UGENIR: UNDERSTANDING-CONDITIONED DISCRETE GENERATION FOR OPEN-WORLD IMAGE RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

All-in-one image restoration is commonly formulated as a closed-set multi-task regression problem: a single model is trained to map a degraded input to a restored output in one shot. Such regressors often fail under *open-world* corruptions, where degradation types, severities, and especially their *compositions* shift beyond the training distribution, making pixel evidence ambiguous and leaving little room for revision. We propose **UGenIR**, which reframes restoration as *understanding-conditioned constrained generation*. Given a degraded image, UGenIR first infers an input-consistent semantic description that acts as a language-level prior to constrain global content. Restoration is then performed in discrete token space via MaskGIT-style *parallel token completion*, where low-confidence tokens are progressively re-masked and refined for efficient iterative recovery of fine details. To further improve robustness under distribution shift, we introduce a **critic-guided self-correction** loop: a critic is trained to predict spatial error maps, and at test time its feedback is converted into token masks to trigger targeted rollback and regeneration on likely-error regions. Experiments across diverse degradation settings demonstrate improved image quality scores, while ablations support the complementary benefits of semantic guidance and critic-guided self-correction.

1 INTRODUCTION

Single-image restoration aims to recover a visually plausible clean image from a degraded observation, *e.g.*, rain, haze, blur, or sensor noise. Despite rapid advances in CNN/Transformer restorers, restoration remains fundamentally ill-posed: multiple clean explanations can fit the same corrupted input, and strong priors are needed to avoid over-smoothing, texture averaging, or implausible artifacts (Zamir et al., 2021; 2022; Ulyanov et al., 2018; Blau & Michaeli, 2018).

A major practical challenge is that real-world degradations are *open-world*. Corruptions vary continuously in type and severity and are often *composed* (*e.g.* haze+rain+blur) with long-tailed combinations under distribution shift. To reduce deployment cost, recent *all-in-one* systems consolidate multiple degradations into a single network, using regression-based (Li et al., 2022; Potlapalli et al., 2023; Conde et al., 2024; Li et al., 2020; Valanarasu et al., 2022) or diffusion-based (Zheng et al., 2024; Jiang et al., 2024) formulations. While effective when test degradations match training coverage, one-shot regressors are brittle under unseen corruptions and, more critically, unseen *compositions*: ambiguous evidence can lead to averaged textures or hallucinated details, and a single forward pass provides limited mechanisms for self-checking or revision (fig. 1). Recent dataset efforts further highlight that OOD generalization remains a central bottleneck for all-in-one restoration (Rajagopalan et al., 2025).

Generative modeling offers a complementary view: instead of directly regressing pixels, restoration can be cast as *conditional generation* under a learned prior, naturally supporting iterative refinement (Kawar et al., 2022; Saharia et al., 2022). However, continuous-space diffusion inference is often expensive (Ho et al., 2020; Song et al., 2020; Lu et al., 2022), and its iterative process does not automatically resolve the key ambiguity of open-world compositions: when pixel evidence is unreliable, the model needs *explicit* constraints (Kawar et al., 2022; Chung et al., 2022) and *targeted* correction (Lugmayr et al., 2022; Avrahami et al., 2022) rather than unconstrained resampling.

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

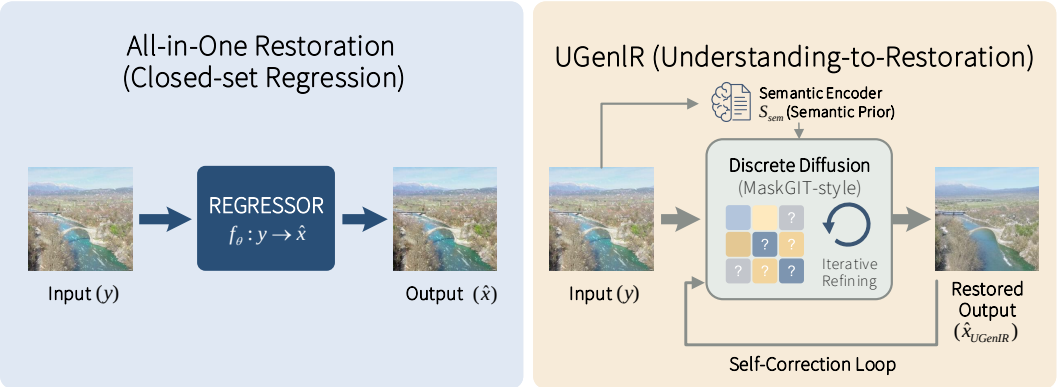

Figure 1: *Left:* Conventional all-in-one restoration formulates image restoration as closed-set regression from a degraded input y to an output $\hat{x}$, which often generalizes poorly to unseen or composite degradations. *Right:* UGenIR reframes restoration as understanding-conditioned generation: it first infers an input-consistent semantic description (language prior), then performs MaskGIT-style discrete diffusion via parallel token prediction with iterative re-masking of low-confidence tokens, guided by a self-correction loop (critic-guided) for more robust restoration under distribution shift.

Discrete token generation provides an attractive alternative. By encoding images into discrete latents, one can perform iterative masked token completion with *parallel* prediction, enabling efficient refinement with confidence-aware re-masking as in MaskGIT-style decoding (Chang et al., 2022; Yu et al., 2024).

Meanwhile, unified multimodal foundation models increasingly integrate *understanding* and *generation* within a single pretrained architecture. These models carry broad semantic priors from large-scale pretraining and can potentially improve robustness under distribution shift (Xie et al., 2025; Wang et al., 2025; Yang et al., 2025; Chameleon Team, 2024; Zhou et al., 2024; Chen et al., 2025a; Lin et al., 2025). Yet, directly repurposing them for low-level restoration is non-trivial: open-world restoration requires (i) conditioning on degraded-image tokens and a generated semantic description, (ii) efficient iterative decoding for high-frequency details, and (iii) a self-correction mechanism that localizes and fixes errors instead of relying on one-shot outputs.

We present **UGenIR**, which reframes restoration as an *understanding-to-generation* procedure (Fig. 2). Given a degraded image, UGenIR first produces an input-consistent semantic description that serves as a language-level prior to constrain global content when pixel evidence is ambiguous. Conditioned on this semantic signal, UGenIR performs restoration via **MaskGIT-style discrete token diffusion**: starting from masked tokens, it predicts missing tokens in parallel and progressively re-masks low-confidence positions to refine structures and textures efficiently. To further improve robustness under distribution shift, we introduce a **critic-guided self-correction loop**. A critic is trained with pixel-level error-map supervision; at test time, its predicted error regions are mapped to token masks to trigger targeted rollback and regeneration, enabling iterative, spatially grounded correction rather than unconditional resampling. We instantiate this procedure by fine-tuning a pretrained unified understanding–generation model. Experiments show improved image quality scores across diverse degradations and complementary benefits from semantic guidance and self-correction.

Overall, our contributions are:

- We propose **UGenIR**, a restoration paradigm that treats open-world restoration as *understanding-conditioned constrained discrete generation* rather than closed-set one-shot regression.
- We introduce **dual conditioning**—degraded-image tokens and an input-consistent semantic description—to condition restoration on both observed image evidence and language-level context.

- We design a **critic-guided self-correction** mechanism with error-map supervision and token-mask feedback, enabling spatially targeted iterative refinement.

2 THE PROPOSED METHOD

2.1 NOTATION AND PROBLEM SETUP

We denote a degraded image by $\mathbf{y} \in \mathbb{R}^{H\times W\times C}$ and its unknown clean counterpart by $\mathbf{x}^\star \in \mathbb{R}^{H\times W\times C}$. Our goal is to estimate a restored image $\hat{\mathbf{x}}$ that is both visually faithful and semantically consistent under *open-world* degradations, including unseen types, severities, and their compositions.

We work in a discrete token space induced by a visual tokenizer. Let $\mathrm{Tok}(\cdot)$ be the tokenizer encoder and $\mathrm{Dec}(\cdot)$ be the tokenizer decoder. The degraded image is mapped to a sequence (or grid-flattened sequence) of discrete tokens

$$\mathbf{X}_{\mathrm{LQ}} = \mathrm{Tok}(\mathbf{y}) \in \{1,\dots,V\}^L, \tag{1}$$

where V is the codebook size and L is the token length. Similarly, the clean image corresponds to a ground-truth token sequence

$$\mathbf{X}^\star = \mathrm{Tok}(\mathbf{x}^\star) \in \{1,\dots,V\}^L. \tag{2}$$

For text, we use a tokenized semantic description $\mathbf{s}^\star \in \{1,\dots,V_{\mathrm{text}}\}^T$ with vocabulary size V_{text} and length T.

We use $\mathcal{M} \subseteq \{1,\dots,L\}$ to denote a set of masked token indices, and $\mathbf{X}_{\mathcal{M}}$ to denote tokens restricted to indices in $\mathcal{M}$. The masking ratio is a scalar $\mu \in (0,1)$. We use lower-case italic symbols for scalars and bold symbols for image and token representations.

2.2 PRELIMINARIES: DISCRETE DENOISING VIA MASKED TOKEN MODELING

MaskGIT-style parallel token completion. MaskGIT trains a bidirectional Transformer to predict masked tokens in parallel. Given a ground-truth token sequence $\mathbf{X}^\star \in \{1,\dots,V\}^L$, we sample a mask set $\mathcal{M}$ and construct a masked sequence $\tilde{\mathbf{X}}$ by replacing positions in $\mathcal{M}$ with a special token $[\mathrm{MASK}]$. The model outputs a categorical distribution for each masked position, enabling parallel prediction.

Iterative inference (discrete diffusion). At inference time, masked denoising starts from an all-mask initialization

$$\mathbf{X}^{(0)} = [\mathrm{MASK}]^L, \tag{3}$$

and proceeds for $t = 1,\dots,T_{\mathrm{den}}$ iterations. At iteration t, the model predicts tokens for all masked positions in parallel and assigns confidence scores (*e.g.*, the probability of the sampled token). A noise schedule determines the number of tokens to *re-mask* for the next iteration by resetting the lowest-confidence subset to $[\mathrm{MASK}]$. This yields an efficient refinement process that explicitly localizes uncertainty in token space.

2.3 UGENIR: UNDERSTANDING-CONDITIONED CONSTRAINED GENERATION

UGenIR reframes restoration as *understanding-conditioned* constrained generation. Here, constraints are provided by image and text conditioning, rather than a hard pixel-wise data-consistency guarantee. We instantiate this procedure by fine-tuning a unified multimodal understanding–generation backbone. The method consists of two coupled components: implicit degradation understanding with semantic prior generation, and token-space restoration.

(i) Implicit Degradation Understanding & Semantic Generation. Our unified backbone directly processes the degraded-image tokens $\mathbf{X}_{\mathrm{LQ}} = \mathrm{Tok}(\mathbf{y})$, which represent the observed image content and its corruptions. Degradation understanding is learned implicitly from these tokens; there is no separate degradation-token sequence or dedicated degradation encoder. Conditioned on $\mathbf{X}_{\mathrm{LQ}}$,

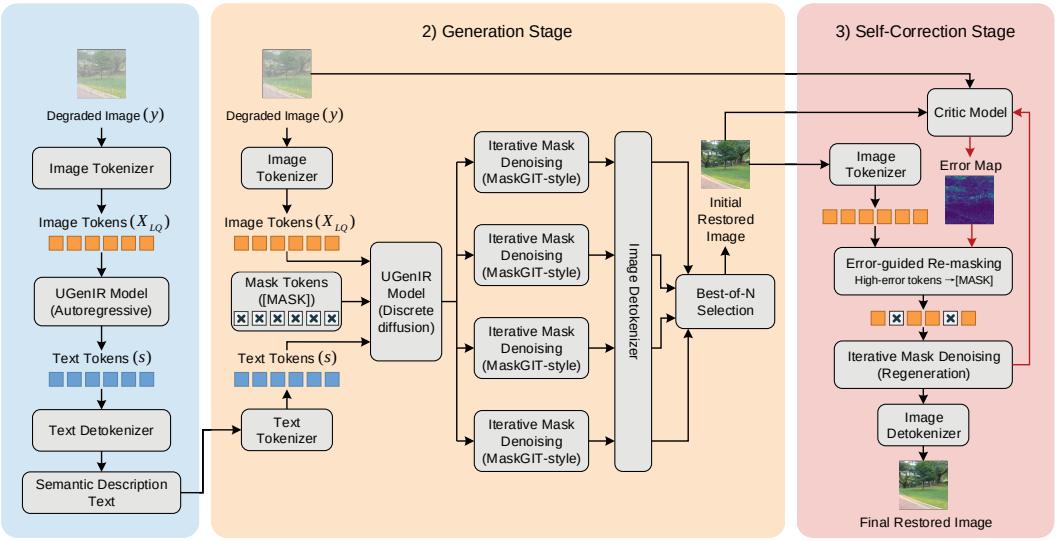

Figure 2: **Overview of the proposed UGenIR framework.** It integrates three distinct phases: a semantic understanding stage that converts degraded images into text descriptions; a generation stage that utilizes discrete diffusion for initial restoration; and a self-correction stage that employs a Critic Model to detect and refine artifacts via error-guided re-masking. Dual conditioning combines degraded-image tokens (visual evidence) with the generated semantic description (language-level guidance).

the model directly generates an input-consistent semantic description:

$$p_\theta(\mathbf{s} \mid \mathbf{X}_{\mathrm{LQ}}) = \prod_{t=1}^{T} p_\theta(s_t \mid \mathbf{s}_{<t}, \mathbf{X}_{\mathrm{LQ}}). \tag{4}$$

This resulting text acts as a high-level language prior to constrain global content when pixel evidence is unreliable.

(ii) Restoration as Constrained Discrete Denoising. Conditioned simultaneously on the degraded-image tokens $\mathbf{X}_{\mathrm{LQ}}$ and the generated semantic prior $\mathbf{s}$, we restore the clean token sequence via iterative masked denoising:

$$p_\theta(\mathbf{X}_{\mathcal{M}} \mid \tilde{\mathbf{X}}, \mathbf{X}_{\mathrm{LQ}}, \mathbf{s}) = \prod_{i \in \mathcal{M}} p_\theta(X_i \mid \tilde{\mathbf{X}}, \mathbf{X}_{\mathrm{LQ}}, \mathbf{s}), \tag{5}$$

where $\tilde{\mathbf{X}}$ denotes the current partially-masked sequence during inference and $\mathcal{M}$ is the current mask set. After denoising, the restored image is obtained by decoding:

$$\hat{\mathbf{x}} = \mathrm{Dec}(\hat{\mathbf{X}}). \tag{6}$$

2.4 Visual Tokenizer

We adopt MAGVIT-v2 to map continuous images into discrete tokens (Yu et al., 2024). Compared with lookup-based vector quantization (van den Oord et al., 2017; Esser et al., 2021), MAGVIT-v2 employs lookup-free quantization (Yu et al., 2024), enabling a large vocabulary size and improved representational expressiveness, which is crucial for preserving fine structures and textures needed in low-level restoration.

2.5 Training Objectives

UGenIR jointly optimizes understanding and restoration with a weighted multi-task objective.

Algorithm 1 UGenIR inference (self-correcting discrete restoration)

Require: degraded image $\mathbf{y}$; denoising steps T_{den}; candidates n; refinement rounds r

$\mathbf{X}_{\text{LQ}} \leftarrow \text{Tok}(\mathbf{y})$
$\mathbf{s} \sim p_\theta(\mathbf{s} \mid \mathbf{X}_{\text{LQ}})$
for $j = 1$ **to** n **do**
  $\hat{\mathbf{X}}^{(j)} \leftarrow \text{MaskDenoise}_\theta(\mathbf{X}_{\text{LQ}}, \mathbf{s}, T_{\text{den}})$
  $\hat{\mathbf{x}}^{(j)} \leftarrow \text{Dec}(\hat{\mathbf{X}}^{(j)})$
end for
$e_j \leftarrow \text{mean}(c_\phi(\hat{\mathbf{x}}^{(j)}, \mathbf{y}))$, $j = 1, \ldots, n$
$k \leftarrow \arg\min_j e_j$; $\hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}}^{(k)}$
for $u = 1$ **to** r **do**
  $\mathbf{E} \leftarrow c_\phi(\hat{\mathbf{x}}, \mathbf{y})$
  $\mathcal{M}_{\text{crit}} \leftarrow \text{Err2Mask}(\mathbf{E})$
  $\hat{\mathbf{X}} \leftarrow \text{Tok}(\hat{\mathbf{x}})$
  $\hat{\mathbf{X}}[\mathcal{M}_{\text{crit}}] \leftarrow [\text{MASK}]$
  $\hat{\mathbf{X}} \leftarrow \text{MaskDenoise}_\theta(\mathbf{X}_{\text{LQ}}, \mathbf{s}, T_{\text{den}}; \hat{\mathbf{X}})$
  $\hat{\mathbf{x}} \leftarrow \text{Dec}(\hat{\mathbf{X}})$
end for
return $\hat{\mathbf{x}}$

Semantic pseudo-labels. Most restoration datasets do not provide dense captions. We therefore generate semantic pseudo-labels $\mathbf{s}^\star$ using an external VLM conditioned on the degraded image $\mathbf{y}$, instructing it to describe both (i) image content and (ii) degradation effects (details in the supplementary material).

Understanding loss. We train the text decoder with standard autoregressive cross-entropy:

$$\mathcal{L}_{\text{under}} = -\sum_{t=1}^{T} \log p_\theta(s_t^\star \mid \mathbf{s}_{<t}^\star,\, \mathbf{X}_{\text{LQ}})\,. \tag{7}$$

Restoration loss (masked token modeling). Given the ground-truth clean token sequence $\mathbf{X}^\star \in \{1, \ldots, V\}^L$, we sample a mask set $\mathcal{M}$ with masking ratio μ, and construct a masked target sequence $\tilde{\mathbf{X}}$. The restoration objective minimizes the negative log-likelihood over masked indices:

$$\mathcal{L}_{\text{rec}} = -\sum_{i \in \mathcal{M}} \log p_\theta\Big(X_i^\star \mid \tilde{\mathbf{X}},\, \mathbf{X}_{\text{LQ}},\, \mathbf{s}^\star\Big)\,. \tag{8}$$

Text-to-image preservation loss. To retain the backbone's text-conditioned generation ability, we also include a masked image-token prediction loss conditioned only on text:

$$\mathcal{L}_{\text{t2i}} = -\sum_{i \in \mathcal{M}} \log p_\theta\Big(X_i^\star \mid \tilde{\mathbf{X}},\, \mathbf{s}^\star\Big)\,. \tag{9}$$

Overall objective. The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{under}} + \alpha\, \mathcal{L}_{\text{rec}} + \beta\, \mathcal{L}_{\text{t2i}}, \tag{10}$$

where α and β are scalar weights.

2.6 Critic-Guided Self-Correction

One-shot restoration is brittle under distribution shift because errors cannot be revised. We introduce a Swin-UNet critic that provides *spatially grounded* feedback and enables targeted token re-generation.

Critic training with error-map supervision. Given a restored image $\hat{\mathbf{x}}$ (decoded from tokens) and its ground-truth $\mathbf{x}^\star$, the critic predicts an error map

$$\mathbf{E} = c_\phi(\hat{\mathbf{x}}, \mathbf{y}) \in \mathbb{R}^{H \times W}, \tag{11}$$

supervised by a per-pixel error target (*e.g.*, $|\hat{\mathbf{x}} - \mathbf{x}^\star|$ aggregated over channels). We use a regression loss

$$\mathcal{L}_{\text{crit}} = \|\mathbf{E} - \mathbf{E}^\star\|_1 . \tag{12}$$

Token-mask feedback at test time. At inference, the critic predicts $\mathbf{E}$ for the current restoration. We convert $\mathbf{E}$ into a token-level mask set $\mathcal{M}_{\text{crit}}$ by pooling error within each token's receptive region and selecting the top-k high-error tokens. We then perform targeted rollback and re-generation by resetting indices in $\mathcal{M}_{\text{crit}}$ to $[\text{MASK}]$ and running additional masked denoising iterations, while keeping low-error tokens fixed.

2.7 Self-Correcting Inference: Generate–Verify–Refine

We formulate inference as a structured *generate–verify–refine* procedure rather than a single forward pass.

Multi-path generation. We sample n candidate restorations $\{\hat{\mathbf{x}}^{(j)}\}_{j=1}^{n}$ by running discrete denoising with different random seeds, yielding diverse hypotheses under ambiguity.

Verification and selection. We rank candidates using the Swin-UNet Critic. For each candidate $\hat{\mathbf{x}}^{(j)}$, we compute the mean of its predicted error map, $e_j = \text{mean}(c_\phi(\hat{\mathbf{x}}^{(j)}, \mathbf{y}))$, and select the candidate with the smallest e_j. If the Critic is unavailable, we instead select the candidate with the highest mean token log-probability recorded during generation. The Critic thus provides both candidate-level ranking and spatial feedback for subsequent refinement (section 3.2).

Critic-guided refinement. Starting from the selected candidate, we run the critic-guided remasking and targeted regeneration loop for r rounds (Algorithm 1).

3 Experiments

In this section, we provide a detailed analysis of the experimental results. Unless otherwise specified, the shorter side of each image is resized to 256, followed by a center crop of 256×256. We evaluate across the listed restoration benchmarks and composite-degradation subsets, and conduct ablations on the GenDS validation set.

3.1 Experimental Settings

Datasets. *Semantic Understanding:* Subsets of CC12M (Changpinyo et al., 2021) and InternVL-SA1B (OpenGVLab, 2024) supervise multimodal understanding and text-to-image generation. *Restoration & Image-Caption Pairs:* We use GenDS (Rajagopalan et al., 2025), which combines synthetic degradations with existing restoration datasets and contains over **750,000** images. Qwen2.5-VL-32B (Bai et al., 2025) generates pseudo-captions describing scene content and degradation effects. We manually reviewed the generated captions and removed those containing incorrect degradation labels or descriptions inconsistent with the visible image content.

Evaluation Metrics. Benchmark comparisons in Table 1 and the FoundIR supplementary table report distributional similarity and predicted image quality. FID (Heusel et al., 2017) compares feature distributions across image sets rather than individual paired images. NIQE (Mittal et al., 2013), MUSIQ (Ke et al., 2021), MANIQA (Yang et al., 2022), and CLIP-IQA (Wang et al., 2023) assess quality without a reference; we additionally report TOPIQ (Chen et al., 2024). For the GenDS ablation in Table 2, we also use PSNR for pixel-wise fidelity and LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2020) for reference-based perceptual similarity. Higher quality scores alone do not establish faithful recovery of the original content. Arrows indicate the preferred direction.

Table 1: Image quality and distributional comparison. Red, blue, and violet indicate the best, second-best, and third-best results; ties share a rank.

Methods	Distribution	Image Quality Scores				
	FID↓	NIQE↓	MUSIQ↑	MANIQA↑	CLIP-IQA↑	TOPIQ↑
***Dataset: REVIDE**(Zhang et al., 2021) (Haze)*						
UGenIR	**66.13**	6.9983	**51.16**	**0.4302**	**0.3362**	**0.5251**
PromptIR	72.19	7.5981	39.66	0.3270	0.2509	0.3464
InstructIR	68.80	7.0674	40.52	**0.3291**	0.2160	**0.3643**
AutoDIR	**69.58**	**6.6641**	41.40	0.3563	0.2873	0.3828
DiffUIR	72.61	7.6004	39.60	0.3237	**0.2602**	0.3446
AirNet	80.57	**7.0654**	**40.81**	0.3204	0.2372	0.3607
***Dataset: RainDS**(Quan et al., 2021) (RainDrop)*						
UGenIR	104.57	4.6466	**56.85**	**0.4203**	**0.5339**	**0.6596**
PromptIR	142.76	**3.8331**	50.27	**0.3363**	**0.3420**	**0.5156**
InstructIR	145.45	4.9243	**50.93**	0.3208	0.2840	0.4959
AutoDIR	**57.35**	4.8523	55.51	0.4059	0.4007	0.6231
DiffUIR	**133.84**	3.8242	50.05	0.3265	0.3358	0.5153
AirNet	150.42	**3.7069**	49.79	0.3259	0.3364	0.5085
***Dataset: RainDS**(Quan et al., 2021) (RainStreak)*						
UGenIR	**98.44**	4.9253	**56.73**	**0.4295**	**0.5366**	**0.6581**
PromptIR	102.67	5.1687	50.94	**0.3466**	**0.4019**	0.5712
InstructIR	**58.36**	**5.1225**	51.63	0.3577	0.3981	0.5783
AutoDIR	85.90	5.4738	**51.62**	0.3138	0.3777	0.5637
DiffUIR	122.17	5.3343	51.05	0.3331	0.4097	0.5587
AirNet	109.43	**4.9120**	51.16	0.3413	0.3864	**0.5729**
***Dataset: RainDS**(Quan et al., 2021) (Mixture)*						
UGenIR	**126.63**	4.5679	**54.49**	**0.3928**	**0.5073**	**0.6352**
PromptIR	201.42	**4.3134**	47.63	0.2837	**0.2983**	0.4406
InstructIR	169.18	4.5049	**48.70**	**0.2884**	0.2728	0.4610
AutoDIR	129.22	4.8414	51.65	0.3263	0.3488	0.5384
DiffUIR	**163.40**	**4.4454**	48.44	0.2854	0.2861	**0.4644**
AirNet	202.90	4.3317	47.33	0.2791	0.2908	0.4395
***Dataset: LOLv1**(Wei et al., 2018) (Low-light)*						
UGenIR	**136.73**	6.0583	**61.61**	**0.4605**	**0.4832**	**0.6996**
PromptIR	190.81	11.1114	32.52	0.2619	0.2587	0.2940
InstructIR	**71.60**	**5.8182**	53.20	0.3836	0.3791	0.5763
AutoDIR	74.47	**6.1292**	**52.05**	**0.3732**	**0.4085**	**0.5555**
DiffUIR	153.24	8.2269	33.37	0.2678	0.4044	0.3406
AirNet	159.93	8.2680	32.63	0.2605	0.4758	0.3342
***Dataset: RVSD**(Chen et al., 2023) (Snow)*						
UGenIR	35.23	5.0690	**45.23**	**0.3574**	**0.3227**	**0.4330**
PromptIR	47.97	4.1773	40.29	0.3106	0.2836	0.3868
InstructIR	57.37	5.0736	43.79	0.2954	0.2841	0.4040
AutoDIR	**47.74**	4.3374	40.69	0.3107	**0.2877**	0.3899
DiffUIR	**23.35**	**4.2373**	41.51	0.3166	0.2754	0.3962
AirNet	52.47	**4.1600**	**41.77**	**0.3153**	0.3085	**0.4011**
***Dataset: GoPro**(Nah et al., 2017) (Motion Blur)*						
UGenIR	50.33	**4.0917**	**45.53**	**0.3798**	**0.3083**	**0.4700**
PromptIR	45.04	5.0289	28.03	0.1786	0.1750	0.2445
InstructIR	**41.00**	4.9456	29.97	0.1956	**0.1956**	0.2630
AutoDIR	39.89	4.8176	35.84	0.2593	0.2071	0.3362
DiffUIR	**33.09**	4.9065	**30.20**	**0.2060**	0.1759	**0.2714**
AirNet	45.65	**4.9032**	28.32	0.1780	0.1951	0.2448

3.2 Implementation Details

Our UGenIR is built upon the unified multimodal model Show-o (Xie et al., 2025). During training, inputs are resized to 256×256. The model is optimized for 500K steps on 6 NVIDIA A100 GPUs using AdamW (initial learning rate 2×10^{-5}, weight decay 0.01, 5K-step warm-up, and cosine

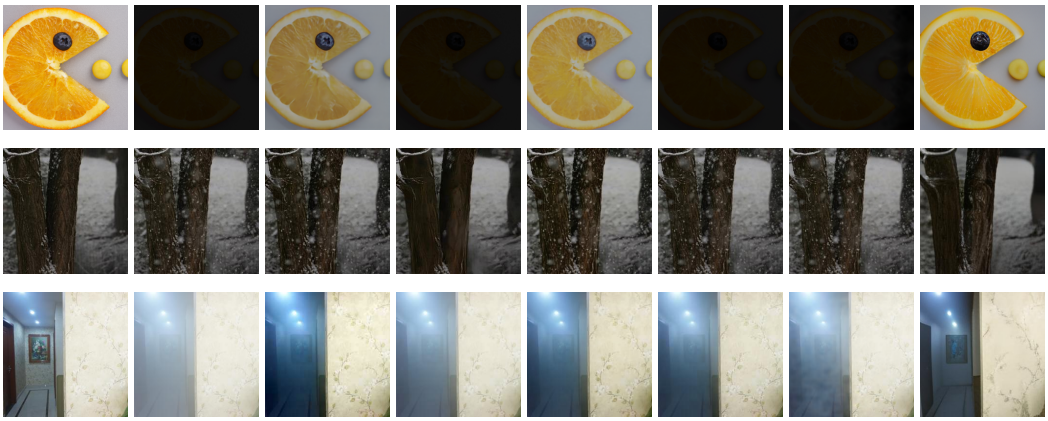

Figure 3: Qualitative comparisons on FoundIR (Li et al., 2025), RealSnow (Lai et al., 2025), and REVIDE (Zhang et al., 2021).

Table 2: Ablation of inference guidance strategies on the GenDS validation set. "Understanding Only" uses semantic guidance, and "SC" denotes self-correction. We report PSNR, LPIPS, DISTS, FID, and NIQE.

Method	PSNR↑	LPIPS↓	DISTS↓	FID↓	NIQE↓
No Guidance	19.31	0.2303	0.1623	28.13	5.6593
Understanding Only	19.54	0.2306	0.1648	28.64	5.7176
Self-Correction Only	19.60	0.2288	0.1616	27.87	5.6682
Understanding + SC	**20.08**	**0.2097**	**0.1513**	**26.54**	**5.4648**

annealing). We employ 6-step gradient accumulation to achieve an effective global batch size of 1,152. At inference, we use $n = 5$ candidates, $T_{\text{den}} = 18$ denoising steps per denoising call, and $r = 1$ refinement round. For candidate selection and self-correction, a Swin-UNet Critic is separately trained to predict pixel-level error maps. At inference, we generate image and text tokens via discrete diffusion and autoregressive sampling, respectively, and select the candidate with the lowest mean Critic-predicted error in a Best-of-N manner. When the Critic is unavailable, candidate selection falls back to maximizing the mean token log-probability.

3.3 Comparative Experiments

Quantitative Comparisons. We compare UGenIR with AirNet (Li et al., 2022), PromptIR (Potlapalli et al., 2023), InstructIR (Conde et al., 2024), DiffUIR (Zheng et al., 2024), and AutoDIR (Jiang et al., 2024), using pretrained baseline checkpoints without additional fine-tuning. This comparison does not isolate architecture from training-data coverage. In Table 1, UGenIR leads MUSIQ, MANIQA, CLIP-IQA, and TOPIQ across all seven settings, achieves the lowest NIQE on GoPro, and obtains the lowest FID on REVIDE and RainDS (Mixture). Its NIQE and FID advantages are not consistent across all settings. On the six FoundIR composite-degradation subsets in the supplementary material, UGenIR leads all five image quality scores but does not achieve the lowest FID. These results support improvements in predicted image quality; they do not establish uniformly better distributional matching or faithful recovery of the original image content.

Qualitative Comparison. Figures 3 and 4 compare results on FoundIR, RealSnow, REVIDE, and RealSnow10k. In the displayed examples, AirNet, PromptIR, and InstructIR leave residual corruptions, while AutoDIR exhibits incomplete removal or color shifts. DiffUIR suppresses corruptions but smooths fine textures around structures such as window frames and the red wall. UGenIR produces cleaner homogeneous regions and sharper boundaries in these examples. These visual comparisons complement the quantitative results; the reference-free RealSnow10k examples illustrate appearance rather than establishing agreement with an unknown clean image.



Figure 4: Qualitative comparisons on RealSnow10k (Lai et al., 2025) (real-world snow removal) without reference/ground truth.

3.4 Ablation Study

Table 2 evaluates semantic guidance and critic-guided self-correction (SC) on GenDS validation data. Understanding alone raises PSNR from 19.31 to 19.54 dB but slightly worsens LPIPS, DISTS, FID, and NIQE. SC alone improves PSNR, LPIPS, DISTS, and FID, while NIQE changes from 5.6593 to 5.6682. Combining both yields the best results across all five metrics, reaching **20.08 dB** PSNR and **0.2097** LPIPS. These results support the complementary benefit of semantic guidance and self-correction in this setting.

4 Conclusion

We presented **UGenIR**, an understanding-conditioned discrete generative framework for open-world image restoration. Degraded-image tokens and a generated semantic description condition MaskGIT-style recovery, while a Swin-UNet Critic guides candidate selection and targeted token regeneration. Experiments demonstrate improvements in image quality scores across diverse degradations, while GenDS ablations support the complementary benefits of semantic guidance and self-correction. These quality scores alone do not establish faithful recovery of the original image content.

AI USE STATEMENT

We used Qwen2.5-VL-32B to generate semantic pseudo-labels from degraded training images, as described in Section 3.1. OpenAI Codex was used only for language editing, translation, LaTeX formatting, and basic consistency checks during manuscript preparation. All research ideas, methodological design, experiments, analysis, and conclusions were developed and verified by the authors.

REFERENCES

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, pp. 18208–18218, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, pp. 6228–6237, 2018.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *CVPR*, pp. 11305–11315, 2022.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pp. 3558–3568, 2021.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE TIP*, 33:2404–2418, 2024.

Haoyu Chen, Jingjing Ren, Jinjin Gu, Hongtao Wu, Xuequan Lu, Haoming Cai, and Lei Zhu. Snow removal in video: A new dataset and a novel method. In *ICCV*, pp. 13165–13176. IEEE, 2023.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025b.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *ECCV*, 2024.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 44(5):2567–2581, 2020.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu. Autodir: Automatic all-in-one image restoration with latent diffusion. In *ECCV*, pp. 340–359, 2024.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *NeurIPS*, 2022.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pp. 5148–5157, 2021.

Jianyu Lai, Sixiang Chen, Yunlong Lin, Tian Ye, Yun Liu, Song Fei, Zhaohu Xing, Hongtao Wu, Weiming Wang, and Lei Zhu. Snowmaster: Comprehensive real-world image desnowing via mllm with multi-model feedback optimization. In *CVPR*, pp. 4302–4312, June 2025.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *CVPR*, 2022.

Hao Li, Xiang Chen, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Foundir: Unleashing million-scale training data to advance foundation models for image restoration. In *ICCV*, pp. 12626–12636, 2025.

Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *CVPR*, pp. 3175–3185, 2020.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, volume 35, pp. 5775–5787, 2022.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, pp. 3883–3891, 2017.

OpenGVLab. InternVL-SA-1B-Caption. `https://huggingface.co/datasets/OpenGVLab/InternVL-SA-1B-Caption`, 2024.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one blind image restoration, 2023.

Ruijie Quan, Xin Yu, Yuanzhi Liang, and Yi Yang. Removing raindrops and rain streaks in one go. In *CVPR*, pp. 9147–9156, 2021.

Sudarshan Rajagopalan, Nithin Gopalakrishnan Nair, Jay N Paranjape, and Vishal M Patel. Gendeg: Diffusion-based degradation synthesis for generalizable all-in-one image restoration. In *CVPR*, 2025.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality, 2023.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *CVPR*, 2022.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *CVPR*, pp. 9446–9454, 2018.

Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *CVPR*, pp. 2353–2363, 2022.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017.

Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, Yang Li, and Qing-Guo Chen. Ovis-U1 technical report. *arXiv preprint arXiv:2506.23044*, 2025. URL https://arxiv.org/abs/2506.23044.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, volume 37, pp. 2555–2563, 2023.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, 2022.

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement, 2018.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *ICLR*, 2025.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025.

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 *CVPR*, pp. 1191–1200, 2022.

Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion: Tokenizer is key to visual generation. In *ICLR*, 2024.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, Candace Ross, Adam Polyak, Russell Howes, Vasu Sharma, Puxin Xu, Hovhannes Tamoyan, Oron Ashual, Uriel Singer, Shang-Wen Li, Susan Zhang, Richard James, Gargi Ghosh, Yaniv Taigman, Maryam Fazel-Zarandi, Asli Celikyilmaz, Luke Zettlemoyer, and Armen Aghajanyan. Scaling autoregressive multi-modal models: Pretraining and instruction tuning, 2023.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, 2020.

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.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

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, 2021.

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 *CVPR*, pp. 25445–25455, 2024.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

A APPENDIX

A.1 RELATED WORK

A.1.1 ALL-IN-ONE IMAGE RESTORATION

Classic deep image restoration typically trains task- or degradation-specific models (e.g., denoising, deblurring, deraining, dehazing), often leveraging increasingly general-purpose backbones such as multi-scale attention networks, U-shaped Transformers/MLPs, or simplified convolutional baselines (Zamir et al., 2020; 2021; 2022; Wang et al., 2022; Tu et al., 2022; Chen et al., 2022). While these architectures can be applied across multiple restoration tasks, they are usually trained and evaluated under a relatively controlled setting where the degradation family is known.

To reduce deployment cost and handle diverse corruptions with a single model, *all-in-one* image restoration aims to recover images from multiple degradation types and levels without requiring explicit degradation labels at inference. A representative line of work treats this as a *closed-set multi-task regression* problem, where one network learns to map a degraded input to a restored output. AirNet proposes to learn a degradation-aware representation via a contrastive degraded encoder and uses it to guide a restoration network, enabling blind restoration across multiple corruption types (Li et al., 2022). PromptIR further introduces prompt-based conditioning, using learnable prompts to encode degradation information and dynamically steer the restoration backbone, improving flexibility across degradation varieties (Potlapalli et al., 2023).

Recent work highlights that the key bottleneck is not only handling multiple degradations, but *generalizing* to **OOD patterns and unseen compositions**. GenDeg tackles this challenge from the data side by using a conditional diffusion model to synthesize diverse degradations on clean images and constructing a large-scale dataset (GenDS), explicitly targeting generalizable all-in-one restoration (Rajagopalan et al., 2025). Despite these advances, most existing all-in-one systems remain regression-centric and one-shot at inference, which can be brittle when confronted with ambiguous evidence under distribution shift and OOD composition. Our work complements this literature by reframing restoration as understanding-conditioned *iterative* generation with explicit self-correction, aiming to improve robustness in OOD composition regimes.

A.1.2 UNIFIED UNDERSTANDING AND GENERATION MODELS

In parallel, the community has rapidly progressed toward *unified multimodal generative models* that support both multimodal understanding (e.g., VQA, captioning) and multimodal generation (e.g., text-to-image, interleaved generation) within a single architecture. Token-based autoregressive mixed-modal models provide a natural unification by representing images as discrete tokens and training a single decoder to model interleaved sequences. (Yu et al., 2023) is a representative decoder-only multimodal language model that can generate and infill both text and images, leveraging large-scale multimodal pretraining and supervised instruction tuning. (Sun et al., 2023) demonstrates generative multimodal pretraining over interleaved multimodal sequences, enabling both image-to-text and text-to-image capabilities in one model. (Chameleon Team, 2024) proposes early-fusion mixed-modal modeling that can understand and generate images and text in arbitrary order, targeting a single-model treatment of multimodal documents. (Wu et al., 2024; Chen et al., 2025b) further studies the conflict between understanding and generation when sharing visual encoding, and proposes decoupled visual encoding pathways while maintaining a unified transformer for mixed-modal processing.

Beyond fully autoregressive mixed-modal models, (Xie et al., 2025) explores a hybrid unification by combining autoregressive modeling for understanding with (discrete) diffusion-style generation in token space, supporting a wide range of understanding and generation tasks within one transformer. A key enabling component for these models is the discrete visual tokenizer, building upon vector-quantized representation learning and VQ-style image tokenization (van den Oord et al., 2017; Esser et al., 2021). (Yu et al., 2024) provides a modern tokenizer that produces concise discrete tokens shared across images and videos and has been adopted in unified multimodal generative systems.

While unified multimodal generative models offer strong semantic priors and broad capabilities, their use for *low-level restoration under open-world degradations* is still underexplored. Our work bridges these lines by fine-tuning a unified understanding–generation model for restoration, condi-

Table 3: Quantitative comparison on the **complex degradation subsets of the FoundIR dataset** (Li et al., 2025). We report FID and five image quality scores. The best, second-best, and third-best results are highlighted in **red**, blue, and **violet**, respectively. Ties at the reported precision share a rank, with subsequent ranks skipped.

Methods	Distribution	Image Quality Scores				
	FID↓	NIQE↓	MUSIQ↑	MANIQA↑	CLIP-IQA↑	TOPIQ↑
Dataset: Lowlight Haze						
AirNet	148.26	5.4209	30.22	0.2489	**0.4596**	0.2978
AutoDIR	131.69	4.5972	42.91	0.3017	0.4736	0.4413
DiffUIR	142.90	5.6535	31.60	0.2585	0.4121	0.3236
InstructIR	**110.43**	**4.6193**	**40.90**	**0.2716**	0.4174	**0.4241**
PromptIR	**142.27**	5.4607	30.93	0.2519	0.4302	0.3095
UGenIR	161.99	**4.3986**	**48.69**	**0.3406**	**0.6546**	**0.5785**
Dataset: Rain Haze						
AirNet	179.73	9.1886	53.35	0.3648	0.3766	0.5065
AutoDIR	196.26	8.3550	**55.71**	0.3523	**0.4081**	0.4907
DiffUIR	**158.02**	5.9016	53.81	0.3526	0.4034	0.5134
InstructIR	182.79	**8.1176**	56.30	**0.3739**	0.4629	**0.5148**
PromptIR	**179.41**	8.7920	54.66	0.3765	0.3845	0.5175
UGenIR	167.20	**5.3893**	**58.87**	**0.4350**	**0.5900**	**0.6292**
Dataset: Lowlight Blur						
AirNet	100.28	6.3104	32.04	0.2125	0.2970	0.3588
AutoDIR	77.65	**7.0769**	38.45	0.2445	0.3480	0.4216
DiffUIR	**86.19**	7.1336	34.16	**0.2244**	0.3002	0.3759
InstructIR	**72.78**	7.0829	**35.51**	0.2196	0.3095	**0.3778**
PromptIR	97.97	7.0862	32.74	0.2214	**0.3129**	0.3614
UGenIR	107.44	**4.8972**	**50.33**	**0.3673**	**0.5464**	**0.5964**
Dataset: Lowlight Noise						
AirNet	147.91	6.3305	40.79	0.3256	0.4151	0.4511
AutoDIR	**87.29**	5.3213	57.07	0.4178	0.4602	0.5984
DiffUIR	149.33	7.3605	41.59	0.3347	**0.4263**	0.4548
InstructIR	104.24	**5.4397**	**53.51**	**0.3817**	0.3814	**0.5649**
PromptIR	156.85	8.2456	41.43	0.3325	0.4093	0.4410
UGenIR	**146.06**	**5.1146**	**63.33**	**0.4647**	**0.4903**	**0.6933**
Dataset: Lowlight Blur Noise						
AirNet	106.96	5.8639	38.27	0.2175	0.3209	0.3634
AutoDIR	81.05	6.3313	45.66	0.2615	0.3752	0.4445
DiffUIR	**98.43**	6.3931	39.35	0.2293	0.3050	0.3799
InstructIR	**76.18**	**6.0562**	**43.77**	0.2281	0.3157	**0.4018**
PromptIR	105.27	6.4291	38.26	**0.2298**	**0.3228**	0.3728
UGenIR	125.11	**4.3159**	**53.08**	**0.3651**	**0.5068**	**0.5754**
Dataset: Lowlight Blur JPEG						
AirNet	96.30	7.2162	35.46	0.2452	0.4150	0.3722
AutoDIR	67.14	7.8862	40.99	0.2691	**0.3977**	0.4400
DiffUIR	**83.86**	8.6474	36.91	**0.2569**	0.3855	0.3948
InstructIR	**62.82**	**7.7979**	**37.92**	0.2439	0.3799	**0.4166**
PromptIR	86.54	8.2432	35.66	0.2510	0.3870	0.3806
UGenIR	122.35	**7.1458**	**50.88**	**0.3772**	**0.4450**	**0.6071**

tioning restoration on degraded-image tokens and a generated semantic description, together with critic-guided self-correction to improve robustness, especially on OOD degradation compositions.

A.2 Robustness Evaluation on Complex Composite Degradations

Table 3 evaluates six composite-degradation subsets of FoundIR. UGenIR leads NIQE, MUSIQ, MANIQA, CLIP-IQA, and TOPIQ in all six subsets. FID performance varies across subsets. These results support gains in predicted image quality, but do not establish faithful recovery of the original image content.

A.3 Details of Semantic Label Generation

We utilize Qwen-2.5-VL-32B to generate semantic descriptions used as pseudo-labels for the understanding module. To ensure the generated captions are helpful for the restoration task, we design the prompt to explicitly request information about both the content and the degradation.

Manual quality control. We manually reviewed the generated captions and removed those containing incorrect degradation labels or descriptions inconsistent with the visible image content.

Prompt Design. The exact system instruction fed to the multimodal model is as follows:

```
Instruction: Please describe the degradation type of the
image and briefly describe its effect on the visual quality.
Also provide a short description of the image content.

The output should be in plain text format and must not
include any Markdown syntax. The output should start with:
Degradation Type: <Type>

Then write 2 to 3 concise sentences about what is visible in
the image and how the degradation affects it.
```

By conditioning on this prompt, the resulting captions $\mathbf{s}^\star$ provide a rich linguistic prior that guides UGenIR to distinguish between the underlying clean content and the corruption patterns during the subsequent restoration process.

Examples of the generated captions are shown below:



Degradation Type: Rain Streaks

The image depicts a scene of a building labeled "Riviera," likely a casino or hotel, surrounded by palm trees under a clear blue sky. The presence of rain streaks across the image introduces a visual distortion, giving the impression of heavy rainfall.
This degradation affects the clarity and sharpness of the image, making details less distinct and creating a sense of motion or blur that detracts from the overall visual quality.
The image content shows a vibrant urban setting with tall buildings, billboards, and palm trees, suggesting a lively and iconic location, possibly in Las Vegas.



Degradation Type: Fog

The image depicts a large, ornate stone vase or urn situated on a grassy area near a paved pathway. The presence of fog significantly reduces visibility, creating a hazy atmosphere that obscures finer details of the vase and surrounding environment.

This foggy degradation diminishes the clarity and sharpness of the scene, making it difficult to discern intricate patterns or textures on the vase and blending the background elements into a uniform misty backdrop.