DUDA: Discovering Unseen Degradations to Adapt Open-World Image Restoration

Boseong Kim, Tae Hyun Kim, Donghyeon Cho*
Hanyang University, Seoul, Republic of Korea
{hosky789, taehyunkim, doncho}@hanyang.ac.kr

Abstract

While all-in-one image restoration models excel in controlled, closed-set environments, they face critical limitations when deployed in open-world settings where paired supervision is unavailable and degradations are unknown and often intricately mixed. To address this challenge, we propose a continual restoration framework that integrates degradation discovery with adaptive restoration. Building on a model pre-trained on known degradation categories, our method discovers unseen degradations in feature space and continually expands the degradation space. We then introduce a discovery- and instance-conditioned descriptor that jointly captures discovered degradation priors and image-specific characteristics, enabling robust restoration under ambiguous and intricately mixed degradations. To further support adaptation under unpaired supervision, we adopt a mean-teacher-based semi-supervised framework equipped with a reliable bank for pseudo-target refinement. Specifically, we propose a discovery-adaptive score that assesses pseudo-target reliability using feature-space distances to degradation categories, enabling pseudo-target selection to remain adaptive to newly discovered degradations rather than relying solely on conventional perceptual quality-based scoring methods. Experiments on both controlled open-world protocols and real-world adverse weather datasets demonstrate that our approach consistently outperforms existing methods across multiple image quality metrics while improving robustness to degradation composition shifts. Project page: https://bo-oseng.github.io/DUDA/.

1 Introduction

Image restoration aims to recover clean images from degraded observations and supports downstream vision systems such as autonomous driving, surveillance, and robotics. All-in-one image restoration (AIOIR) models improve deployability by handling multiple degradations in one network. Recent methods use visual prompts or instructions for degradation-aware conditioning [31, 5], and OneRestore [10] extends this approach to composite degradations. Yet most AIOIR methods assume a fixed degradation space defined during supervised training. In practice, such systems face open-world environments where weather, lighting, and sensor noise vary unpredictably. Consequently, real-world degradation patterns, severities, and compositions may differ from those seen during base supervised training, motivating adaptation beyond the closed-set setting in Fig. 1(a).

To bridge the gap between static training and dynamic real-world deployment, recent works on continual and semi-supervised restoration have explored more adaptive learning paradigms, as summarized in Fig. 1(b). Continual restoration methods mitigate forgetting across sequential weather tasks via distillation [25], whereas semi-supervised methods learn from unpaired real-world images using vision-language models [45]. Yet each addresses a different aspect of adaptation, leaving the joint setting of unlabeled and unpaired incoming images unresolved. Adaptation should therefore

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

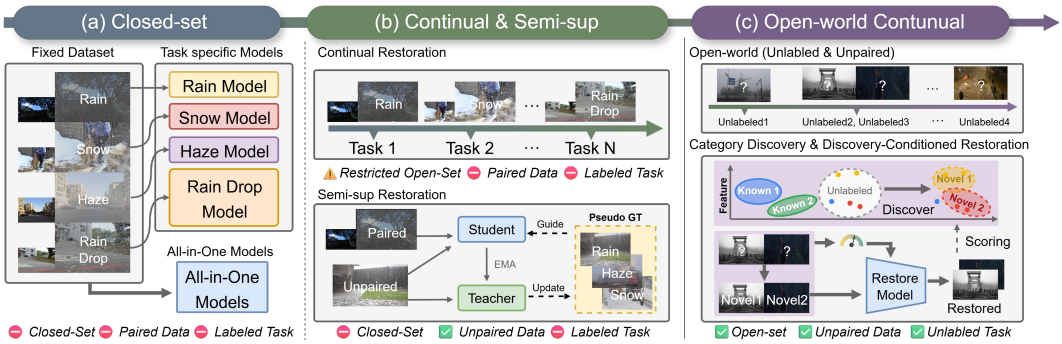

Figure 1: Comparison of image restoration paradigms. **(a) Closed-set:** Traditional restoration models assume a fixed degradation dataset during training and testing. **(b) Continual & Semi-supervised:** Existing methods adapt to sequential tasks or real-world data but still heavily rely on known task boundaries, explicit labels, or paired supervision. **(c) Open-world Continual:** Our proposed framework tackles unlabeled streams where unknown degradations continuously emerge.

discover emerging degradation structure and use it to guide restoration. Continual generalized category discovery (CGCD) [18] can identify known and novel degradation groups in unlabeled streams, but does not convert them into pixel-level restoration guidance for mixed degradations.

Motivated by this challenge, we formalize open-world continual restoration from unlabeled and unpaired streams containing degradation patterns not fully represented during base training, as illustrated in Fig. 1(c). In this setting, the model should continually discover emerging degradation structure while adapting the restoration network without paired clean references. Accordingly, degradation discovery is treated as a continual category discovery problem, whereas restoration is learned in a semi-supervised manner from unpaired real-world streams. Here, labeled/unlabeled denotes the availability of degradation-category annotations of the degraded input, whereas paired/unpaired denotes the availability of a clean ground-truth (GT) image corresponding to the input.

This setting presents three core challenges. First, *unlabeled degradations* make predefined degradation labels inapplicable. Second, *unpaired real-world streams* require semi-supervised restoration, as incoming degraded images do not have corresponding GT references. Finally, *degradation composition bias* causes the restoration model to systematically drift toward the initial training distribution of base degradations (e.g., globally brighter outputs if the base set was low-light dominated). To address these issues, we propose a framework that couples degradation discovery with descriptor-conditioned restoration. Built on a degradation feature space defined by known degradation categories, our method discovers emerging degradation clusters from incoming samples and updates their prototypes to generate discovery- and instance-conditioned descriptors. These descriptors steer the restoration of unknown corruptions without explicit labels. We further combine NR-IQA with a discovery-adaptive score to select reliable pseudo-targets, addressing scoring ambiguity and composition bias.

In summary, our contributions are threefold: (1) We introduce an open-world continual restoration setting where real-world degraded images are simultaneously unlabeled and unpaired. (2) We propose a discovery- and instance-conditioned descriptor that leverages discovered degradation information to enable adaptive restoration of unknown degradations. (3) We introduce a discovery-adaptive pseudo-target refinement strategy that adapts target selection to newly discovered degradation categories, reducing the bias of relying solely on NR-IQA for refinement. Even without semi-supervised adaptation, the proposed discovery- and instance-conditioned descriptor improves restoration performance over recent baselines trained under the same setting on real-world datasets. After semi-supervised adaptation with unpaired data, our method achieves performance competitive with WResVLM [45], a state-of-the-art semi-supervised method that relies on degradation-category labels.

2 Related Work

All-in-one Image Restoration. In low-level vision, recovering clear content from corrupted observations has been treated as isolated problems. Early deep learning was dominated by specialized archi-

tectures optimized for singular weather conditions or lighting artifacts, such as removing fog [12, 52], rain streaks [9, 3], or low light [19, 53]. However, deploying this fragmented approach in the wild imposes severe operational bottlenecks because changing environments require switching between a massive ensemble of task-specific networks. Driven by the necessity for scalable deployment in real-world systems, the community has increasingly shifted toward AIOIR. By processing diverse corruptions through shared weights [21, 2], AIOIR provides a foundation for versatile restoration.

Prompt-based Image Restoration. As unified models must map heavily diverse and sometimes conflicting corrupted inputs to a clean manifold, researchers quickly identified the need for dynamic network modulation. This motivated the emergence of mechanisms driven by prompts and explicit degradation awareness, injecting targeted context directly into the restoration pipeline. Works like PromptIR [31], InstructIR [5], and DCPT [13] guide the recovery process using visual queries, linguistic instructions, or classification pretraining. For composite degradations, OneRestore [10] aligns the restoration process with rich scene descriptors to remove mixed effects. CURE [17] extends this setting toward controllable restoration, enabling selective and intensity-adjustable removal of individual degradation factors within composite mixtures. However, these methods rely on static closed world training environments. Assuming all degradation types are known a priori causes them to fail against unknown real-world corruptions.

Continual and Semi-sup Restoration. To address closed-world constraints, recent studies introduce continual learning to restoration by sequentially learning new tasks without catastrophic forgetting. CLAIO [4] utilizes a knowledge replay buffer while ILAWR [25] employs distillation guided by degradation awareness to stabilize adaptation. To handle real-world scenarios where GT targets are unavailable, Semi-UIR [15] introduces a mean-teacher architecture to enable training on unpaired data. Building on this, WResVLM [45] leverages semantic priors from large vision-language models to guide the recovery process. Despite their progress, these methods typically assume structured streams with explicit labels or rely on a limited vocabulary of predefined texts. Such restrictive assumptions are mismatched with the unpredictable nature of completely open data streams.

Category Discovery. Generalized category discovery aims to identify known and novel visual categories from unlabeled data [39]. Moving beyond offline setups, continual extensions such as PACGCD [18] and VB-CGCD [8] address scenarios where known and novel semantic classes are encountered across learning stages. These methods typically estimate category structures by detecting novel clusters and updating their prototypes. However, they are primarily designed for discriminative image classification. How to translate the statistical properties of discovered clusters, including their means and covariance structures, into restoration guidance remains underexplored. Our work bridges this gap by coupling continual category discovery with a prompt-conditioned restorer. By leveraging discovered degradation clusters, we reduce the dependency on explicit labels, mitigate degradation bias, and enable continual adaptation.

3 Method

To restore images from unlabeled and unpaired streams in open-world continual settings, our framework builds on continual category discovery and semi-supervised image restoration. In this section, we briefly review the background on these two paradigms. We then introduce a discovery- and instance-conditioned descriptor that encodes degradation-aware and instance-specific guidance for an AIOIR model. Furthermore, for unpaired data without clean GT targets, we adopt a mean teacher-based semi-supervised framework [15] and propose a discovery-adaptive score to curate reliable pseudo-targets. An overview of the overall pipeline is illustrated in Fig. 2.

3.1 Preliminary

Continual Category Discovery. We adopt VB-CGCD [8] to model degradation features with a Gaussian mixture model (GMM), retaining previously learned clusters while expanding the mixture to accommodate newly discovered degradation patterns. Initialized from labeled base data, the cluster structure evolves with incoming unlabeled observations, representing not only new degradation types and compositions but also variations within familiar degradation families. Consequently, discovered clusters need not correspond one-to-one with human-defined degradation labels. We represent this evolving structure by C known and discovered clusters, each with a mean $\boldsymbol{\mu}_c$ serving as its prototype and a covariance $\boldsymbol{\Sigma}_c$ describing its spread and shape, for $c \in \{1, \ldots, C\}$. Note that labeled/unlabeled

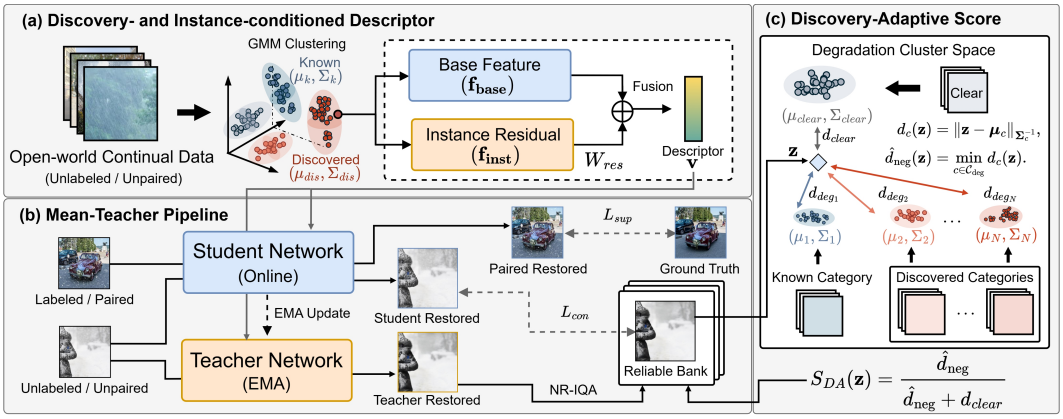

Figure 2: Overview of the proposed open-world continual image restoration framework. The **(a) Discovery- and Instance-conditioned Descriptor** module translates discovery statistics into targeted, instance-aware restoration guidance. The **(b) Mean-Teacher Pipeline** integrates these components with a reliable bank to enable robust semi-supervised learning on unpaired data streams. The **(c) Discovery-Adaptive Score** evaluates the reliability of candidate pseudo-targets based on their relative Mahalanobis distances to the anchored $clear$ cluster and the nearest degradation cluster.

denotes whether an input image has a degradation-category annotation. Separately, known/novel indicates whether a degradation category is available at the initial stage, and a novel category becomes discovered once identified as a cluster.

To describe the discovery process more concretely, we first initialize known clusters in the feature space using labeled base data. For each incoming unlabeled sample, we extract its feature representation $\mathbf{z}$ and evaluate its proximity to each cluster using the Mahalanobis distance:

$$d_c(\mathbf{z}) = \|\mathbf{z} - \boldsymbol{\mu}_c\|_{\boldsymbol{\Sigma}_c^{-1}} = \sqrt{(\mathbf{z} - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1} (\mathbf{z} - \boldsymbol{\mu}_c)}\,. \tag{1}$$

The corresponding Gaussian log-likelihood and hard assignment over the current C clusters are

$$\ell_c(\mathbf{z}) = -\frac{1}{2}\left(d_c^2(\mathbf{z}) + \log|\boldsymbol{\Sigma}_c| + D\log(2\pi)\right), \quad \hat{c} = \underset{k\in\{1,\dots,C\}}{\operatorname{argmax}}\ \ell_k(\mathbf{z})\,, \tag{2}$$

where D is the feature dimension and $\hat{c}$ denotes the assigned cluster index. To accommodate emerging degradation patterns, VB-CGCD additionally forms candidate clusters from incoming unlabeled features and incorporates them alongside previously learned clusters through variational inference. The number of incoming candidate clusters K can be specified by the update protocol or estimated through a separate model-selection procedure. In the specified-count setting, only K is provided, while the groups' semantic identities and sample assignments remain unknown. We use this setting for the main component analyses to isolate their effects from category-count estimation errors. For the estimated-count setting, we augment VB-CGCD with the Akaike information criterion (AIC) [1]. Appendix B.3 describes the AIC-based count-estimation procedure, while Appendix E.1 analyzes its effects on discovery and restoration.

Semi-Supervised Image Restoration. In the open-world image restoration setting, newly encountered samples are not only unlabeled with respect to degradation categories but also unpaired with clean GT images. To address the unpaired aspect of this setting, we adopt the mean-teacher-based semi-supervised framework of Semi-UIR [15], which learns from degraded inputs without paired clean GT images. The framework consists of a student network and a teacher network. The student parameters $\boldsymbol{\theta}_S$ are optimized through supervised learning on paired data and consistency-based learning on unpaired data. The teacher parameters $\boldsymbol{\theta}_T$ are then updated as an exponential moving average (EMA) of the student parameters:

$$\boldsymbol{\theta}_T \leftarrow \alpha\boldsymbol{\theta}_T + (1-\alpha)\boldsymbol{\theta}_S, \tag{3}$$

where $\alpha \in [0,1]$ is the smoothing coefficient controlling the EMA update. For each unpaired degraded input, the teacher network generates a restored image as an initial pseudo-target candidate.

Rather than directly adopting this prediction as a pseudo-target, the framework utilizes a reliable bank to selectively store only reliable pseudo-targets. The reliable bank evaluates each candidate using MUSIQ [16] and is updated with the teacher prediction only when it scores higher than both the current student prediction and the stored pseudo-target. The student network is jointly optimized using supervised learning on paired data and pseudo-target supervision on unpaired data, where the reliable bank provides pseudo-targets for guidance.

3.2 Discovery- and Instance-Conditioned Descriptor

To address all-in-one image restoration in open-world settings, where degradation types are unknown and clean GT pairs are unavailable, we use the VB-CGCD clusters introduced in Section 3.1. Following the scene-descriptor conditioning used in OneRestore [10], a straightforward approach would use the embedding or prototype of one discovered cluster as the descriptor for each input. However, this approach has several limitations. First, newly discovered degradations may be ambiguous, with multiple degradation factors coexisting within a single input. This makes one-to-one degradation embeddings, which assign each image to a single category-specific representation, insufficient. Second, degradation-level embeddings alone cannot capture instance-specific features, such as degradation severity or local image content, limiting adaptation to each input. To overcome these issues, we introduce a discovery- and instance-conditioned descriptor composed of a base degradation embedding and instance-specific features, as illustrated in Fig. 2(a). Specifically, for each unlabeled input, we extract its feature representation $\mathbf{z}$ using a pre-trained DINO [37] backbone fine-tuned for known degradation classification via adapters. We then compute its soft assignment to each cluster c by applying a softmax function to the log-likelihoods in Eq. (2), resulting in soft assignment weights $s \in \mathbb{R}^C$, where C denotes the total number of categories. To translate these statistical properties into a restoration guidance signal, we introduce learnable class embeddings $\mathbf{E} \in \mathbb{R}^{C\times D}$, initialized from the cluster means. We aggregate cluster-level embeddings using the resulting soft assignment weights:

$$s_c(\mathbf{z}) = \frac{\exp(\ell_c(\mathbf{z}))}{\sum_{j=1}^{C}\exp(\ell_j(\mathbf{z}))}, \quad \mathbf{f}_{base}(\mathbf{z}) = \sum_{c=1}^{C} s_c(\mathbf{z})\mathbf{E}_c, \tag{4}$$

where s_c is the soft assignment weight of input feature $\mathbf{z}$ to cluster c, $\mathbf{E}_c$ is the embedding of cluster c, and $\mathbf{f}_{base}$ is the aggregated base degradation embedding. This soft aggregation avoids forcing each input into a single degradation category, allowing the descriptor to capture ambiguous degradations through a combination of cluster-level embeddings. While $\mathbf{f}_{base}$ captures global degradation characteristics through the aggregation of cluster-level embeddings, it does not explicitly encode instance-specific variations as deviations from the cluster centers. To capture such variations, we model instance-specific residual features using covariance-normalized deviations from the cluster centers. Specifically, we compute normalized residual vectors for each cluster as follows:

$$\mathbf{r}_c(\mathbf{z}) = \mathbf{\Sigma}_c^{-1/2}(\mathbf{z} - \boldsymbol{\mu}_c), \quad \mathbf{f}_{inst}(\mathbf{z}) = \sum_{c=1}^{C} s_c(\mathbf{z})\mathbf{r}_c(\mathbf{z}), \tag{5}$$

where $\mathbf{r}_c$ represents the residual for cluster c standardized by its covariance to account for varying scales and correlations, and $\mathbf{f}_{inst}$ is the aggregated instance-level feature. To combine the global base feature and the instance-level feature in a stable and adaptive manner, we introduce an adaptive feature modulation mechanism based on linear projection as follows:

$$\mathbf{v}(\mathbf{z}) = \mathbf{f}_{base}(\mathbf{z}) + \mathbf{W}_{res}\mathbf{f}_{inst}(\mathbf{z}), \tag{6}$$

where $\mathbf{W}_{res}$ is a learnable linear projection that selectively filters the instance-specific information, and $\mathbf{v}$ denotes the resulting discovery- and instance-conditioned descriptor. Crucially, inspired by zero-initialized conditioning in recent generative models such as ControlNet [49], $\mathbf{W}_{res}$ is strictly initialized to zero. This conservative initialization encourages the network to rely on the stable $\mathbf{f}_{base}$ at the early stage of training, enabling a gradual incorporation of $\mathbf{f}_{inst}$ while preserving the representational stability of the base prototypes. Finally, we integrate the descriptor $\mathbf{v}$ into the restoration backbone via cross-attention, following the approach in OneRestore. Specifically, $\mathbf{v}$ serves as the query in place of textual embeddings, while intermediate image features are adaptively modulated by the discovered visual priors. This discovery-driven guidance enables the restoration backbone to exploit cluster priors even for degradations beyond the known categories.

3.3 Discovery-Adaptive Score

While the discovery- and instance-conditioned descriptor provides effective guidance for restoration, the reliability of pseudo-targets in unpaired settings remains a critical challenge. As described in Section 3.1, the reliable bank is updated by comparing each teacher prediction with the student prediction and stored bank target using NR-IQA. However, because this criterion primarily assesses aesthetic image quality, a high score does not necessarily indicate better degradation removal. Thus, pseudo-targets closer to the *clear* cluster and farther from degradation clusters may be preferable candidates from a restoration perspective, but can be discarded when selection relies solely on NR-IQA. To address this issue, as illustrated in Fig. 2(c), we propose a discovery-adaptive score for evaluating candidate pseudo-targets. First, we construct an anchored *clear* cluster from clean GT images in the known paired training set. Then, for each unpaired input, we extract the feature representation $\mathbf{z}$ from the teacher prediction and compare $\mathbf{z}$ against the anchored *clear* cluster as well as the known and discovered degradation clusters. Mathematically, the discovery-adaptive score is defined as follows:

$$\hat{d}_{\text{neg}}(\mathbf{z}) = \min_{c \in \mathcal{C}_{\text{deg}}} d_c(\mathbf{z}), \quad S_{DA}(\mathbf{z}) = \frac{\hat{d}_{\text{neg}}(\mathbf{z})}{\hat{d}_{\text{neg}}(\mathbf{z}) + d_{\text{clear}}(\mathbf{z})}, \tag{7}$$

where d_c and d_{clear} denote the Mahalanobis distances in Eq. (1) to the c-th degradation cluster and the anchored *clear* cluster, respectively, and $\mathcal{C}_{\text{deg}}$ denotes the set of identified degradation categories, including both known and novel categories. A larger S_{DA} indicates that the candidate pseudo-target is closer to the *clear* cluster and farther from its nearest degradation cluster. We use S_{DA} as an additional criterion for reliable bank updates, where the bank is initialized with degraded inputs and progressively updated with teacher predictions selected by MUSIQ or the discovery-adaptive score. To avoid using initial degraded inputs as pseudo-targets, samples are included in the semi-supervised objective only after their bank entries have been updated at least once. Detailed update procedures are provided in Algorithm 1 in the Appendix.

3.4 Training Objectives

We define the overall training objective as a combination of losses for labeled/paired data and unlabeled/unpaired data streams. At the initial stage of training, both the student and teacher networks are initialized with weights pre-trained on known degradation categories. For labeled/paired data, the supervised loss $\mathcal{L}_{sup}$ is formulated as:

$$\mathcal{L}_{sup} = \lambda_1 \mathcal{L}_1(\hat{\mathbf{y}}, \mathbf{y}) + \lambda_2 \mathcal{L}_2(\hat{\mathbf{y}}, \mathbf{y}) + \lambda_{per} \mathcal{L}_{per}(\hat{\mathbf{y}}, \mathbf{y}) + \lambda_{ce} \mathcal{L}_{ce}(\hat{\mathbf{u}}, c_{gt}), \tag{8}$$

where $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_{per}$, and $\mathcal{L}_{ce}$ denote the L_1, mean squared error (MSE), perceptual, and cross-entropy losses, respectively. Moreover, $\hat{\mathbf{y}}$ and $\mathbf{y}$ represent the student's prediction and the GT image, while $\hat{\mathbf{u}}$ denotes the log-likelihoods from Eq. (2), and $\lambda_1, \lambda_2, \lambda_{per}$, and λ_{ce} are fixed balancing weights. The combination of pixel-wise and perceptual losses ensures both accurate reconstruction and high-frequency texture restoration. Additionally, $\mathcal{L}_{ce}$ is applied on the labeled set to stabilize the descriptor representation during online discovery, providing a structural constraint via the GT labels c_{gt}. For unpaired data, we apply a consistency loss using the pseudo-targets $\mathbf{y}^B$ stored in the reliable bank. The total objective function is defined as:

$$\mathcal{L}_{con} = \mathcal{L}_1(\hat{\mathbf{y}}_u, \mathbf{y}^B), \quad \mathcal{L}_{total} = \mathcal{L}_{sup} + \omega_u \mathcal{L}_{con}, \tag{9}$$

where $\hat{\mathbf{y}}_u$ denotes the student's prediction for the unlabeled/unpaired input, and ω_u is an epoch-dependent weight that is gradually increased during training.

4 Experiments

4.1 Datasets

To evaluate the efficacy of our framework in open-world scenarios, we strictly adhere to the data collection and evaluation protocols established in WResVLM [45]. Since no standardized benchmark has been established for open-world restoration with unknown degradations and unavailable paired supervision, adopting this protocol enables direct comparison on real-world adverse weather evaluation. For labeled/paired supervised training, we use a comprehensive set of synthetic benchmarks

Table 1: Overall quantitative comparisons of adverse weather removal on real-world sets. Zero-shot baselines are trained on paired data and evaluated without semi-supervised adaptation. **Bold**, underline, and *italic* indicate the best, second-best, and third-best performance, respectively.

Method	MUSIQ ↑ / CLIP-IQA ↑	LIQE ↑ / Q-Align ↑
Zero-shot		
Restormer [47]	56.38 / 0.438	2.456 / 3.503
WeatherDiff [28]	53.35 / 0.396	2.173 / 3.352
MWDT [29]	54.62 / 0.366	2.230 / 3.326
PromptIR [31]	56.07 / 0.443	2.437 / 3.491
DA-CLIP [27]	55.59 / 0.412	2.438 / 3.480
OneRestore [10]	56.23 / 0.425	2.520 / 3.452
AdaIR [7]	55.33 / 0.422	2.329 / 3.370
MoCE-IR [46]	55.58 / 0.423	2.364 / 3.392
VLU-Net [48]	56.53 / 0.415	2.397 / 3.454
DFPIR [38]	55.60 / 0.406	2.406 / 3.369
Ours (Zero-shot)	56.64 / 0.442	2.602 / 3.515
Semi-sup		
WResVLM [45] (Non-blind Semi-supervised)	**59.34** / *0.456*	*2.640* / *3.574*
Ours (Blind Semi-supervised, AIC)	*58.38* / 0.489	2.681 / 3.586
Ours (Blind Semi-supervised)	58.79 / **0.494**	**2.702** / **3.593**



Figure 3: Qualitative comparisons with competing models on real-world datasets. Ours better removes fine rain streaks and subtle snowflakes in the highlighted regions.

as defined in the WResVLM framework to establish base restoration capabilities. Specifically, for deraining, the setup includes Outdoor-Rain [22], RainDrop [32], and SPA [42], while OTS [20] and Snow100K [24] are employed for dehazing and desnowing respectively. For unlabeled/unpaired adaptation, the real-world stream is composed of images reflecting diverse atmospheric degradations. This includes 2,318 hazy images from the URHI [20] dataset, along with manually curated internet rainy and snowy collections containing 2,433 and 2,018 images, respectively. By adopting this established data stream, we simulate realistic deployment conditions where GT targets are unavailable. For evaluation, we employ the real-world adverse weather datasets aligned with the established protocol. These include the RTTS [20] dataset with 4,322 haze images, DDN-SIRR [43] and the realistic subset of Real3000 [23] comprising a total of 2,320 rain images, and the Snow100K Realistic [24] dataset containing 1,329 snow images.

4.2 Comparative Analysis

Evaluation Metrics. Quantitative validation on real-world degraded images is challenging because GT images are unavailable. We evaluate restoration quality using four representative NR-IQA metrics: MUSIQ [16], CLIP-IQA [41], LIQE [50], and Q-Align [44]. For paired CDD-11 evaluation, we additionally report PSNR and CIEDE2000 [26] to assess reconstruction fidelity and color error,

Table 2: Ablation study of the proposed components and reliable bank update strategy.

Ablation	$\mathbf{f}_{base}$	$\mathbf{f}_{inst}$	S_{DA}	MUSIQ ↑	CLIP-IQA ↑	Δ (MUSIQ / CLIP-IQA)
Component	×	×	—	56.07	0.433	—
	✓	×	—	56.11	0.434	+0.04 / +0.001
	✓	✓	—	56.64	0.442	+0.57 / +0.009
Reliable Bank	×	×	✓	57.67	0.485	—
	✓	✓	×	58.24	0.491	—
	✓	✓	✓	**58.79**	**0.494**	**+0.55 / +0.003**

Table 3: Ablation across continual learning stages. Forgetting$_{\mathrm{L/Q}}$ denotes the average LIQE/Q-Align drop on previous categories.

Stage Setup	LIQE ↑ / Q-Align ↑	Forgetting$_{\mathrm{L/Q}}$ ↓
$Base + 3$	2.702 / 3.593	– / –
$Base + 1 + 2$	2.701 / 3.587	0.011 / 0.018
$Base + 1 + 1 + 1$	2.721 / 3.589	0.017 / 0.025
Avg.	2.708 / 3.590	0.014 / 0.022

Table 4: Quantitative evaluation on the paired CDD-11 benchmark under biased degradation distributions.

Dataset	S_{DA}	CIEDE2000 ↓	Q-Align ↑	PSNR ↑
Balanced	—	10.81	3.322	18.23
	✓	**9.87**	**3.372**	**18.76**
Low-light Biased	—	11.75	3.696	17.20
	✓	**10.71**	**3.780**	**17.59**
Haze Biased	—	9.53	3.094	**20.74**
	✓	**8.73**	**3.166**	20.57

respectively. These metrics allow evaluation under both unpaired real-world settings and paired CDD-11 settings.

Baselines and Evaluation Protocols. We compare the performance of our framework against an extensive range of representative restoration models to demonstrate its competitive advantages. The selected baselines include a general restoration architecture, Restormer [47], and representative all-in-one restoration methods, including WeatherDiff [28], MWDT [29], PromptIR [31], DA-CLIP [27], OneRestore [10], AdaIR [7], MoCE-IR [46], VLU-Net [48], and DFPIR [38]. Our evaluation considers a more challenging unlabeled/unpaired open-world setting. Following the WResVLM dataset protocol, we evaluate all methods on the same real-world rain, haze, and snow benchmarks used in WResVLM [45]. Unlike WResVLM, which relies on degradation labels and textual descriptions during training, our framework discovers degradation categories directly from unlabeled real-world inputs. To further clarify the performance gains, we evaluate three comparison setups on the same real-world benchmark. First, *Zero-shot* compares our descriptor-conditioned model with baselines trained on paired data, without adaptation to unpaired real-world data. Next, *Non-blind Semi-supervised* denotes the labeled/unpaired setting of WResVLM [45], where the category of each unpaired real-world sample is known. Lastly, *Blind Semi-supervised* represents our open-world setting, where the model adapts to unlabeled/unpaired real-world data without using degradation labels or paired clean GT images. Since VB-CGCD can operate with either a specified or automatically estimated number of discovered categories, we report both variants. For the main analysis and ablation studies, we use the specified-category variant to isolate the effect of each component without confounding factors from category-count estimation. The automatic variant, denoted as *Blind Semi-supervised, AIC*, estimates this number using AIC, with details provided in Appendix B.3. To the best of our knowledge, this is among the first attempts to address semi-supervised all-in-one image restoration without degradation category information.

Quantitative Results. As summarized in Table 1, *Zero-shot*, which performs inference without semi-supervised learning, is competitive with representative restoration baselines on several metrics. This suggests that the proposed descriptor provides useful guidance before unpaired adaptation. With *Blind Semi-supervised* learning, our model outperforms the labeled/unpaired WResVLM on three metrics while remaining competitive on MUSIQ, despite operating under the more challenging setting. Compared with the specified-count variant, *Blind Semi-supervised, AIC* shows slightly lower restoration performance, while top-1 category accuracy falls from 94.72% to 81.37%. Nevertheless, the AIC variant remains competitive with the specified-count variant and also surpasses WResVLM on all metrics except MUSIQ. Appendix E.1 analyzes the relationship between category-count estimation, discovery accuracy, and restoration quality. Overall, these results show that our framework remains competitive in real-world restoration without degradation-category labels or paired clean targets for the incoming stream.

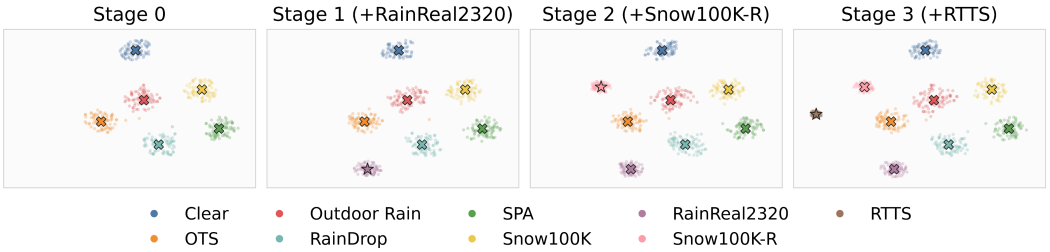

Figure 4: t-SNE visualization of the degradation cluster space under the *Base + 1 + 1 + 1* scenario. Stage 0 shows the base clusters, followed by the sequential introduction of real-world rain, snow, and haze in Stages 1–3. Cross markers denote existing cluster anchors, while star markers indicate newly introduced clusters. Newly introduced groups form distinct clusters alongside existing ones.

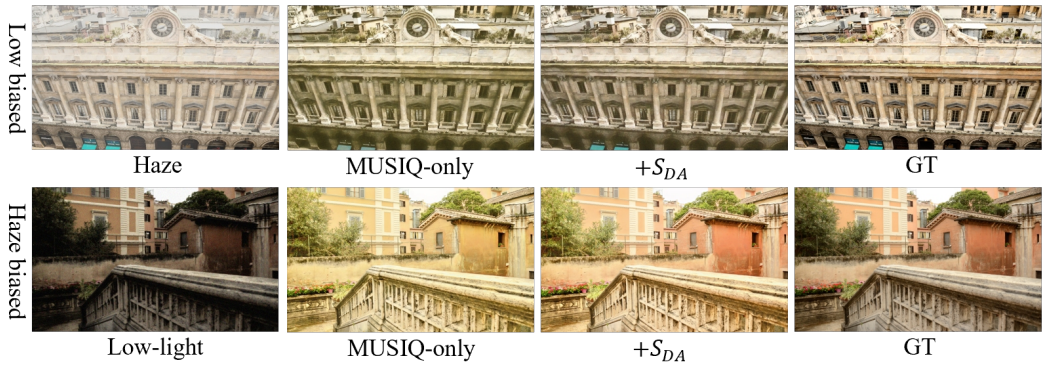

Figure 5: Qualitative comparison on CDD-11 under biased degradation distributions. Models are pre-trained on biased subsets and undergo unpaired semi-supervised adaptation to degradation categories excluded from base training. From left to right, the degraded input, the result after reliable-bank updates based on MUSIQ alone, the result after updates based on both MUSIQ and S_{DA} (Ours), and the GT image.

Qualitative comparison. As shown in Fig. 3, our method achieves visually faithful restoration compared with competing methods on real-world degraded images. Across rain, haze, and snow examples, our method effectively removes rain streaks, alleviates haze without introducing overly dark distortions, and suppresses subtle snowflakes more thoroughly. In particular, the highlighted regions show that competing methods often leave fine rain streaks or snowflakes unremoved, whereas our method removes these degradation-specific artifacts more effectively. Additional qualitative results are provided in Appendix D.1.

4.3 Ablation Studies

Effectiveness of Proposed Components. We conduct an ablation study to examine the contribution of the base degradation embedding $\mathbf{f}_{base}$, the instance-level feature $\mathbf{f}_{inst}$, and the discovery-adaptive score S_{DA} in our framework. As shown in Table 2, adding $\mathbf{f}_{base}$ alone yields a modest gain, while further incorporating $\mathbf{f}_{inst}$ produces larger improvements in the *Zero-shot* setting. With the same reliable-bank update rule incorporating S_{DA}, adding the descriptor raises MUSIQ from 57.67 to 58.79 and CLIP-IQA from 0.485 to 0.494. With the descriptor fixed, adding S_{DA} to MUSIQ-only bank selection raises MUSIQ from 58.24 to 58.79 and CLIP-IQA from 0.491 to 0.494. These matched comparisons show that the descriptor and discovery-adaptive score make distinct contributions during semi-supervised adaptation.

Robustness to Continual Learning Stages. In the open-world setting, the model may receive novel degradation categories over sequential learning stages rather than observing all of them at once. To evaluate robustness across continual learning stages, we compare $Base + 3$, $Base + 1 + 2$,

Table 5: Downstream object-detection results and mean MUSIQ gains on paired adverse-weather datasets. The detector weights are fixed, and only the input images differ between the degraded and restored conditions. AP_{50} values are percentages, and AP50 gains are reported in percentage points. MUSIQ gains are averaged over the same degraded/restored image pairs for each dataset and are therefore the same for both detectors.

		AP_{50}			
Detector	Dataset	Degraded	Restored	Δ	MUSIQ Δ
Faster R-CNN [34]	Foggy Cityscapes-DBF [36]	44.12	**45.96**	**+1.84**	**+4.96**
	RainCityscapes [14]	44.01	**46.15**	**+2.14**	**+2.40**
Mask R-CNN [11]	Foggy Cityscapes-DBF [36]	41.73	**43.71**	**+1.98**	**+4.96**
	RainCityscapes [14]	38.74	**40.62**	**+1.88**	**+2.40**

and $Base + 1 + 1 + 1$, which introduce real-world rain, snow, and haze together, in two stages, and one by one, respectively. As reported in Table 3, our method maintains comparable LIQE/Q-Align performance across these scenarios with low forgetting. Here, forgetting denotes the drop on previously introduced groups from their best previous scores after introduction, averaged over applicable stages. Additional stage-wise results are provided in Appendix D.2. We further visualize the evolving degradation cluster space under the *Base + 1 + 1 + 1* scenario in Fig. 4. Starting from the supervised base clusters at Stage 0, the model progressively incorporates real-world rain, snow, and haze in Stages 1–3. Each incoming group forms a distinct cluster alongside the base clusters, while previously introduced groups remain visually separated as subsequent groups arrive. These observations provide qualitative support for the progressive expansion of the discovery space to accommodate new real-world degradation distributions while retaining the previously discovered cluster structure.

Synthetic Dataset Evaluation. We evaluate on the synthetic CDD-11 [10] benchmark to examine robustness under biased degradation distributions. Unlike the real-world benchmarks, CDD-11 provides paired GT images, enabling reference-based evaluation. We construct biased subsets by skewing the training data toward specific degradations, such as haze or low-light conditions, and evaluate on degradation classes excluded from paired training. As reported in Table 4, incorporating S_{DA} consistently reduces CIEDE2000 [26] and improves Q-Align across all settings, with a PSNR trade-off only in the haze-biased setting. The qualitative comparisons in Fig. 5 further illustrate the degradation composition bias discussed in the introduction. With MUSIQ-only reliable-bank updates, restored images exhibit color shifts relative to GT, whereas incorporating S_{DA} produces colors closer to GT in these examples. Together with the lower CIEDE2000 errors, these observations suggest that incorporating S_{DA} into reliable-bank updates helps mitigate degradation composition bias compared with relying solely on MUSIQ. Detailed subset configurations and additional quantitative results are provided in Appendices C.3 and D.3, respectively.

Downstream Object Detection. We further evaluate degraded and restored images from Foggy Cityscapes-DBF [36] and RainCityscapes [14] using Faster R-CNN [34] and Mask R-CNN [11] with Cityscapes-trained R50-FPN backbones [6]. The official detector checkpoints remain fixed without fine-tuning or weather-domain adaptation, and each degraded/restored image pair is evaluated using the same GT boxes. As shown in Table 5, restoration improves AP_{50} across both detectors and weather conditions, while mean MUSIQ also increases on the same image pairs. These results show concurrent improvements in image quality and downstream detection under the evaluated conditions.

5 Conclusion

In this paper, we introduced an open-world continual image restoration framework for unpaired data with unknown degradations. Our method couples degradation discovery with a discovery- and instance-conditioned descriptor, enabling restoration guidance without degradation category labels. We further proposed a discovery-adaptive score to complement NR-IQA for reliable bank updates under degradation composition bias. Experiments on real-world adverse weather benchmarks demonstrate competitive restoration performance without paired clean targets. A current limitation is that cumulative replay increases training cost as more real-world data are accumulated. Future work will explore replay-free adaptation with LoRA-based updates for scalable continual restoration.

Acknowledgements

This work was partly supported by Korea planning & Evaluation institute of Industrial Technology(KEIT) grant funded by the Korea government(MOTIE) (No.RS-2026-25545342, Development of Dataset Construction and Virtual Simulation-Based Verification Technology for VLA Model Training and Validation, 25 %), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2025-25422680, Metacognitive AGI Framework and its Applications, 25 %), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University), 25 %) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-00521432, 25 %).

References

[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE TAC*, 19(6):716–723, 1974.

[2] Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, and Sy-Yen Kuo. Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In *CVPR*, pages 17653–17662, 2022.

[3] Xiang Chen, Jinshan Pan, and Jiangxin Dong. Bidirectional multi-scale implicit neural representations for image deraining. In *CVPR*, pages 25627–25636, 2024.

[4] De Cheng, Yanling Ji, Dong Gong, Yan Li, Nannan Wang, Junwei Han, and Dingwen Zhang. Continual all-in-one adverse weather removal with knowledge replay on a unified network structure. *IEEE TMM*, 26:8184–8196, 2024.

[5] Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *ECCV*, pages 1–21, 2024.

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[7] 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. In *ICLR*, pages 57335–57356, 2025.

[8] Hao Dai and Jagmohan Chauhan. Continual generalized category discovery: Learning and forgetting from a bayesian perspective. In *ICML*, pages 11884–11903, 2025.

[9] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *CVPR*, pages 3855–3863, 2017.

[10] Yu Guo, Yuan Gao, Yuxu Lu, Huilin Zhu, Ryan Wen Liu, and Shengfeng He. Onerestore: A universal restoration framework for composite degradation. In *ECCV*, pages 255–272, 2024.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[12] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. In *CVPR*, pages 1956–1963, 2009.

[13] JiaKui Hu, Lujia Jin, Zhengjian Yao, and Yanye Lu. Universal image restoration pre-training via degradation classification. In *ICLR*, 2025.

[14] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8014–8023. IEEE, 2019.

[15] Shirui Huang, Keyan Wang, Huan Liu, Jun Chen, and Yunsong Li. Contrastive semi-supervised learning for underwater image restoration via reliable bank. In *CVPR*, pages 18145–18155, 2023.

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

[17] Boseong Kim and Donghyeon Cho. Cure: Controllable unified image restoration for complex degradations. In *International Conference on Pattern Recognition*, pages 135–150. Springer, 2026.

[18] Hyungmin Kim, Sungho Suh, Daehwan Kim, Daun Jeong, Hansang Cho, and Junmo Kim. Proxy anchor-based unsupervised learning for continuous generalized category discovery. In *ICCV*, pages 16688–16697, 2023.

[19] Edwin H. Land. The retinex theory of color vision. *Scientific American*, 237(6):108–129, 1977.

[20] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE TIP*, 28(1):492–505, 2018.

[21] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *CVPR*, pages 17452–17462, 2022.

[22] Ruoteng Li, Loong-Fah Cheong, and Robby T Tan. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *CVPR*, pages 1633–1642, 2019.

[23] Yang Liu, Ziyu Yue, Jinshan Pan, and Zhixun Su. Unpaired learning for deep image deraining with rain direction regularizer. In *ICCV*, pages 4753–4761, 2021.

[24] Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *IEEE TIP*, 27(6):3064–3073, 2018.

[25] Xin Lu, Jie Xiao, Yurui Zhu, and Xueyang Fu. Continuous adverse weather removal via degradation-aware distillation. In *CVPR*, pages 28113–28123, 2025.

[26] M Ronnier Luo, Guihua Cui, and Bryan Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application*, 26(5):340–350, 2001.

[27] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for universal image restoration. In *ICLR*, 2024.

[28] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE TPAMI*, 45(10):12403–12415, 2023.

[29] Prashant W Patil, Sunil Gupta, Santu Rana, Svetha Venkatesh, and Subrahmanyam Murala. Multi-weather image restoration via domain translation. In *ICCV*, pages 21696–21705, 2023.

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[31] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *NeurIPS*, 36:71275–71293, 2023.

[32] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *CVPR*, pages 2482–2491, 2018.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[35] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420, 2007.

[36] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *European conference on computer vision*, pages 707–724. Springer, 2018.

[37] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[38] Xiangpeng Tian, Xiangyu Liao, Xiao Liu, Meng Li, and Chao Ren. Degradation-aware feature perturbation for all-in-one image restoration. In *CVPR*, pages 28165–28175, 2025.

[39] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, pages 7492–7501, 2022.

[40] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080, 2009.

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

[42] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *CVPR*, pages 12270–12279, 2019.

[43] Wei Wei, Deyu Meng, Qian Zhao, Zongben Xu, and Ying Wu. Semi-supervised transfer learning for image rain removal. In *CVPR*, pages 3877–3886, 2019.

[44] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *ICML*, 2024.

[45] Jiaqi Xu, Mengyang Wu, Xiaowei Hu, Chi-Wing Fu, Qi Dou, and Pheng-Ann Heng. Towards real-world adverse weather image restoration: Enhancing clearness and semantics with vision-language models. In *ECCV*, pages 147–164, 2024.

[46] Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration. In *CVPR*, pages 12753–12763, 2025.

[47] 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*, pages 5728–5739, 2022.

[48] Haijin Zeng, Xiangming Wang, Yongyong Chen, Jingyong Su, and Jie Liu. Vision-language gradient descent-driven all-in-one deep unfolding networks. In *CVPR*, pages 7524–7533, 2025.

[49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023.

[50] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *CVPR*, 2023.

[51] Ying Zhao and George Karypis. Criterion functions for document clustering: Experiments and analysis. 2001.

[52] Yu Zheng, Jiahui Zhan, Shengfeng He, Junyu Dong, and Yong Du. Curricular contrastive regularization for physics-aware single image dehazing. In *CVPR*, pages 5785–5794, 2023.

[53] Mingliang Zhou, Xingtai Wu, Xuekai Wei, Tao Xiang, Bin Fang, and Sam Kwong. Low-light enhancement method based on a retinex model for structure preservation. *IEEE TMM*, 26:650–662, 2023.

[54] Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, and Timothy Large. Image restoration for under-display camera. In *CVPR*, pages 9179–9188, 2021.

A Appendix Overview

This appendix provides implementation details, experimental protocols, additional results, and analyses supporting the main paper. We first describe implementation details in Section B, including the training settings, reliable bank update procedure, and estimation of the number of discovered categories using AIC. We then provide additional experimental details in Section C, including the evaluation settings for text-conditioned baselines, continual training scenarios, and the construction of CDD-11 [10] biased subsets. Detailed quantitative and additional qualitative results are presented in Section D, covering real-world benchmarks, various continual training scenarios, CDD-11, and unseen degradations. Finally, Section E examines the relationship between category discovery and restoration quality through AIC-based count estimation, cluster structure analysis, and a specified-K oracle comparison.

B Implementation Details

B.1 Training Settings

We select OneRestore [10] as our restoration backbone due to its lightweight 5.98M-parameter design and prompt-conditioned structure, which naturally accommodates our descriptor as guidance. Following VB-CGCD [8], we construct the discovery feature space by parameter-efficiently tuning adapters on a pre-trained DINO [37] backbone using known base degradation categories. VB-CGCD is then applied in this feature space to discover degradation categories during semi-supervised adaptation. Our training setup for the restoration model is divided into a supervised base phase for known degradation categories and a semi-supervised adaptation phase for newly discovered categories. During the supervised known phase, the model is trained on synthetic paired data using only the L_1 loss with a weight of 1.0. We train this phase for 500 epochs with a batch size of 128 and 224×224 patches. The model is optimized using AdamW with a learning rate of 1×10^{-4}, a weight decay of 1×10^{-4}, and a cosine annealing scheduler without warm-up. This supervised phase takes approximately 22 hours on three NVIDIA A100 GPUs.

The semi-supervised adaptation phase starts from the final checkpoint of the supervised base phase and optimizes the semi-supervised objective in Eq. 9. This objective combines supervised learning on labeled synthetic data with consistency-based training on reliable-bank pseudo-targets. Following the official implementation of Semi-UIR [15], we combine the L_1 and MSE losses with weights of 0.8 and 0.2, respectively. The perceptual loss $\mathcal{L}_{per}$ is computed as the MSE between VGG16 features of the restored image and its target. For this phase, we adopt a mean-teacher architecture in which both the student and teacher networks are initialized from the final supervised checkpoint. The teacher network is updated using an exponential moving average (EMA) of the student parameters with a momentum coefficient of 0.996. This coefficient is dynamically adjusted during the initial training steps to ensure stable convergence of the teacher network. During semi-supervised adaptation, unlabeled real-world images are divided into non-overlapping 224×224 patches. We use a total batch size of 56, the largest stable batch size under the memory constraints, comprising both labeled synthetic samples and unlabeled real-world patches. The network is optimized using AdamW with a weight decay of 1×10^{-4} and momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. This is coupled with a cosine annealing learning rate scheduler operating between a maximum learning rate of 1×10^{-4} and a minimum learning rate of 1×10^{-6}. To ensure stable training dynamics, we apply a linear ramp-up period spanning the first 5 epochs. The entire adaptation process is conducted over 10 epochs, which takes approximately 20 hours to complete.

B.2 Reliable Bank Update

Algorithm 1 summarizes the reliable bank update procedure used in the semi-supervised adaptation phase. The update constructs a union reliable bank by combining MUSIQ with the discovery-adaptive score in Eq. 7. For each sample, the teacher prediction is accepted as a new bank target if it improves over both the student prediction and the current bank target under either criterion. For computational efficiency, we implement the union update sequentially, applying the MUSIQ-based criterion first and evaluating S_{DA} only for the remaining candidates (reversing the order yields the same final selected set). The selected indices are used to update the bank targets y^B. The active mask $\mathbf{m}$ is initialized with $m_i = 0$ for all samples before adaptation and is set to $m_i = 1$ once the corresponding bank

Algorithm 1 Reliable bank update with MUSIQ and the discovery-adaptive score

Require: teacher prediction y^T, student prediction y^S, current bank target y^B
Require: active mask $\mathbf{m} \in \{0,1\}^N$, initialized as $m_i = 0$ for all samples before adaptation
Require: MUSIQ score $s_{\mathrm{M}}(\cdot)$, discovery-adaptive score $s_{\mathrm{DA}}(\cdot)$
Initialize updated set $\mathcal{B} \leftarrow \emptyset$
Compute s^T_{M}, s^S_{M}, and s^B_{M}
$\mathcal{B}_{\mathrm{M}} \leftarrow \{\, i \mid s^T_{\mathrm{M}}(i) > s^S_{\mathrm{M}}(i)$ **and** $s^T_{\mathrm{M}}(i) > s^B_{\mathrm{M}}(i) \,\}$
$y^B_i \leftarrow y^T_i$ and $m_i \leftarrow 1$ for all $i \in \mathcal{B}_{\mathrm{M}}$
$\mathcal{B} \leftarrow \mathcal{B}_{\mathrm{M}}$
Define remaining candidates $\mathcal{F} \leftarrow \{\, i \mid i \notin \mathcal{B} \,\}$
Compute s^T_{DA}, s^S_{DA}, and s^B_{DA} for $i \in \mathcal{F}$
$\mathcal{B}_{\mathrm{DA}} \leftarrow \{\, i \in \mathcal{F} \mid s^T_{\mathrm{DA}}(i) > s^S_{\mathrm{DA}}(i)$ **and** $s^T_{\mathrm{DA}}(i) > s^B_{\mathrm{DA}}(i) \,\}$
$y^B_i \leftarrow y^T_i$ and $m_i \leftarrow 1$ for all $i \in \mathcal{B}_{\mathrm{DA}}$
$\mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{B}_{\mathrm{DA}}$
Define active training set $\mathcal{A} \leftarrow \{\, i \mid m_i = 1 \,\}$
Use only samples in $\mathcal{A}$ for semi-supervised training
return updated bank target y^B, active mask $\mathbf{m}$, and active set $\mathcal{A}$

entry has been updated at least once. Only activated samples are used for semi-supervised training, preventing the initial degraded inputs from directly supervising the model.

B.3 Estimation of Category Counts

In the default setting, we specify the number of incoming candidate categories according to the incremental discovery protocol. We additionally evaluate an automatic variant that estimates this number using the Akaike information criterion (AIC) [1]. Instead of manually fixing the number of new candidate categories, we run VB-CGCD inference with multiple candidate values and select the one with the lowest AIC, defined as

$$\mathrm{AIC} = 2k - 2\log \mathcal{L}, \tag{10}$$

where $\mathcal{L}$ is the likelihood of the fitted candidate model and k denotes the number of estimated parameters. The selected candidate number is then used for VB-CGCD inference in the AIC variant.

C Additional Experimental Details

C.1 Details of Text-Conditioned Baselines

Some baselines in the main comparison require predefined text or embedding inputs to specify degradation categories. These baselines are trained only with paired supervised data and are not updated using the unpaired real-world stream. For OneRestore [10], we follow its GloVe [30] text embedding initialization strategy before restoration training. As in its initialization of *low+haze* by averaging the embeddings of *low* and *haze*, we initialize the OTS [20] haze embedding by averaging *synthetic* and *haze*, and the RTTS [20] haze embedding by averaging *real* and *haze*. For DFPIR [38], whose framework operates on predefined input texts, we add "Rain degradation with rain drops" and "Snow degradation with snow flakes" following the original DFPIR text format. For real-world evaluation, each sample is manually assigned to the corresponding degradation text based on its dataset category, as DFPIR does not include an automatic text-selection mechanism. For VLU-Net [48], which fine-tunes CLIP [33] before restoration training, we add "image with raindrop" and "image with snow" to the predefined caption set, following the original VLU-Net caption style. These modifications are made to ensure a fair comparison by providing each text-conditioned baseline with the category conditioning required by its original design. Since these baselines require category information for their conditioning, this setting can be viewed as mildly favorable rather than fully blind.

C.2 Various Continual Training Scenarios

We denote by *Base* the supervised base phase trained on paired data from known degradation categories. This phase uses the paired synthetic datasets adopted in the main paper, including Outdoor-Rain [22], RainDrop [32], SPA [42], OTS [20], and Snow100K [24]. After this base training, semi-supervised adaptation is performed on unpaired real-world rain, snow, and haze datasets. Following the WResVLM [45] setup, unpaired real-world data are used for semi-supervised adaptation, while separate real-world benchmark data are used for evaluation. To analyze robustness across continual learning stages, we consider three adaptation scenarios: *Base + 3*, *Base + 1 + 2*, and *Base + 1 + 1 + 1*. Here, each + term indicates the number of unpaired real-world degradation groups introduced at a given adaptation stage. Thus, unlike standard continual restoration settings that introduce one new degradation category per stage, our setting also considers stages where multiple unpaired degradation groups are introduced in the same adaptation stage. In the *Base + 3* setting, rain, snow, and haze are introduced together after the supervised base phase, and the model is adapted for 10 epochs. In the *Base + 1 + 2* setting, real-world rain is introduced first, followed by real-world snow and haze in the next stage. In the *Base + 1 + 1 + 1* setting, rain, snow, and haze are introduced sequentially over three stages. Each adaptation stage uses the same training settings described in Appendix B.1 and is trained for 10 epochs. All sequential settings use cumulative replay, where each stage trains on both newly introduced data and all previously introduced real-world data.

C.3 Synthetic Biased Subsets

We use the synthetic CDD-11 benchmark [10] to evaluate robustness to degradation-composition bias. CDD-11 consists of eleven degradation categories: *haze*, *low*, *rain*, *snow*, *haze+rain*, *haze+snow*, *low+haze*, *low+rain*, *low+snow*, *low+haze+rain*, and *low+haze+snow*. To construct biased base training distributions, we exclude selected categories from the supervised base phase and introduce the excluded categories during semi-supervised adaptation. Specifically, for the *balanced* setting, we exclude *haze+rain*, *haze+snow*, and *low+haze+snow*, providing a less one-sided composition shift. For the *low-light-biased* setting, we exclude *haze*, *haze+rain*, and *haze+snow*, which tests whether the model tends to produce overly bright restorations. For the *haze-biased* setting, we exclude *low*, *low+rain*, and *low+snow* from base training, which tests whether the model tends to produce overly dark restorations. The excluded categories are then used for semi-supervised learning in each setting, allowing us to evaluate the effect of the discovery-adaptive score in Eq. 7 under biased base distributions.

D Additional Experimental Results

D.1 Real-world Weather Benchmark Results

Table 6 provides a degradation-wise breakdown of the real-world adverse weather results summarized in the main paper. Overall, our method remains competitive across the metrics, while the degradation-wise results show that the gains are most pronounced on haze. In particular, *Ours (Blind Semi-supervised)* achieves the best haze scores across all metrics. The AIC variant shows a similar trend with slightly lower scores, suggesting that automatic category-count estimation preserves the overall behavior despite some discovery noise. Additional qualitative comparisons on RealRain, RealSnow, and RealHaze are provided in Figs. 6 to 8. A qualitative comparison with WResVLM using the limited sample results provided in its paper is shown in Fig. 9.

D.2 Continual Training Scenario Results

Stage-wise performance under various continual training scenarios is reported in Table 7. Forgetting is computed for each metric as the drop on previously introduced categories from their best previous values. The sequential settings maintain performance comparable to the joint *Base + 3* setting, with generally limited forgetting across the four metrics. Negative MUSIQ forgetting in some stages indicates that later adaptation can also improve previously introduced categories. Overall, these results suggest that the proposed framework remains stable under sequential adaptation to newly introduced degradation groups.

Table 6: Quantitative comparisons of adverse weather removal on Real-world sets. **Bold**, underline, and *italic* indicate the best, second-best, and third-best performance, respectively.

Method	MUSIQ [16] ↑ / CLIP-IQA [41] ↑			
	Overall	Rain	Haze	Snow
Restormer [47]	56.38 / 0.438	54.69 / 0.437	53.27 / 0.366	61.18 / 0.510
WeatherDiff [28]	53.35 / 0.396	51.82 / 0.395	47.70 / 0.326	60.52 / 0.466
MWDT [29]	54.62 / 0.366	52.47 / 0.377	51.23 / 0.315	60.16 / 0.407
PromptIR [31]	56.07 / 0.443	53.48 / 0.439	53.88 / 0.372	60.86 / 0.517
DA-CLIP [27]	55.59 / 0.412	52.98 / 0.412	53.23 / 0.325	60.57 / 0.499
OneRestore [10]	56.23 / 0.425	55.70 / 0.450	54.59 / 0.379	62.47 / *0.529*
AdaIR [7]	55.33 / 0.422	55.22 / 0.441	53.58 / 0.383	61.21 / 0.520
MoCE-IR [46]	55.58 / 0.423	55.32 / 0.445	53.82 / *0.385*	61.73 / 0.510
VLU-Net [48]	56.53 / 0.415	56.18 / 0.437	54.77 / 0.377	**62.88** / 0.499
DFPIR [38]	55.60 / 0.406	55.16 / 0.432	54.05 / 0.358	61.39 / 0.520
WResVLM [45] (Non-blind Semi-supervised)	**59.34** / *0.456*	**59.80** / **0.477**	*56.09* / 0.371	62.12 / 0.519
Ours (Zero-shot)	56.64 / 0.442	54.97 / 0.435	53.82 / 0.372	61.12 / 0.519
Ours (Blind Semi-supervised, AIC)	*58.38* / 0.489	55.80 / *0.470*	57.58 / 0.446	61.76 / 0.549
Ours (Blind Semi-supervised)	58.79 / **0.494**	*56.15* / 0.474	**58.09** / **0.452**	*62.14* / **0.556**

Method	LIQE [50] ↑ / Q-Align [44] ↑			
	Overall	Rain	Haze	Snow
Restormer [47]	2.456 / 3.503	2.277 / 3.795	1.918 / 3.068	3.172 / 3.646
WeatherDiff [28]	2.173 / 3.352	2.050 / 3.640	1.520 / 2.843	2.950 / 3.573
MWDT [29]	2.230 / 3.326	2.068 / 3.548	1.720 / 2.861	2.903 / 3.569
PromptIR [31]	2.437 / 3.491	2.250 / 3.770	1.941 / 3.093	3.121 / 3.609
DA-CLIP [27]	2.438 / 3.480	2.250 / 3.732	2.014 / 3.071	3.050 / 3.637
OneRestore [10]	2.520 / 3.452	**2.646** / **3.874**	*2.194* / 3.139	**3.359** / 3.733
AdaIR [7]	2.329 / 3.370	2.508 / 3.767	1.965 / 3.080	3.197 / 3.619
MoCE-IR [46]	2.364 / 3.392	2.535 / 3.778	2.005 / 3.097	3.231 / 3.674
VLU-Net [48]	2.397 / 3.454	*2.577* / 3.851	2.020 / 3.150	3.306 / **3.747**
DFPIR [38]	2.406 / 3.369	2.485 / 3.764	2.116 / 3.077	3.209 / 3.627
WResVLM [45] (Non-blind Semi-supervised)	*2.640* / *3.574*	2.563 / *3.843*	2.064 / *3.176*	*3.293* / 3.702
Ours (Zero-shot)	2.602 / 3.515	2.518 / 3.783	2.105 / 3.109	3.182 / 3.652
Ours (Blind Semi-supervised, AIC)	2.681 / 3.586	2.566 / 3.801	2.221 / 3.242	3.264 / 3.713
Ours (Blind Semi-supervised)	**2.702** / **3.593**	2.586 / 3.804	**2.243** / **3.254**	3.278 / *3.722*

D.3 CDD-11 Benchmark

We provide additional quantitative results on the CDD-11 biased subsets. As shown in Table 8, we compare the degraded inputs and the restored outputs of our method on the novel categories introduced during semi-supervised adaptation. These novel categories correspond to the categories excluded from the supervised base phase for each biased subset, and the reported metrics are averaged over them. Across all biased subsets, the restored outputs improve over the degraded inputs on PSNR, SSIM, MUSIQ, CLIP-IQA, and LIQE. These results suggest that the model maintains generalization to newly introduced degradation types, even without using predefined degradation category labels or textual descriptions. We further compare paired PSNR and SSIM alongside MUSIQ in Table 9 for the zero-shot model, MUSIQ-guided adaptation using the Semi-UIR framework [15], and the same adaptation with S_{DA}. Adding S_{DA} to Semi-UIR improves PSNR, SSIM, and MUSIQ in the balanced and low-light-biased settings. In the haze-biased setting, MUSIQ improves and CIEDE2000 decreases as reported in Table 4, while PSNR and SSIM decrease slightly relative to Semi-UIR.

D.4 Unseen Degradations

We further evaluate generalization to unseen artifact types using the POLED and TOLED datasets from the UDC benchmark [54]. As shown in Table 10, Ours improves over the input on POLED in PSNR, SSIM, and CLIP-IQA, and achieves the best scores among restoration methods for these metrics. On TOLED, the input already has high full-reference scores, and the best restoration

Table 7: Stage-wise results under various continual training scenarios on real-world sets. Overall scores are averaged over introduced degradation groups. $\text{Forgetting}_{\text{M/C}}$ and $\text{Forgetting}_{\text{L/Q}}$ denote MUSIQ/CLIP-IQA and LIQE/Q-Align drops on previous categories. Negative values indicate improvement.

Configuration	Stage	MUSIQ ↑ / CLIP-IQA ↑				$\text{Forgetting}_{\text{M/C}}$ ↓
		Overall	Rain	Snow	Haze	
$Base + 3$	1	58.79 / 0.494	56.15 / 0.474	62.14 / 0.556	58.06 / 0.452	– / –
$Base + 1 + 2$	1	55.98 / 0.458	55.98 / 0.458	– / –	– / –	– / –
$Base + 1 + 2$	2	58.84 / 0.467	56.03 / 0.453	62.16 / 0.536	58.30 / 0.410	-0.056 / 0.004
$Base + 1 + 1 + 1$	1	56.05 / 0.458	56.05 / 0.458	– / –	– / –	– / –
$Base + 1 + 1 + 1$	2	59.02 / 0.492	55.76 / 0.449	62.27 / 0.534	– / –	0.287 / 0.008
$Base + 1 + 1 + 1$	3	59.34 / 0.468	56.27 / 0.452	62.21 / 0.531	59.54 / 0.419	-0.078 / 0.004

Configuration	Stage	LIQE ↑ / Q-Align ↑				$\text{Forgetting}_{\text{L/Q}}$ ↓
		Overall	Rain	Snow	Haze	
$Base + 3$	1	2.702 / 3.593	2.586 / 3.804	3.278 / 3.722	2.243 / 3.254	– / –
$Base + 1 + 2$	1	2.600 / 3.809	2.600 / 3.808	– / –	– / –	– / –
$Base + 1 + 2$	2	2.701 / 3.587	2.589 / 3.790	3.279 / 3.725	2.234 / 3.244	0.011 / 0.018
$Base + 1 + 1 + 1$	1	2.612 / 3.818	2.612 / 3.818	– / –	– / –	– / –
$Base + 1 + 1 + 1$	2	2.936 / 3.762	2.589 / 3.790	3.281 / 3.732	– / –	0.022 / 0.027
$Base + 1 + 1 + 1$	3	2.721 / 3.589	2.591 / 3.784	3.279 / 3.722	2.293 / 3.260	0.011 / 0.021

Table 8: Degraded inputs and our restored outputs on the paired CDD-11 biased subsets, averaged over categories excluded from supervised base training.

Split	Setting	PSNR ↑	SSIM ↑	MUSIQ ↑	CLIP-IQA ↑	LIQE ↑
Haze-biased	Input	13.36	0.380	43.49	0.39	1.47
	Output	20.57	0.665	53.57	0.45	2.07
Low-light-biased	Input	15.14	0.770	65.90	0.59	2.86
	Output	17.59	0.873	69.81	0.66	4.05
Balanced	Input	14.73	0.670	57.88	0.50	2.24
	Output	18.76	0.793	61.61	0.57	3.14

results are distributed across different methods. These results indicate that performance on unseen degradation types depends on the artifact type and evaluation metric.

D.5 VB-CGCD Classification Accuracy

We additionally report the stage-wise top-1 category prediction accuracy of VB-CGCD on the real-world evaluation sets in Table 11. We evaluate three continual training configurations, *Base + 3*, *Base + 1 + 2*, and *Base + 1 + 1 + 1*, using either the specified number of incoming categories or AIC-based category-count estimation. For each image, we extract the DINO feature and assign the degradation category using the top-1 VB-CGCD prediction, with dataset-level labels used only as evaluation references. With the specified number of incoming categories, VB-CGCD maintains high accuracy across stages, although performance slightly decreases as more categories are introduced. In contrast, AIC-based category-count estimation leads to substantially lower accuracy, especially in multi-stage settings.

E Analysis of Category Discovery and Restoration Quality

E.1 Effect of AIC-Based Category-Count Estimation on Restoration

In the specified-count setting, K gives the number of incoming groups at each update, but neither their semantic identities nor sample assignments. This can represent a planned, user-initiated update in which the operator specifies how many new condition groups are expected in the collected batch.

Table 9: Comparison of restoration settings on the paired CDD-11 biased subsets, averaged over categories excluded from supervised base training. Semi-UIR denotes MUSIQ-only reliable bank updates, and $+S_{\mathrm{DA}}$ denotes our final model, corresponding to Output in Table 8.

Split	Setting	PSNR ↑	SSIM ↑	MUSIQ ↑
Haze-biased	Zero-shot	19.33	0.604	48.11
	Semi-UIR	**20.74**	**0.670**	52.75
	$+S_{\mathrm{DA}}$	20.57	0.665	**53.57**
Low-light-biased	Zero-shot	16.65	0.850	67.77
	Semi-UIR	17.20	0.871	69.34
	$+S_{\mathrm{DA}}$	**17.59**	**0.873**	**69.81**
Balanced	Zero-shot	17.01	0.750	58.51
	Semi-UIR	18.23	0.787	61.31
	$+S_{\mathrm{DA}}$	**18.76**	**0.793**	**61.61**

Table 10: Quantitative evaluation on unseen degradation datasets from the UDC benchmark [54]. **Bold** indicates the best performance among restoration methods for each dataset and metric.

Dataset	Method	PSNR ↑	SSIM ↑	MUSIQ ↑	CLIP-IQA ↑	LIQE ↑
POLED	Input	15.55	0.643	29.91	0.266	1.023
	VLU-Net [48]	15.32	0.591	25.35	0.224	1.023
	AdaIR [7]	15.01	0.550	**29.14**	0.270	1.017
	MoCE-IR [46]	15.26	0.610	28.10	0.327	**1.033**
	Ours	**17.52**	**0.674**	27.51	**0.372**	1.026
TOLED	Input	29.07	0.867	20.31	0.198	1.018
	VLU-Net [48]	**27.94**	**0.849**	**21.82**	0.210	1.017
	AdaIR [7]	25.11	0.843	19.74	0.235	1.016
	MoCE-IR [46]	22.28	0.830	20.35	**0.335**	**1.030**
	Ours	23.97	0.818	20.48	0.193	1.020

Table 11: Stage-wise top-1 category prediction accuracy of VB-CGCD on the real-world evaluation sets. Accuracy is measured using dataset-level degradation labels as label references.

Category-count Setting	Continual Training Scenario	Stage 1	Stage 2	Stage 3
Specified K	$Base + 3$	94.72%	–	–
Specified K	$Base + 1 + 2$	98.46%	95.47%	–
Specified K	$Base + 1 + 1 + 1$	98.46%	97.76%	95.63%
AIC-estimated K	$Base + 3$	81.37%	–	–
AIC-estimated K	$Base + 1 + 2$	85.62%	61.75%	–
AIC-estimated K	$Base + 1 + 1 + 1$	86.52%	68.40%	54.91%

VB-CGCD [8] likewise uses specified counts in its main experiments. Estimating an unknown count is a distinct challenge in CGCD, and VB-CGCD reports a separate evaluation with estimated counts. We therefore use specified K to isolate discovery-conditioned restoration from count-estimation errors and evaluate AIC when the count must be inferred. We compare both settings at the final stage of matched arrival schedules in Table 12. In *Base + 3*, ACC falls from 94.72% to 81.37%, while MUSIQ declines from 58.79 to 58.38 and CLIP-IQA from 0.494 to 0.489. At the final stage of *Base + 1 + 1 + 1*, ACC falls from 95.63% to 54.91%, whereas MUSIQ rises from 59.34 to 59.92 and CLIP-IQA from 0.468 to 0.483. In *Base + 1 + 2*, ACC similarly falls from 95.47% to 61.75%, while MUSIQ changes only from 58.84 to 58.80 and each degradation-wise change remains within 0.29. Thus, reductions in one-to-one category accuracy do not consistently correspond to reductions in restoration quality.

To interpret the different changes in ACC and restoration quality, we examine the AIC-discovered clusters in Table 13. NMI [40] measures overall agreement between cluster assignments and dataset labels. Purity [51] measures how strongly each cluster is dominated by a single dataset category. Completeness [35] measures whether samples of the same dataset category remain together in one

Table 12: Final-stage category accuracy and restoration quality under specified and AIC-estimated incoming group counts. ACC measures one-to-one agreement with dataset-level degradation labels.

Arrival Schedule	Count Setting	ACC ↑	MUSIQ ↑	CLIP-IQA ↑
$Base + 3$	Specified K	94.72%	58.79	0.494
$Base + 3$	AIC	81.37%	58.38	0.489
$Base + 1 + 1 + 1$	Specified K	95.63%	59.34	0.468
$Base + 1 + 1 + 1$	AIC	54.91%	59.92	0.483

Table 13: Stage-wise AIC category-count estimation and cluster structure under the *Base + 1 + 1 + 1* schedule. K_{GT} and $\hat{K}_{\mathrm{AIC}}$ refer to each incoming update, whereas NMI, purity, completeness, and ACC are evaluated over all groups observed up to that stage. NMI, purity, and completeness are omitted at Stage 1 because the reference partition contains only one degradation group.

Setting	Stage	K_{GT}	$\hat{K}_{\mathrm{AIC}}$	NMI	Purity	Completeness	ACC
Base + 1	1	1	2	–	–	–	86.52%
Base + 1 + 1	2	1	2	0.688	0.982	0.568	68.40%
Base + 1 + 1 + 1	3	1	2	0.729	0.948	0.664	54.91%

Table 14: Effect of actual VB-CGCD misclassification on restoration quality. Results are computed only on samples whose top-1 VB-CGCD prediction differs from the dataset-level degradation label. *Predicted* denotes restoration using the predicted, but incorrect, category condition from the original pipeline. *Oracle* denotes restoration after replacing only the category condition with the dataset-level degradation label.

Dataset	Incorrect / Total	Predicted MUSIQ ↑ / CLIP-IQA ↑	Oracle MUSIQ ↑ / CLIP-IQA ↑	Δ MUSIQ / CLIP-IQA
RealRain	285 / 2320	49.39 / 0.428	49.82 / 0.429	+0.43 / +0.002
RTTS	139 / 4322	57.31 / 0.410	57.54 / 0.413	+0.24 / +0.004
Snow100K-R	235 / 1329	58.04 / 0.459	58.24 / 0.462	+0.20 / +0.003
Overall	**659 / 7971**	**54.15 / 0.435**	**54.45 / 0.438**	**+0.31 / +0.003**

cluster. At Stages 2 and 3, AIC estimates more incoming clusters than the reference count, while high purity and limited completeness are consistent with subdivisions of dataset-level categories (e.g., heavy rain and drizzle both labeled as rain). One-to-one ACC penalizes such subdivisions because only one discovered cluster can be matched to each dataset-level category. Together with the matched restoration results, this indicates that a drop in ACC does not necessarily imply weaker restoration guidance. ACC counts both potentially useful within-category subdivisions and cross-category mismatches as errors, although their effects on restoration can differ.

E.2 Effect of Actual CGCD Misclassification

To isolate the effect of category misclassification from category-count estimation errors, we conduct this oracle analysis in the specified-K setting. This analysis is conducted on the real-world evaluation sets, including RealRain, RTTS, and Snow100K-R. For each image, we extract the DINO feature, obtain the top-1 VB-CGCD prediction, and collect samples whose predicted category differs from the dataset-level degradation label. For these misclassified samples, we compare the restoration result produced by the original pipeline using the incorrectly predicted category condition with an oracle result obtained by replacing only the category condition with the ground-truth degradation category. The input image, restoration model, and checkpoint are kept fixed in both cases, using the same checkpoint as our main reported model. Thus, this analysis changes only the category condition at inference time. This comparison helps separate the effect of incorrect category conditioning from the intrinsic difficulty of the misclassified samples. Table 14 summarizes the results. Among 7,971 real-world evaluation images, 659 samples are misclassified by the top-1 VB-CGCD prediction. Replacing the predicted descriptor with the oracle degradation descriptor consistently improves restoration quality across all datasets on the misclassified subset. Under specified K, the oracle improvement shows that correcting category-level mismatches can improve restoration without

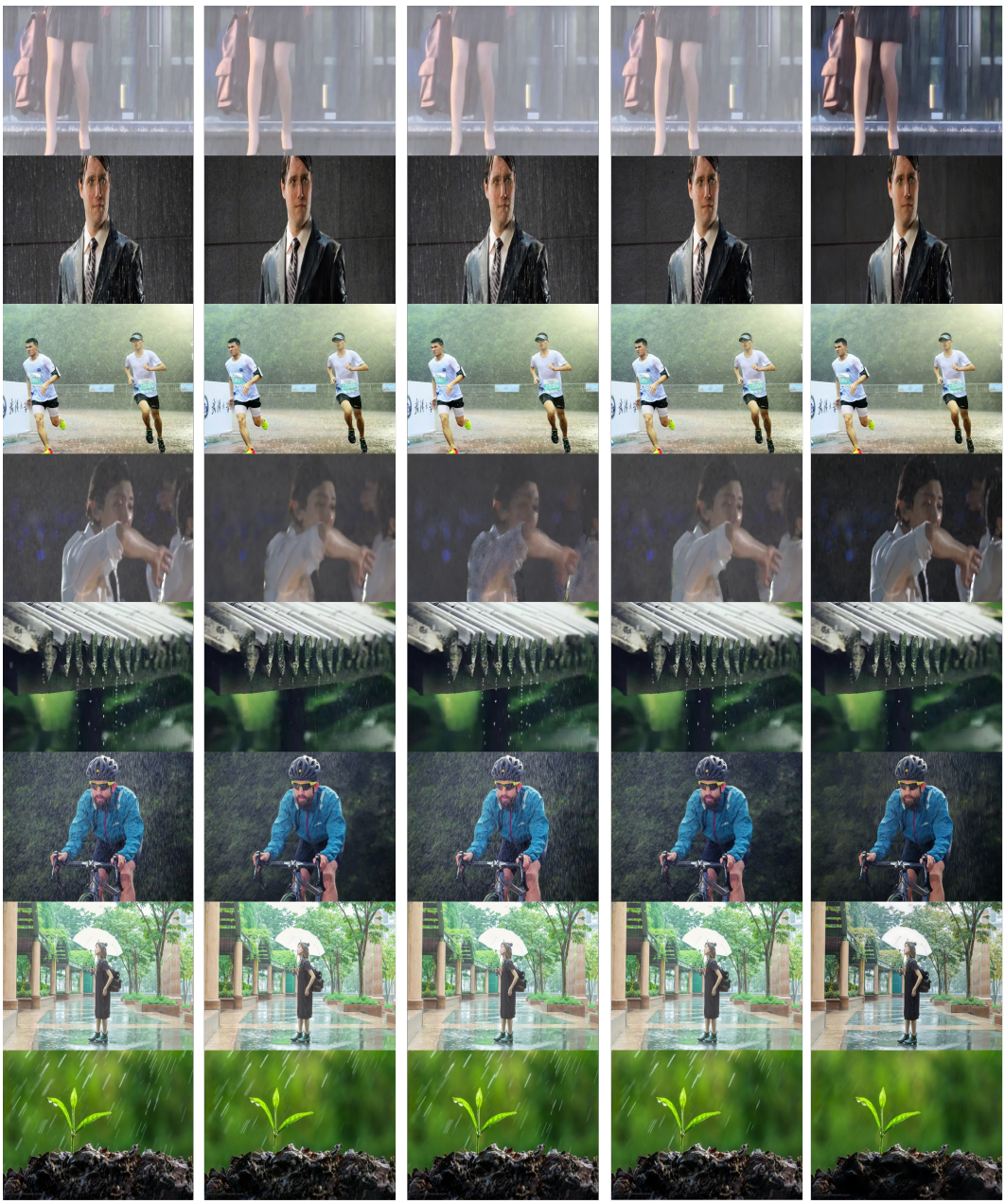

Figure 6: Qualitative comparison with competing baselines on RealRain. Ours produces cleaner restored images with fewer remaining rain streaks.

category-count estimation errors. Under AIC-estimated K, one-to-one ACC drops substantially while restoration quality remains comparatively stable, as shown in Table 12. Together, these results seem to indicate that within-category subdivisions can retain useful restoration guidance, whereas mismatched category conditions can impair restoration.

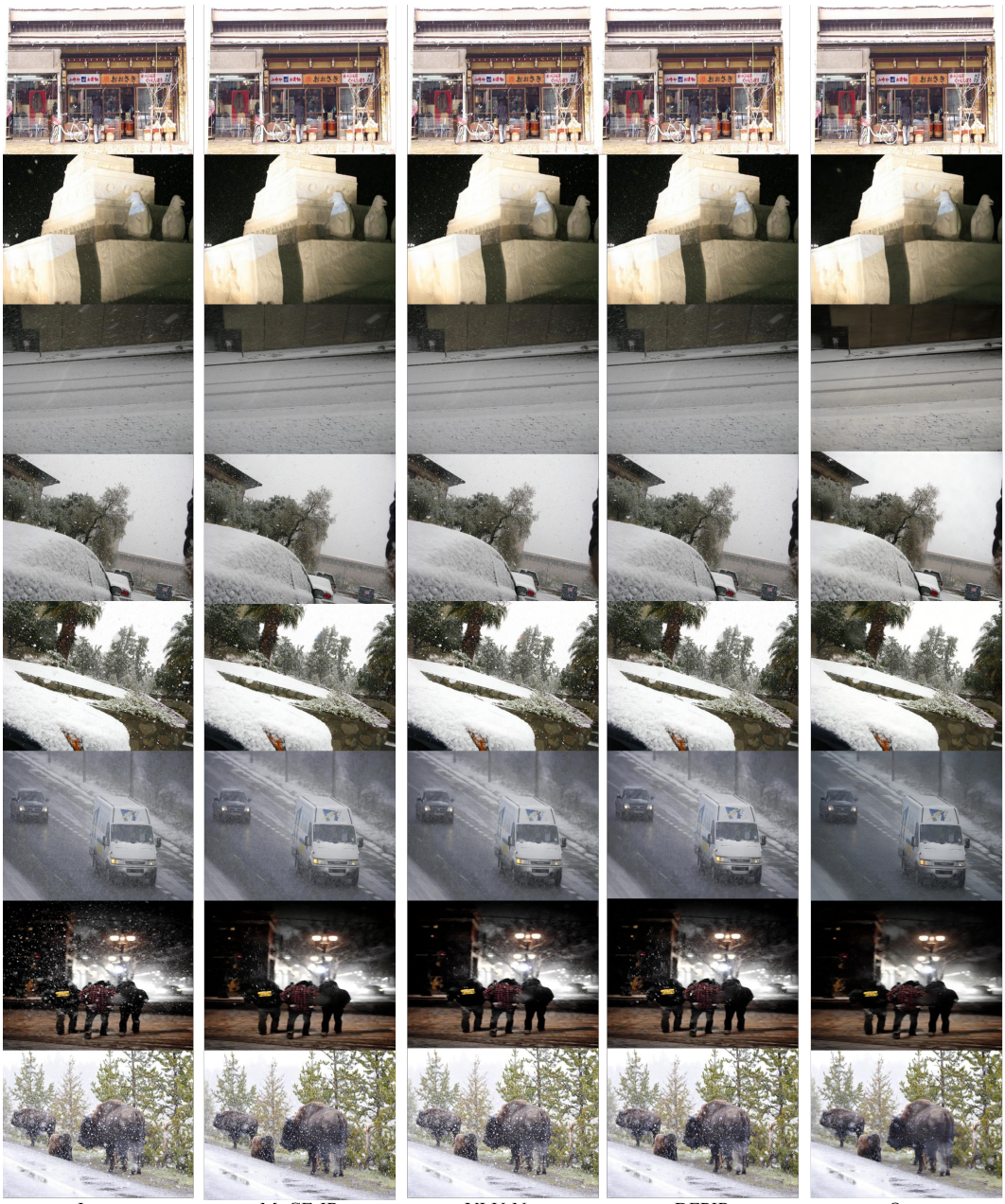

Figure 7: Qualitative comparison with competing baselines on real snow. Ours produces cleaner restored images with fewer remaining snowflake.

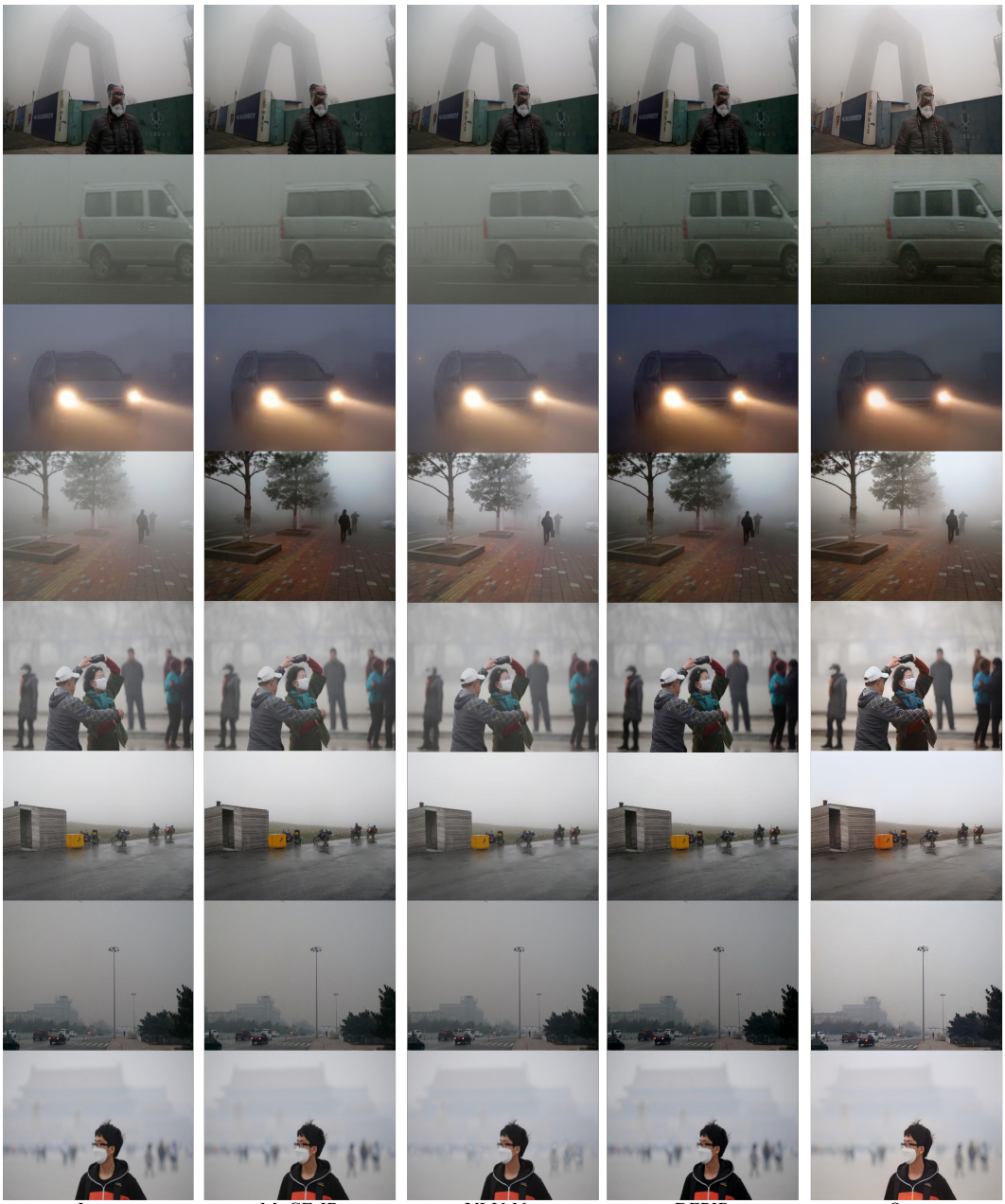

Figure 8: Qualitative comparison with competing baselines on real haze. Ours alleviates haze while preserving natural brightness and image details.



Input WResVLM Ours

Figure 9: Qualitative comparison with WResVLM on real snow and rain images using the limited sample results provided in the WResVLM paper.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly outline our focus on open world image restoration and the challenges of unlabeled and unpaired data streams. We explicitly summarize our core methodological contributions including discovery conditioned descriptors and the discovery adaptive score at the end of the introduction. These claims are directly supported by our experimental results and ablation studies.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses a current limitation in the Conclusion. Specifically, the proposed sequential adaptation setting relies on cumulative replay, which increases training cost as more real-world data are accumulated across stages. We also discuss future work on replay-free adaptation strategies, such as LoRA-based updates, to improve scalability for long-term continual restoration.

   Guidelines:

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

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: This paper does not contain theoretical results requiring formal proofs. Our work is entirely empirical and focuses on model design and experimental validation.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper provides the details required to reproduce the main experimental results, including the experimental environment, datasets, evaluation metrics, training protocols, hyperparameter settings, and implementation details. These details are described in the Appendix, ensuring that the main claims and findings are reproducible.

   Guidelines:

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

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No]

   Justification: While we plan to release the code upon publication, it is not available at the time of submission. The main contribution of this work lies in the proposed open-world continual restoration framework and its empirical validation, rather than in introducing a new benchmark or open-source framework. To support reproducibility, we provide detailed descriptions of the implementation, datasets, training protocols, and evaluation procedures in the main paper and Appendix Section B.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We comprehensively describe the experimental settings in Appendix. This includes the complete list of training and testing datasets along with their specific configurations. We also provide exhaustive implementation details such as the optimizer choice, learning rate schedules, batch sizes, patch sizes, and hardware specifications. These thorough descriptions ensure our experimental pipeline is fully transparent and understandable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: Our primary focus was demonstrating the overall effectiveness of the proposed continual framework across diverse open world degradation streams using consistent evaluation protocols. Due to the substantial computational cost associated with repeated training and inference for our large scale restoration model we were unable to conduct multiple independent runs to report statistical variability. We plan to incorporate comprehensive statistical evaluations including variance across runs and confidence intervals in our future work to further strengthen the empirical analysis.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We explicitly detail the computational resources required to reproduce our experiments within the implementation details section. We specify that all training and evaluation procedures were conducted using three NVIDIA A100 GPUs with 80GB of memory each operating on the PyTorch framework. Additionally we provide the approximate execution time required to complete the full adaptation process to give a clear estimate of the total compute.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in this paper strictly adheres to the NeurIPS Code of Ethics. Our work focuses on developing a framework for open world image restoration using standard publicly available benchmark datasets. These datasets do not contain sensitive personal information or involve human subjects. Furthermore our proposed methodology aims to improve the reliability and robustness of vision systems in unpredictable environments presenting no foreseeable societal harm or ethical concerns.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: This work focuses on open-world continual image restoration under unknown and unpaired degradations. While improved restoration could benefit downstream vision systems operating under adverse conditions, the paper does not involve real-world deployment, human subjects, sensitive attributes, or decision-making systems. The method is evaluated on standard restoration benchmarks and is not intended for applications that directly affect individuals or social outcomes. Therefore, we consider the direct societal impact of this work to be limited at this stage.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [N/A]

Justification: This paper does not release high-risk models or datasets, such as large generative models, language models, or scraped datasets intended for broad downstream use. The proposed method is an image restoration framework evaluated on synthetic and standard real-world degradation benchmarks. Since the work does not introduce resources with a high risk of misuse or dual-use concerns, specific release safeguards are not applicable.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The paper uses existing datasets, pretrained models, and baseline methods only for research and evaluation purposes. We properly cite the original works for all datasets, benchmarks, models, and evaluation metrics used in the experiments, including the real-world adverse weather datasets, CDD-11, UDC, DINO, VB-CGCD, OneRestore, and competing restoration baselines. We do not redistribute or repackage these assets as part of the submission, and our use follows the intended academic research setting of the corresponding resources.

Guidelines:

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

13. **New assets**

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

Answer: [N/A]

Justification: This paper does not introduce or release a new dataset, benchmark, or public model asset at submission time. The main contribution is the proposed open-world continual image restoration framework and its empirical validation. We plan to release code and pretrained models upon publication, and the corresponding documentation will be provided with the released assets. Therefore, documentation for newly released assets is not applicable at submission time.

Guidelines:

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

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

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

Answer: [N/A]

Justification: This work does not involve crowdsourcing, user studies, or research with human subjects. All experiments are conducted using existing image restoration benchmarks and synthetic or real-world degradation datasets. Therefore, participant instructions, screenshots, and compensation details are not applicable.

Guidelines:

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

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

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

Answer: [N/A]

Justification: This research does not involve human subjects, crowdsourced participants, or user studies. The experiments are performed on existing image restoration datasets and benchmarks. Therefore, IRB approval or an equivalent human-subjects review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: This work does not use LLMs as an important, original, or non-standard component of the proposed method. The core methodology is based on continual degradation discovery, descriptor-conditioned restoration, and semi-supervised adaptation. Therefore, declaration of LLM usage is not applicable under the checklist guidelines.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.