a single step costs

$$C_1^{\text{Diff}} = (n^2)^2 = n^4.$$

With the same number of forward steps as VAR, namely $K = \log_a n + 1$, we get

$$C^{\text{Diff}} = \sum_{k=1}^{K} n^4 = (\log_a n + 1)n^4 = \mathcal{O}(n^4 \log n).$$

Comparison. From the above,

$$\frac{C^{\mathrm{Diff}}}{C^{\mathrm{VAR}}} \sim \frac{\left(\log_a n\right) n^4}{\frac{(a^4-1)(a^2-1)^2}{a^8} \, n^4} = \frac{a^8}{(a^4-1)(a^2-1)^2} \, \log_a n = \mathcal{O}(\log n).$$

That is, with the same number of forward passes, VAR totals $\mathcal{O}(n^4)$ complexity while diffusion totals $\mathcal{O}(n^4 \log n)$, yielding a $\mathcal{O}(\log n)$ speedup for VAR.

D PERFORMANCE ANALYSIS: APE vs. RoPE

As discussed in Sec. 3.3, we replace the absolute position embeddings (APE) used in VAR (Tian et al., 2024) with Rotary Positional Embeddings (Su et al., 2024) (RoPE). We found that using RoPE yields better performance when scaling the resolution from 256×256 to 512×512 .

To demonstrate this, we conducted an ablation where both APE and RoPE-based variants were fine-tuned at 512×512 resolution for 10 epochs. As shown in Fig. 10, the RoPE-based model achieves higher validation accuracy compared to the APE-based model, indicating its effectiveness.

APE RoPE APE APE APE ROPE APE APE ROPE ROPE APE ROPE ROPE APE ROPE ROPE APE ROPE ROPE

E ADDITIONAL ARCHITECTURAL DETAILS

We now provide additional architectural details for the RestoreVAR framework. We first describe the details for the RestoreVAR transformer, followed by the Latent Refiner Transformer (LRT).

Figure 10: Validation accuracy comparison of APE and RoPE-based fine-tuning at 512×512 resolution. RoPE demonstrates better performance.

RestoreVAR Transformer. We adopted the VAR model with a transformer depth of 16, i.e., the architecture consists of 16 transformer

blocks. The structure of each block is illustrated in Fig. 11(a). The embedding dimension was set to 1024, and the number of attention heads used was 16. Furthermore, the transformer predicted discrete latents at the following spatial resolutions in the latent space: 1×1 , 2×2 , 3×3 , 4×4 , 6×6 , 9×9 , 13×13 , 18×18 , 24×24 , and 32×32 . The start-of-sequence (SOS) token is constructed by augmenting the class embedding with the mean value (along spatial dimensions) of the features obtained after a learnable projection applied on $f_{\text{cont}}^{\text{deg}}$. Specifically,

$$SOS = class_{emb} + g_{sos} \times Mean(Proj(f_{cont}^{deg}), SOS \in \mathbb{R}^{1 \times C}.$$

Here, class_{emb} is the class token embedding and g_{sos} is initialized as 0 for gradual incorporation of degradation conditioning. Other notations follow Sec. 3.

Latent Refiner Transformer. The LRT follows a similar structure for the blocks as the RestoreVAR transformer, as shown in Fig. 11(b). It was configured with a depth of 12, six attention heads, and an embedding dimension of 384.

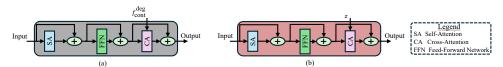


Figure 11: Illustration of a transformer block in (a) RestoreVAR transformer and (b) Latent Refiner Transformer.

F RUNTIME AND PARAMETER BREAKDOWN

In this section, we provide a breakdown of the runtime and parameter count for the following components of the RestoreVAR framework: the VAE, the RestoreVAR transformer, and the Latent Refiner Transformer (LRT). This analysis provides insights into the distribution of the computational cost across the pipeline. As shown in Table 6, the majority of inference time is taken by the autoregressive Restore-VAR transformer.

Component	VAE	Transformer	Refiner
Time (s)	0.0086	0.1863	0.0061
Parameters (M)	108.95	273.98	22.97

Table 6: Compute time and parameter count breakdown for each component of Restore-VAR. VAE time includes both encoding and decoding.

G ADDITIONAL VISUAL RESULTS

We now present additional visualizations for some of the ablations discussed in Sec. 4.4, along with more qualitative comparisons across methods.

G.1 CONTINUOUS VS. DISCRETE CONDITIONING

As shown in Sec. 4.4, conditioning RestoreVAR on the continuous latent of the degraded image yields significantly better performance compared to using the quantized or discrete latent. Fig. 12 further illustrates this using visual comparisons between the model trained with discrete and continuous conditioning. The model trained with discrete conditioning exhibits noticeably more hallucinations than the one trained with continuous conditioning.

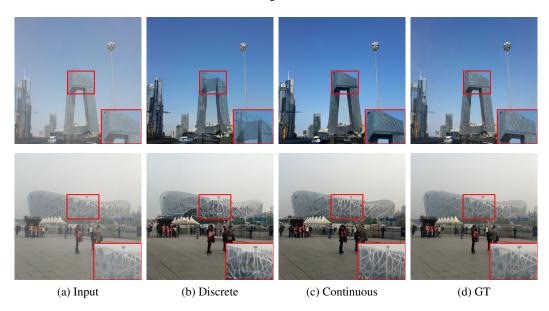


Figure 12: Qualitative comparisons of RestoreVAR under discrete vs. continuous conditioning. RestoreVAR with discrete conditioning exhibits more hallucinations than the variant with continuous conditioning.

G.2 QUALITATIVE COMPARISONS OF VAE DECODERS

As mentioned in Sec. 3.3.2, our fine-tuned VQ-VAE decoder achieves superior reconstruction performance compared to the decoders of VAR and HART. Fig. 13, provides qualitative results to illustrate the same. Our decoder produces the best reconstruction.

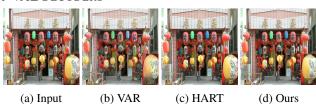


Figure 13: Qualitative comparisons of the input reconstructed using VAR (Tian et al., 2024), HART (Tang et al., 2024) and Our VAE decoder. Our result has minimal distortions.

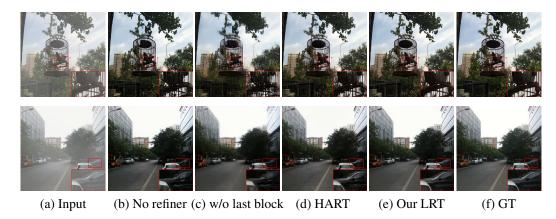


Figure 14: Qualitative results for the ablation on latent refiner configurations. Our proposed LRT preserves fine details better than the other configurations.

G.3 VISUAL COMPARISON OF REFINER VARIANTS

In Sec. 4.4, we demonstrated that our proposed Latent Refiner Transformer (LRT) achieves the best performance compared to using no refiner, a refiner without last-block conditioning, and HART (Tang et al., 2024)'s diffusion-based refiner. Quantitative results, reported in Table 4, included mean PSNR and SSIM scores on the RESIDE (Li et al., 2019) test set. Fig. 14 presents qualitative comparisons for these configurations. It can be observed that our LRT preserves fine details more effectively than the other variants.

G.4 ADDITIONAL QUALITATIVE COMPARISONS

In this section, we provide additional qualitative comparisons of RestoreVAR with state-of-the-art LDM-based and non-generative all-in-one image restoration (AiOR) methods. Fig. 15 presents results from RestoreVAR alongside Diff-Plugin(Liu et al., 2024), AutoDIR (Jiang et al., 2023), and PixWizard (Lin et al., 2024) on the RESIDE (Li et al., 2019), Snow100k (Liu et al., 2018), Rain13K (Zamir et al., 2021), LOLv1 (Wei et al., 2018), and GoPro (Nah et al., 2017) datasets. RestoreVAR consistently produces outputs that are more semantically aligned with the ground truth (see zoomed-in patches).

Fig. 16 provides comparisons with non-generative methods—PromptIR(Potlapalli et al., 2024), InstructIR (Conde et al., 2025), AWRaCLe (Rajagopalan & Patel, 2024), DCPT (Hu et al., 2025) and DFPIR (Tian et al., 2025)—for real-world, mixed and unseen degradation generalization on LHP (Guo et al., 2023), REVIDE (Zhang et al., 2021b), TOLED (Zhou et al., 2021), POLED (Zhou et al., 2021), CDD (Guo et al., 2024) and LOLBlur (Zhou et al., 2022) datasets. RestoreVAR generates sharper, more realistic outputs with fewer artifacts than non-generative models. For instance, for the TOLED and POLED cases, RestoreVAR outputs are visibly sharper than non-generative methods. Similarly, the results of RestoreVAR are superior in the case of mixed degradations.

H MORE DETAILS ABOUT AUTODIR COMPARISON

As mentioned in the main paper, comparisons with AutoDIR (Jiang et al., 2023) were conducted without its structure correction module (SCM). AutoDIR consists of a latent diffusion model (LDM) for initial restoration, followed by an SCM which is a non-generative post-processing network. The intuition behind this approach is that the SCM predicts a residual based on the degraded input image and the restored output of the LDM, to correct the VAE-induced distortions. In short,

$$I_{\text{result}} = I_{\text{sd}} + \mathcal{F}\left(\left[I_{\text{sd}}, I_{\text{deg}}\right]\right),$$

where $I_{\rm sd}$ is the restored output from the LDM, $I_{\rm deg}$ is the original degraded input image, and $\mathcal{F}(\cdot)$ denotes the SCM which operates on the concatenated inputs $[I_{\rm sd},I_{\rm deg}]$. However, we found that instead of slightly modulating the structural details in $I_{\rm sd}$, the SCM behaves like a separate nongenerative restoration model which directly restores $I_{\rm deg}$. To show this, we evaluated AutoDIR with

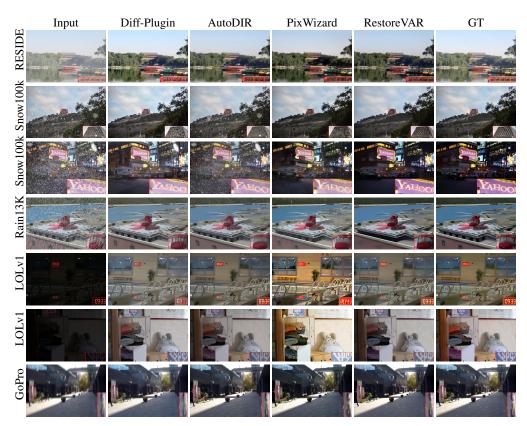


Figure 15: Additional qualitative comparisons of RestoreVAR with LDM-based AiOR approaches. RestoreVAR consistently preserves fine details more effectively than the LDM-based methods.

the SCM on the RESIDE (Li et al., 2019) dataset for two cases: (1) using $I_{\rm sd}$ as the actual LDM output and (2) setting $I_{\rm sd}=0$, effectively removing any structural information from the LDM. If the SCM were functioning as a corrective module, performance in the second case should deteriorate significantly. However, we found that the SCM was able to independently restore the degraded input in the second case, as shown in Fig. 18. This suggests that the SCM largely ignores the LDM output and instead performs direct restoration on $I_{\rm deg}$, thereby behaving as a non-generative restoration network. Therefore, to ensure a fair comparison with other generative models, we evaluated only the LDM output of AutoDIR.

I LIMITATIONS AND SCOPE FOR FUTURE WORK

Despite the strengths of RestoreVAR, there remains scope for improvement. First, its performance is inherently constrained by the latent refiner transformer (LRT) and the VAE decoder. While the LRT significantly improves results over using no refiner, it does not reach the upper bound set by directly decoding from ground-truth continuous latents. Exploring improved VQVAE and refiner architectures could help address this. Another promising direction is to employ our non-generative LRT in fully generative VAR models, given its strong performance for AiOR. Finally, future work can investigate how the performance of RestoreVAR scales with larger VAR backbones.

J LLM USAGE

LLM was used only for polishing writing in parts of the main paper and supplementary.

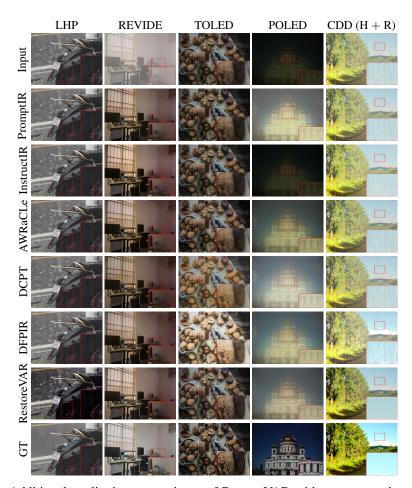


Figure 16: Additional qualitative comparisons of RestoreVAR with non-generative methods on real-world, unseen and mixed degradations. RestoreVAR achieves better results, highlighting its superior generalization.



Figure 17: Qualitative comparisons with non-generative methods on samples from the real mixed-degradation dataset LOLBlur.

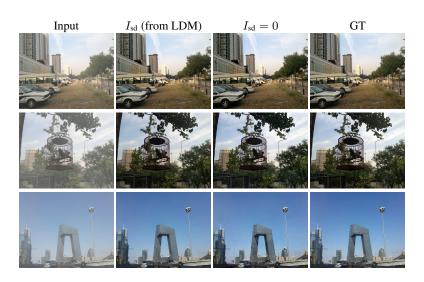


Figure 18: Illustration of the behavior of AutoDIR's (Jiang et al., 2023) structure correction module (SCM). The second column shows outputs when the SCM is applied to the LDM output $I_{\rm sd}$, while the third column shows results when $I_{\rm sd}$ is set to zero. Despite no structural information (third column), the SCM still restores the image, indicating that it functions as a separate non-generative restoration model.