A Code and Data Availability

We make all data augmentation, training, and evaluation scripts and baseline experiment source code available here: `https://anonymous.4open.science/r/PRISM-E4E3/README.md`.

B LLM Usage Statement

The authors of this paper minimally (fewer than ten times) used open-source LLM tools in order to provide editing suggestions for the paper (paraphrasing, synonyms, etc.).

C Training Details

PRISM is trained on 8×40GB NVIDIA A100 GPUs using mixed-precision training (FP16) with distributed data parallelism.

Hyperparameters. For our final model we use:

1. Batch size: 264 (global, distributed across 8 GPUs)
2. Learning rate: 1×10^{-4} (AdamW optimizer, $\beta_1 = 0.9, \beta_2 = 0.999$, weight decay $= 0.01$)
3. No learning-rate warm-up; cosine decay schedule
4. Training epochs: 500
5. Input resolution: 256×256
6. Gradient clipping: 1.0
7. EMA decay: 0.9999 for model weights

We initialize the backbone from publicly available Stable Diffusion v1.5 weights (Rombach et al., 2022).

CLIP Fine-Tuning. For the embedding space, we initialize from OpenAI CLIP ViT-B/32 (Radford et al., 2021) pretrained weights. We fine-tune only the final projection layers and cross-attention adapters, freezing the base vision and text encoders to preserve semantic alignment. Fine-tuning uses:

- Batch size: 512
- Learning rate: 5×10^{-5} (AdamW)
- Epochs: 50
- Temperature parameter in contrastive loss initialized at 0.07 and annealed to 0.04

D Datasets and Evaluation

Mixed Degradations Dataset As stated in Section 3, our composite degradation dataset used for ground truth during training was drawn from a diverse collection of datasets spanning multiple scientific and environmental imaging domains. Table 5 summarizes the datasets used.

Data Augmentation Pipeline Each clean image is transformed into a distorted counterpart using a multi-step augmentation strategy designed to simulate diverse, compounding visual degradations. First, for each image, the number of distortions to apply, N is sampled. To balance across simple and complex cases, degradations are sampled according to a multinomial distribution: 50% single distortions, 30% two-way mixtures, 20% three-way mixtures.

We found that $N > 3$ degraded the signal significantly and made the task of restoration too difficult, so we limit N to a maximum of 3 distortions. Given the sampled number N, distinct distortion types are drawn uniformly from a predefined library $\mathcal{D}$ of transformations. Our distortion set spans 14 categories: including geometric distortions (motion blur, warping, refraction, defocus blur), photometric degradations (contrast, color shifts, brightness, low light), occlusions (clouds, haze, rain,

Table 5: Summary of Training Datasets. PRISM is trained on a diverse set of natural and scientific domains spanning ecological, medical, astronomical, and remote sensing imagery.

Dataset	Description	Size
ImageNet (Deng et al., 2009)	1000-class benchmark of natural images with visually diverse scenes.	1.2M
Sen12MS (Sentinel-2) (Schmitt et al., 2019)	RGB satellite image patches with and without clouds, used for land-cover and cloud-removal tasks.	720K
iWildCam 2022 (JohnBeuving et al., 2022)	Camera trap sequences for wildlife monitoring under challenging lighting and environmental conditions.	28.8K
EUVP (Islam et al., 2020)	Paired underwater photos with clear vs. distorted conditions (enhanced clean split).	3.7K
Cityscapes (Cordts et al., 2016)	Urban street scenes captured from vehicle-mounted cameras (includes 5K labeled and 20K "extra").	25K
BioSR (Gong et al., 2021)	Fluorescence microscopy slides (wide-field vs. SIM ground truth) for super-resolution and denoising tasks.	14K
Brain Tumor MRI (Nickparvar, 2021)	Clinical MRI scans for tumor detection and segmentation with paired clean/noisy modalities.	7K
AstroSR HSC Surveys (Miao et al., 2024)	Wide-field sky survey images from the Hyper Suprime-Cam (HSC), used for astrophysical source recovery.	2K

snow), and noise-based effects (additive noise, compression). Parameter ranges for each degradation type are uniformly sampled within physically realistic bounds (e.g., haze density $\alpha \sim U(0.1, 0.5)$; Gaussian blur kernel size $\sigma \sim U(0.5, 3.0)$). Further information on the implementation of this distortion library is provided in the codebase. Finally, the selected distortions are randomly ordered and sequentially applied to the clean image. Randomizing the application order reflects the non-commutative nature of compound distortions and further increases the diversity and realism of the visual outcomes. This data augmentation process is summarized in Fig. 7.

Prompts p describing distortions are auto-generated with GPT-4 (Hurst et al., 2024) to simulate the variability in natural-language queries that may be provided as input. We sample multiple phrasings per distortion type to encourage robustness to linguistic variation (e.g., "remove haze," "dehaze the image," "clear atmospheric fog"). In addition, we generate *compound prompts* that explicitly describe multiple simultaneous degradations (e.g., "remove blur and color shift"), ensuring that the model is trained on realistic mixtures rather than isolated categories.

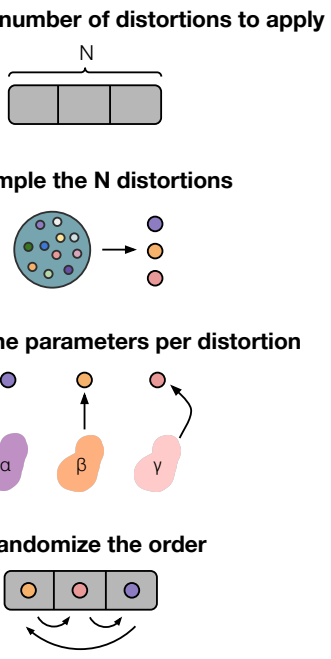

Figure 7: An overview of the data augmentation pipeline of diverse compound degradations.

To improve controllability, we further incorporate:

- **Partial prompts**: instructing the model to remove only a subset of degradations present in I_{dist}, requiring the model to learn selective restoration (e.g., input degraded with haze + rain + blur, prompt: "remove haze and blur").
- **Negative prompts**: instructing the model to remove degradations that are *absent*, which enforces that restoration actions are conditional on both image evidence and textual prompts. For instance, if the input is degraded with haze + blur and the prompt is "remove snow," the model should leave the image unchanged with respect to snow.

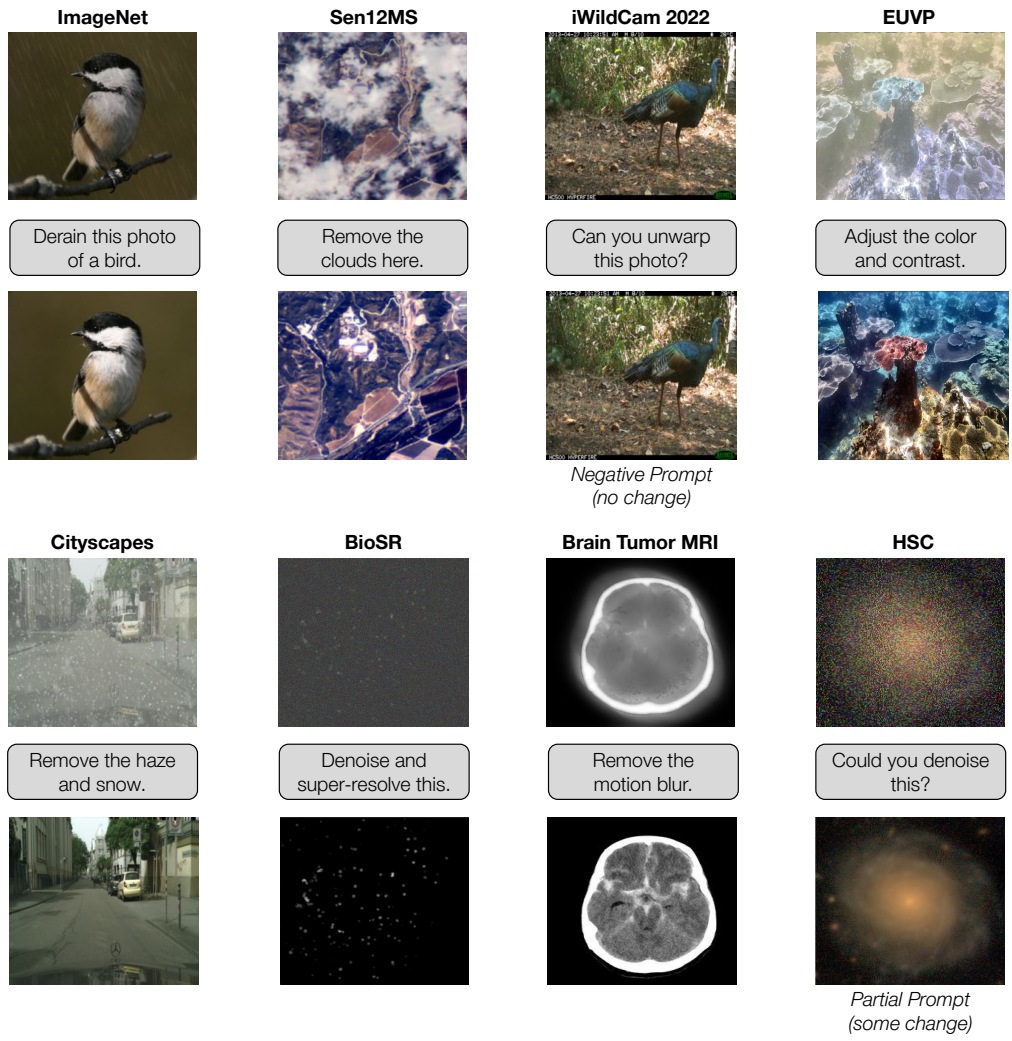

Figure 8: *MDB Examples.* Samples from our compound degradation dataset, with paired "clean" (top) and "distorted" (bottom) images, with corresponding prompts in between.

Approximately 20% of the training samples are partial prompts and 10% are negative prompts. The inclusion of partial and negative prompts is critical for teaching PRISM to respect expert instructions. Without them, the model tends to over-restore, indiscriminately removing all degradations it detects. By explicitly training on examples where the correct action is *not* to remove a present or absent degradation, PRISM learns to balance restoration fidelity with adherence to prompts.

The final training corpus consists of approximately 2.5M triplets, split $80/19/1$ across training, validation, and held-out test sets. For each clean image, multiple degraded variants and prompts are generated, increasing coverage of both degradation mixtures and linguistic variability. Fig. 8 demonstrates our dataset diversity, with examples of clean, distorted, and prompt triplets.

E Evaluation

As discussed in Section 3, we evaluate PRISM along three complementary axes: (1) restoration under synthetic compound degradations, (2) downstream utility in real scientific datasets, and (3) zero-shot robustness to unseen real-world distortions. A summary of each evaluation testbed is provided in Table 6. Unless noted otherwise, all outputs are generated using a fixed random seed=42.

Table 6: We group evaluations into (1) synthetic compound degradations, (2) downstream utility in real scientific datasets, and (3) zero-shot robustness to unseen real-world distortions.

Evaluation Setting	Task / Dataset	Dataset Size	Metric
(1) Synthetic Compound Degradations			
MDB	Synthetic mixtures from clean datasets	25K	PSNR, SSIM, FID, LPIPS
(3) Zero-Shot Robustness to Unseen Real-World Distortions			
Underwater Imaging	UIEB (Li et al., 2019)	890	PSNR, SSIM, LPIPS
Under-Display Cameras	POLED (UDC) (Zhou et al., 2021)	512	PSNR, SSIM, LPIPS
Fluid Lensing	ThapaSet	600	PSNR, SSIM, LPIPS
(2) Downstream Utility in Real Scientific Datasets			
Land Cover Classification	Sen12MS-CR (cloudy Sentinel-2 patches) (Ebel et al., 2020; Schmitt et al., 2019)	200	Classification accuracy (ResNet50)
Wildlife Classification	iWildCam 2022 (camera traps) (JohnBeuving et al., 2022)	200	Classification accuracy (SpeciesNet)
Microscopy Segmentation	BioSR (WF vs. SIM microscopy) (Gong et al., 2021)	10K	Instance Segmentation mIoU (MicroSAM)
Urban Scene Understanding	Rooftop Cityscapes (haze/low-light)	5K	Panoptic segmentation mIoU (RefineNet)

Downstream Utility. Here, we provide specific details about how we constructed our novel benchmarking suite for evaluation over downstream utility. We re-purpose real datasets with distortions that present known challenges for models across remote sensing, wildlife monitoring, microscopy and weather, where ground truth labels are available *not* because the distorted images are annotated, but because undistorted views exist at different points in time or are collected from a more sophisticated sensor.

Rather than training task-specific models for each of these downstream tasks, we deliberately use off-the-shelf pretrained models. This design choice reflects a realistic scenario: domain experts are far more likely to apply widely available models for segmentation than to train bespoke models for each experimental setup. Using off-the-shelf models therefore provides a conservative estimate of restoration utility in practice and avoids confounding performance gains from joint training on dataset-specific distributions. If restoration improves the outputs of a generic model, this strongly suggests practical downstream utility beyond controlled benchmarks. In each of the four domains, we examine restoration performance on a specified distortion against the default set of automatically-detected distortions present in the input image. We do not compare against domain-specific restoration models (e.g., dedicated cloud removal networks) because our goal is to evaluate generalist models that can flexibly handle a wide variety of distortions; this broader applicability makes them more useful in scientific imaging, where degradations are diverse, overlapping, and often domain-shifted.

1. **Land Cover Classification:** To assess performance on land cover classification over satellite data, we select 200 cloudy satellite images degraded by cloud cover from the Sen12MS-CR dataset Ebel et al. (2020), with land cover labels derived from temporally aligned, cloud-free views in the Sen12MS dataset Schmitt et al. (2019). We evaluate using a ResNet50, trained on satellite imagery that includes minimal cloud cover Papoutsis et al. (2023). It is important to evaluate on land cover classification because accurate identification of surface types (e.g., forests, croplands, urban areas) under degraded conditions like cloud cover directly underpins large-scale monitoring of climate change, biodiversity, and resource management.
2. **Wildlife Classification** Camera trap classification is critical for ecological monitoring, enabling large-scale biodiversity surveys without direct human observation. We evaluate our model on the task of species identification using *iWildCam 2022 Camera Trap* dataset (JohnBeuving et al., 2022). Specifically, we use 200 nighttime wildlife images, after filtering out blank frames (no species) and sample frames with low confidence species predictions (< 0.70). Ground truth annotations are sourced from expert labels of alternate frames in the same camera trap sequence. Classification is evaluated using SpeciesNet (Gadot et al., 2024), a classifier trained on over 6 million camera trap images.
3. **Microscopy Segmentation** Next, we evaluate our model on the task of microscopy image segmentation, which informs the quantification of organelle morphology and dynamics, which are central to understanding cell physiology and disease. We build on the *BioSR dataset* (Gong et al., 2021) introduced by Qiao et al. This dataset was acquired using

paired low resolution wide-field (WF, diffraction-limited) and super-resolved structured illumination microscopy (SIM) images of cellular structures (clathrin-coated pits) across a wide range of signal-to-noise ratios. In our setting, the WF images serve as noisy "distorted" inputs, while the corresponding high-SNR SIM sensor data provide the undistorted reference. This setup allows us to evaluate restoration not against simulated degradations, but against experimentally aligned ground truth. We measure performance by applying restored images to the downstream task of segmentation, using the microscopy foundation model MicroSAM model (Archit et al., 2025) to generate cell-structure masks, and report segmentation accuracy compared to the high-quality SIM annotations. This mirrors real-world use, where quantitative biological conclusions (e.g., about organelle morphology or cytoskeletal organization) depend critically on reliable segmentation.

4. **Urban Scene Understanding** We also evaluate our model on the task of cityscape scene understanding, which enables reliable monitoring of urban forests. To do so, we collected, labeled, and processed a novel *Rooftop Cityscapes* dataset for an additional setting: and haze in urban scenes. Specifically, we deployed fixed-position cameras on several building rooftops across multiple days under varying weather and lighting conditions. From each sequence, we manually identified and labeled frames with clear conditions to serve as the ground truth. We applied an off-the-shelf panoptic segmentation model (pre-trained on the original Cityscapes (Cordts et al., 2016) dataset) to each distorted and restored frame. To ensure reliable comparison, we restricted evaluation to "stationary" classes (buildings, vegetation, and sky) while ignoring dynamic objects such as cars or pedestrians, which may change across frames and introduce label inconsistency. See Fig. 20 for qualitative examples from this custom dataset.

Statistical Significance Evaluation To assess whether Selective Restoration provided a statistically significant improvement over Full Restoration, we conducted paired hypothesis tests across repeated experimental runs. Each model was trained and evaluated with multiple random seeds on the same dataset splits (seeds 2, 42, and 420), yielding a distribution of results for each condition.

For every domain and downstream task, we computed paired differences between the two methods:

$$d_i = \text{Selective}_i - \text{Full}_i, \quad i = 1, \ldots, n$$

where n is the number of runs (seeds/splits). This controls for variability due to dataset sampling and ensures that each comparison is made under identical conditions.

We then applied a two-tailed paired t-test to the differences d_i:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \quad s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2},$$

with $n-1$ degrees of freedom. The null hypothesis H_0 is that Selective and Full Restoration perform equally ($\mu_d = 0$). The p-value is computed as:

$$p = 2 \cdot P(T_{n-1} \geq |t|)\,,$$

where T_{n-1} is a Student's t distribution with $n-1$ degrees of freedom.

1. If $p < 0.05$, we reject H_0 and conclude that Selective Restoration significantly outperforms Full Restoration.
2. If $p \geq 0.05$, we fail to reject H_0, indicating that the observed difference may be due to random variation.

In Table 4, we report mean ± standard deviation across runs for each method, alongside the resulting p-values. The results show that Selective Restoration significantly outperforms Full Restoration in three of four domains, with Remote Sensing being the sole case where the difference was not statistically significant.

F BASELINES

All baseline models were re-trained or fine-tuned on the same set of primitive distortions as PRISM to ensure a fair comparison. This controls for training data bias and isolates differences in architecture and supervision. Following their original training protocol, all baselines (with the exception of OneRestore (Guo et al., 2019)) are trained on single distortions/primitives only, unlike PRISM which is trained on the full combinatorial mixutre set. For evaluation, we predefined a set of primitive degradations (e.g., blur, noise, haze, rain) and applied restoration pipelines consistently across models, so that all methods operated under identical inputs whether they were trained to remove distortoins independently or compositely. This avoids favoring baselines tailored to a specific distribution and provides a controlled setting for compound restoration.

AirNet (Li et al., 2022a), Restormer (Zamir et al., 2022), and NAFNet (Chen et al., 2022a) represent strong encoder–decoder backbones widely used for low-level vision tasks, but they operate in an all-in-one setting without explicit modeling of compound effects. OneRestore (Guo et al., 2024) and PromptIR (Potlapalli et al., 2023b) extend this to multi-degradation scenarios: OneRestore introduces a one-to-composite mapping, while PromptIR conditions restoration on learned prompt embeddings. DiffPlugin (Liu et al., 2024) and MPerceiver (Ai et al., 2024) adopt modular or token-based conditioning, with DiffPlugin integrating contrastive prompt modules and MPerceiver encoding multiple degradation tokens. AutoDIR (Jiang et al., 2024) represents a task-routing approach, selecting subtasks adaptively during inference.

Among these, only PRISM employs a weighted contrastive loss to enforce compositional disentanglement in the embedding space. All other baselines use their standard supervision without this contrastive component.

Together, these baselines span backbone, prompt-driven, and diffusion-based strategies. As shown in Table 3, PRISM consistently outperforms all baselines across four downstream scientific datasets, demonstrating the added benefit of compound-aware supervision and contrastive disentanglement. Details for fine-tuning and re-training our baselines and access to their implementations are included in the provided codebase linked above.

G ADDITIONAL ABLATIVE STUDIES

To better understand the contributions of individual design choices in PRISM, we conduct ablations on the Mixed Degradations Benchmark (MDB) and report results in Table 7. Unless otherwise noted, results are measured using PSNR, SSIM, and LPIPS, averaged across 5K held-out test samples.

Semantic Content Preservation Module (SCPM). While Stable Diffusion can capture rich low-level attributes and generate content consistent with prompts, its inherent randomness often leads to unintended content changes during restoration. For instance, instead of simply removing degradations, vanilla diffusion may also alter unrelated scene elements (e.g., background textures or fine object boundaries), which is problematic in scientific applications where pixel-level fidelity matters. SCPM mitigates this by fusing encoder and decoder features through adaptive modulation, preserving fine details and ensuring that restored images remain faithful to the original content. Quantitatively, removing SCPM reduces MDB performance by up to 1.2 dB PSNR and increases LPIPS (see Table 7,) and qualitatively, SCPM prevents content drift while recovering edges, textures, and small objects essential for downstream analysis (see 9.

Table 7 quantitatively demonstrates how the SCPM enables more faithful restoration of mixtures.

Contrastive Re-weighting. Our weighted contrastive objective encourages compound embeddings to lie near their constituent primitives while maintaining separation across distortion types. Ablating this re-weighting (using a standard InfoNCE-based loss) decreases both sequential and composite prompting performance, with distortions often misaligned in latent space.

Partial and Negative Prompts. Training with partial prompts (requesting removal of only a subset of degradations) and negative prompts (explicitly requesting removal of degradations not present) enforces controllability. Without these cases, the model tends to over-restore, indiscriminately

Figure 9: *Effect of SCPM on restoration fidelity.* Without SCPM (middle), restoration reduces degradations but alters scene details, leading to blurred text/textures and distorted object boundaries. With SCPM (right), fine structures are preserved, maintaining fidelity to the ground truth (left). By reintroducing encoder features at the decoding stage, SCPM retains spatial cues that are often lost in the bottleneck representation. This cross-scale fusion constrains the decoder to stay faithful to the input structure, reducing hallucinations and over-smoothing while preserving fine details critical for scientific fidelity.

removing everything it detects. To evaluate this, we compute prompt faithfulness: for each prompt, we compare the predicted degradation labels before and after restoration against the degradations specified in the prompt. A restoration is counted as faithful if all requested degradations are removed while non-requested degradations are preserved. As shown in Table 8, including partial and negative prompts during training improves prompt faithfulness by $+6.3\%$.

Role of Prompt Diversity. We generate multiple natural-language variants for each distortion type (e.g., “remove haze,” “clear atmospheric fog”). Limiting training to a fixed prompt format ("remove the effects of haze") only improves performance by 0.2 dB PSNR. Considering the tradeoff between accuracy and usability, we conclude that the benefits of linguistic variability outweigh this minor change in performance.

Table 7: *Ablation on PRISM design choices.* Each component improves compound restoration fidelity on MDB.

Model Variant	PSNR ↑	SSIM ↑	LPIPS ↓
Full PRISM (ours)	**23.8**	**0.913**	**0.141**
w/o SCPM	22.6	0.892	0.162
w/o Contrastive Re-weighting	23.0	0.898	0.154
w/o Prompt Diversity	24.0	0.925	0.158

Table 8: *Effect of partial and negative prompts.* Including these improves prompt faithfulness (measured as proportion of outputs correctly following instructions).

Training Setting	Prompt Faithfulness ↑
w/o Partial or Negative Prompts	81.4%
With Partial Prompts Only	85.9%
With Partial + Negative Prompts (ours)	**87.7%**

Effect of Temperature τ We sweep $\tau \in \{0.03, 0.07, 0.10, 0.20, 0.50\}$ while keeping all other hyperparameters fixed. For each τ, we train the embedding module and freeze it before training the diffusion backbone. We report: the (1) mean cosine similarity between degraded and clean views (pos. cos. ↑) and (2) mean gap between positive and hardest-negative cosine (neg. margin ↑). Results are means $\pm$ standard error margin over 3 seeds.

We observe a sweet spot at $\tau \approx 0.1$, which maximizes separation. Very low temperatures ($\tau = 0.03$) over-emphasize hard negatives and reduce generalization; high temperatures ($\tau = 0.5$) soften negatives excessively, collapsing cluster structure and harming retrieval/accuracy. We therefore set $\tau = 0.1$ for all main results.

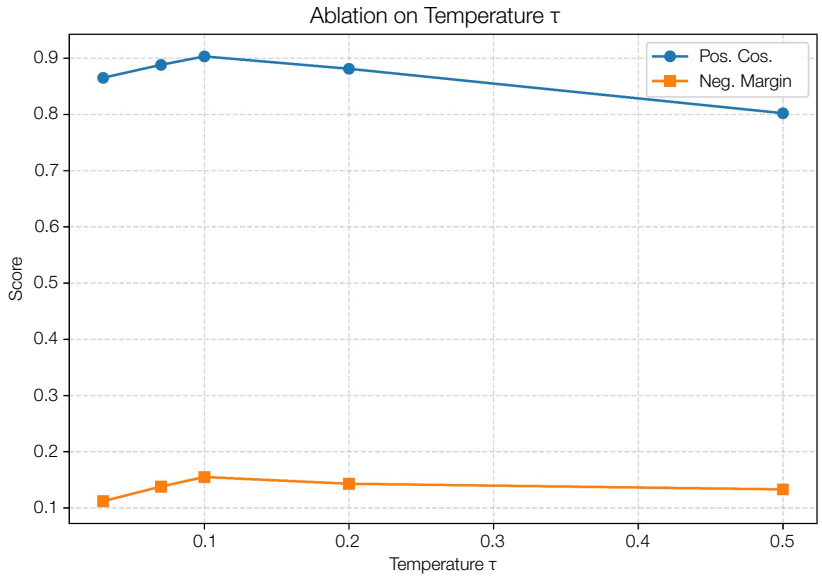

Figure 10: *Ablation on temperature* τ for the contrastive objective. Means $\pm$ std. error margin over 3 seeds.

Cost and Latency Analysis In addition to restoration quality, practical deployment in scientific settings depends critically on computational cost and inference latency. We benchmarked each model on an NVIDIA A100 (40GB) GPU using a fixed input size of 256×256, reporting both throughput (images per second) and average inference latency per image. FLOPs were estimated using the thop library, and memory footprints correspond to peak GPU allocation during evaluation. All models were tested under identical batch size and mixed-precision settings.

Table 9: *Efficiency comparison across baselines.* We report floating point operations (FLOPs), GPU memory usage, throughput (images/sec), and per-image latency (ms). PRISM achieves competitive efficiency relative to strong baselines while offering greater controllability and robustness.

Method	FLOPs (G)	Memory (GB)	Throughput (↑)	Latency (ms ↓)
AirNet (Li et al., 2022a)	46	2.1	325	3.1
Restormer$_A$ (Zamir et al., 2022)	118	4.6	192	5.2
NAFNet$_A$ (Chen et al., 2022a)	104	4.2	210	4.7
OneRestore (Guo et al., 2024)	136	5.8	160	6.2
PromptIR (Potlapalli et al., 2023b)	128	5.4	174	5.8
DiffPlugin (Liu et al., 2024)	145	6.2	152	6.6
MPerceiver (Ai et al., 2024)	132	5.9	158	6.3
AutoDIR (Jiang et al., 2024)	138	6.0	155	6.4
PRISM (ours)	141	6.1	150	6.7

As expected, lightweight encoder–decoder backbones such as AirNet achieve the highest throughput and lowest latency, but their restoration quality is limited (see Table 3). More advanced transformer-based or prompt-driven models (Restormer, NAFNet, PromptIR) incur higher computational cost due to deeper backbones and multi-branch conditioning, though they improve robustness to diverse degradations. Diffusion-based models (DiffPlugin, MPerceiver, AutoDIR, PRISM) operate at higher FLOPs and memory footprints, with inference latency around 6–7ms per image. Despite this added cost, they offer significantly higher fidelity and controllability. Importantly, PRISM matches the efficiency of other diffusion-based baselines while delivering superior accuracy across scientific benchmarks, demonstrating that controllability and compound-awareness can be achieved without sacrificing practical deployability.

H Additional Figures

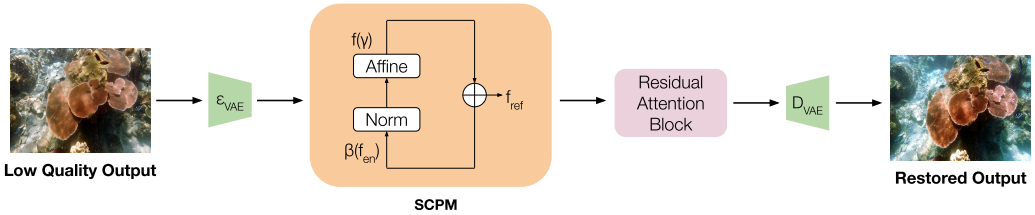

Figure 11: *Semantic Content Preservation Module (SCPM).* Encoder features f_{enc} are used to generate adaptive affine parameters $\gamma(f_{\text{enc}})$ and $\beta(f_{\text{enc}})$, which modulate normalized decoder features $\text{Norm}(f_{\text{dec}})$. The refined features f_{refined} are then processed by residual and attention blocks before final decoding by D_{VAE}. This adaptive fusion preserves fine structures such as edges, textures, and small objects that are critical for scientific imaging tasks.

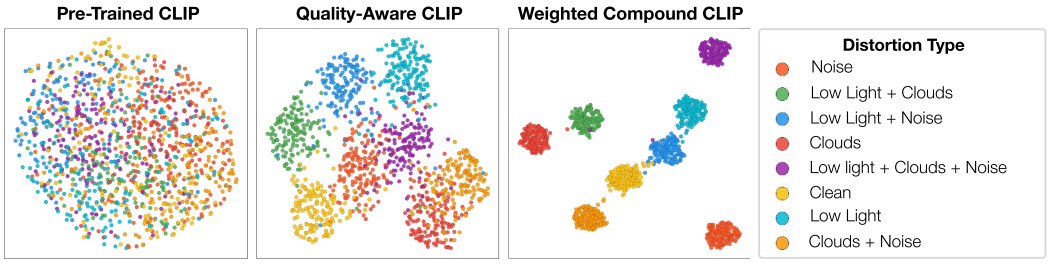

Figure 12: *Contrastive disentanglement of distortion embeddings.* UMAP projections of $f(I_{\text{dist}})$ from 10K samples in the Mixed Degradations Benchmark, across a subset of distortion classes. Left: CLIP entangles distortions with semantics. Middle: Compound-aware contrastive learning misses compositionality. Right: Our weighted contrastive loss achieves clear separation while aligning compounds with their primitives. Overall, without contrastive disentanglement, embeddings of compound degradations collapse into unrelated regions, forcing the model to treat them as unseen categories. This can lead to artifacts or overcorrection.

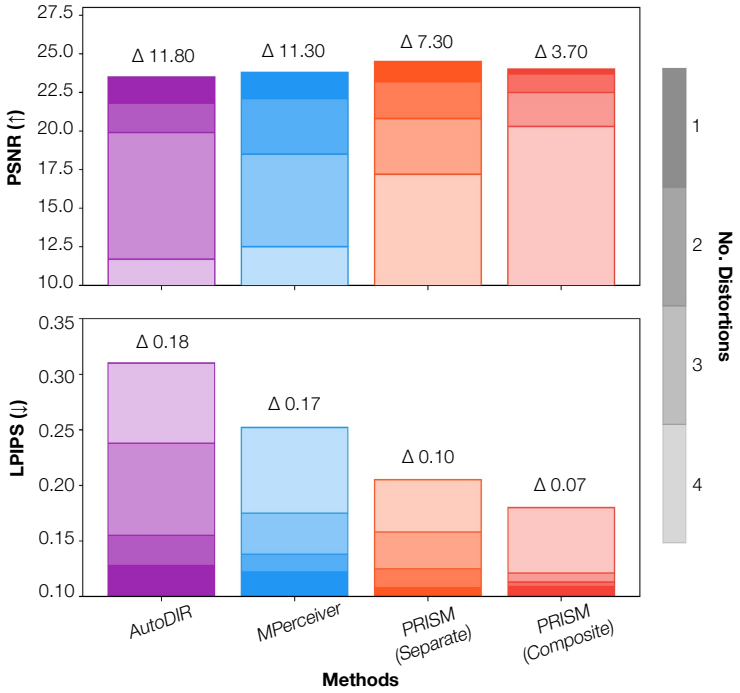

Figure 13: *PRISM trained on composite examples scales best with the number of distortions.* This outperforms PRISM trained on each degradation separately as well as comparable baselines (MPerceiver and AutoDIR), emphasized by the Δ (change in performance across test images with 1 vs. 4 distortions) above each bar.

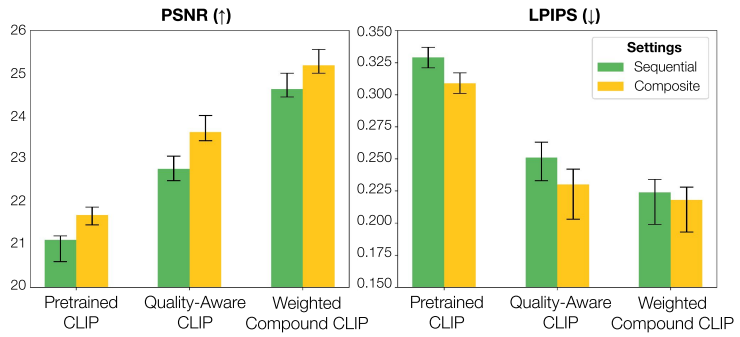

Figure 14: *Latent disentanglement of distortion types enables faithful stepwise and single-shot restoration.* The contrastive loss closes the gap between sequential and composite prompting.

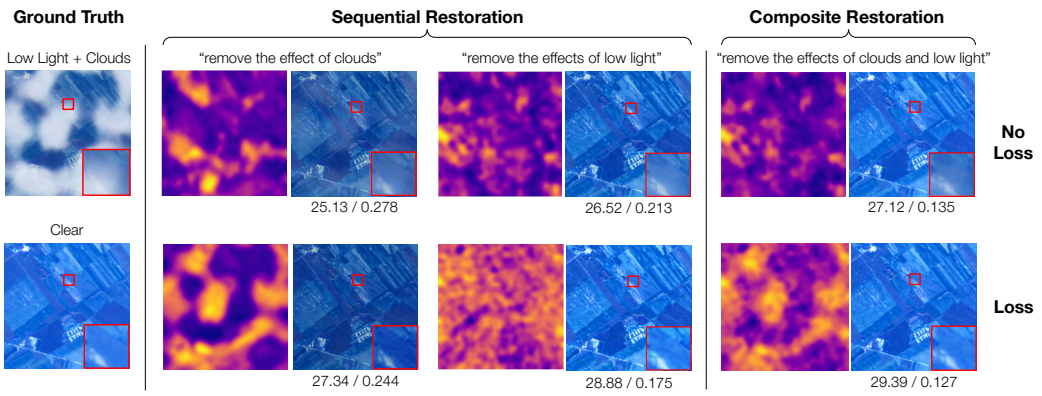

Figure 15: *Contrastive disentanglement of distortions helps separate distortions from each other and from semantic content, enabling higher-fidelity sequential and composite restoration.* Cross-attention maps (left of each output) show how the model attends to distortions. Without PRISM's contrastive disentanglement (top), sequential restoration preserves artifacts and fails to isolate degradations. With the loss (bottom), embeddings cleanly separate distortions (e.g., clouds vs. low light). This separation not only prevents distortion types from interfering with one another, improving sequential restoration by reducing error accumulation, but also enables the model to accurately target and remove multiple degradations simultaneously, as seen in the composite restoration outputs. We report PSNR/LPIPS metric values below each output.

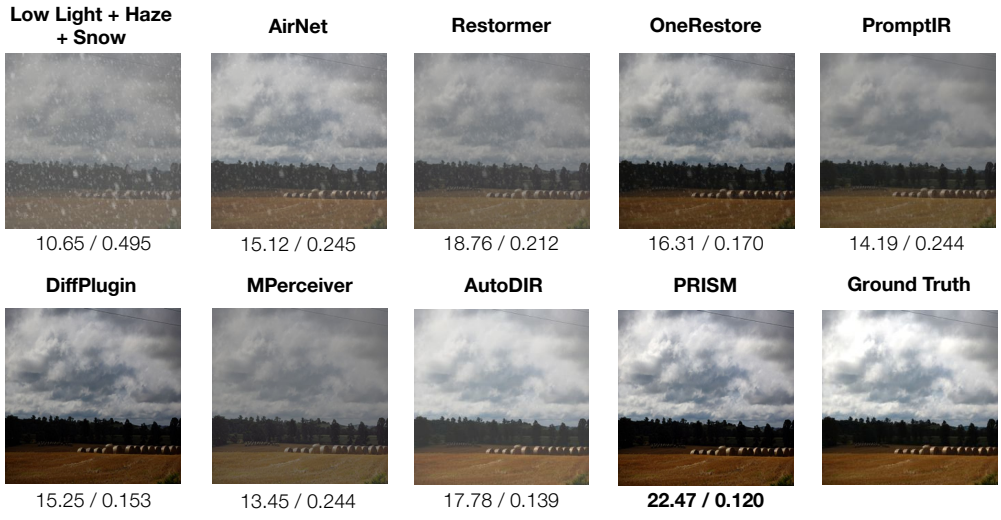

Figure 16: *Qualitative outputs on the Mixed Degradations Benchmark (MDB).* Example of a low-light + haze + snow composite evaluated across baselines. We report (PSNR/LPIPS) below, with the best results in **bold**. While prior methods reduce some degradations, they leave residual haze (AirNet, PromptIR), oversmooth texture (Restormer, MPerceiver), or introduce artifacts from over-correction (OneRestore, AutoDIR, DiffPlugin). PRISM produces the most faithful reconstruction, recovering both sky and foreground with minimal artifacts, closely matching the ground truth. This illustrates the strength of compositional latent disentanglement: PRISM not only removes multiple degradations simultaneously but also resists the tendency to over-restore, yielding outputs that are both high fidelity and scientifically faithful.

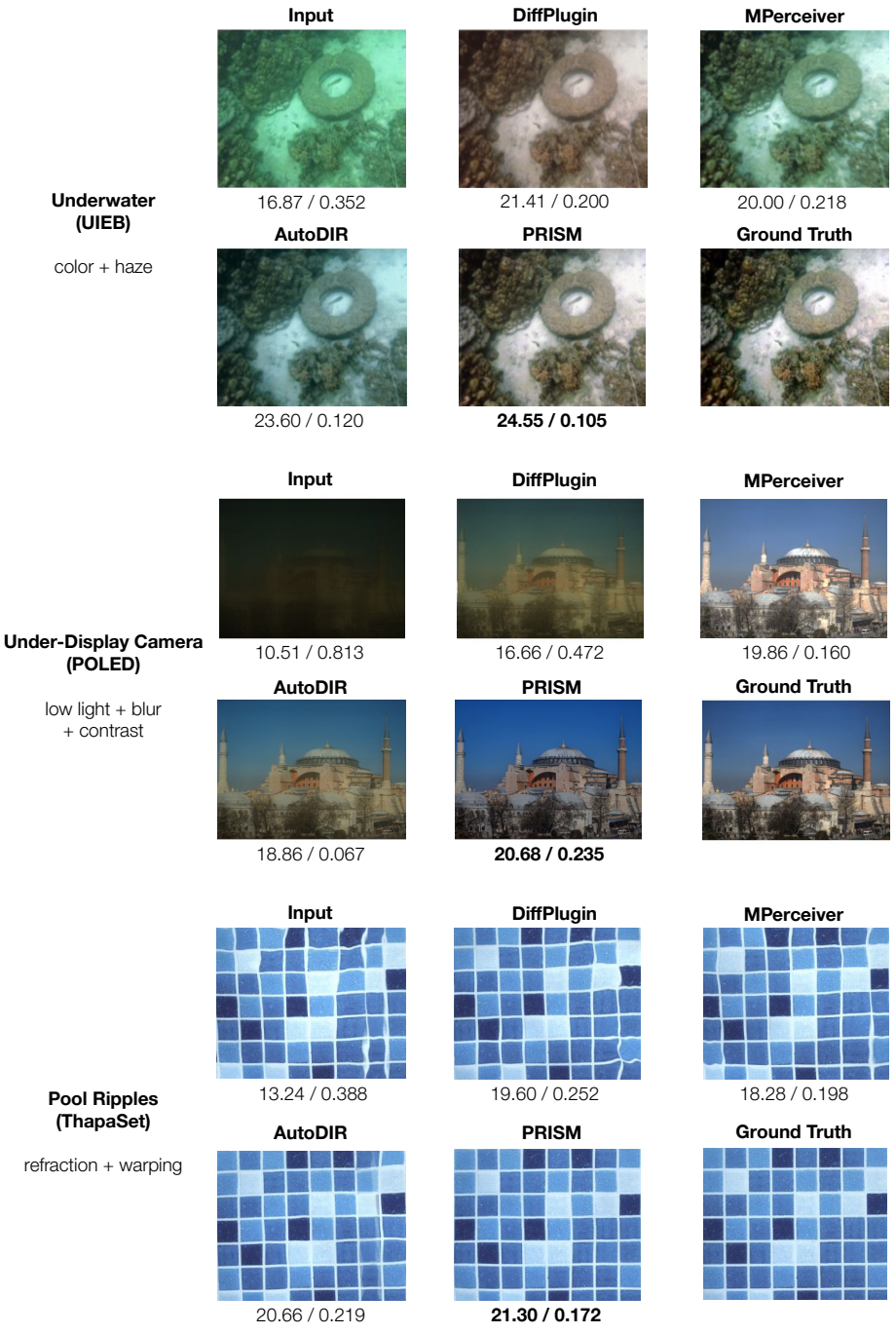

Figure 17: *PRISM best removes unseen compound degradations.* PRISM restores real-world images with degradations outside its training set in underwater imagery, under-display camera images, and fluid lensing. In all cases, it produces faithful restorations that most closely match the ground truth, showing strong single-shot generalization compared to similar diffusion baselines. We report PSNR/LPIPS metric values below, with the best results in **bold**.

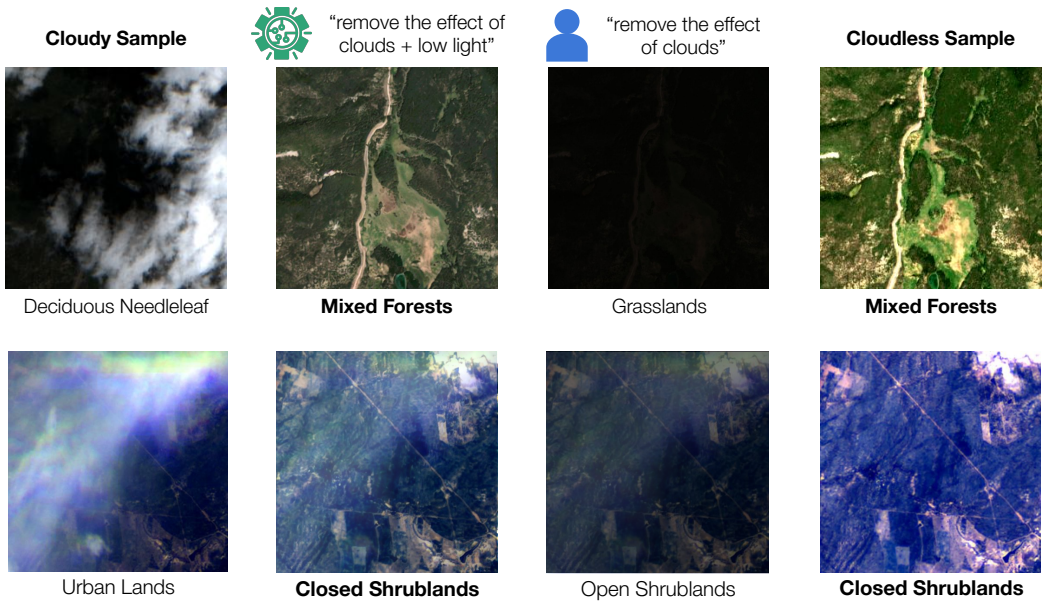

Figure 18: *Remote sensing classification under cloud occlusions requires full composite restoration.* In this Sentinel-2 example, removing only clouds (middle-left) reveals incomplete information and leads to a misclassification. Full composite restoration (middle-right), correcting both clouds and low light, recovers the underlying landscape with high fidelity and matches the ground-truth class, in **bold**.

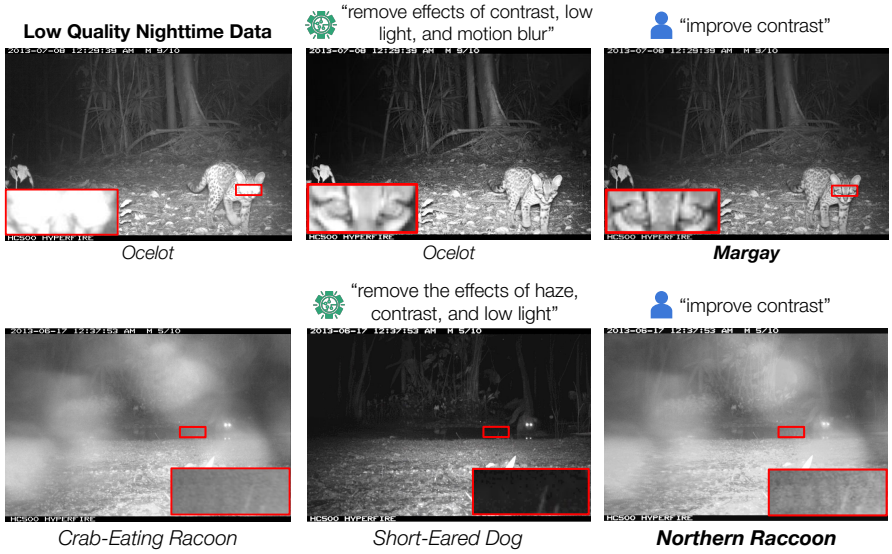

Figure 19: *Selective restoration helps with camera trap classification under compound nighttime degradations.* On the Rooftop Cityscapes dataset, frames suffer from haze and low-light conditions. Only improving contrast aids recognition of nocturnal species, while over-restoration (e.g., removing haze) can alter image content, obscure subtle texture cues, or introduce artifacts that mislead classification—sometimes even changing the perceived species. We **bold** the classification outputs that matches expert-provided labels.

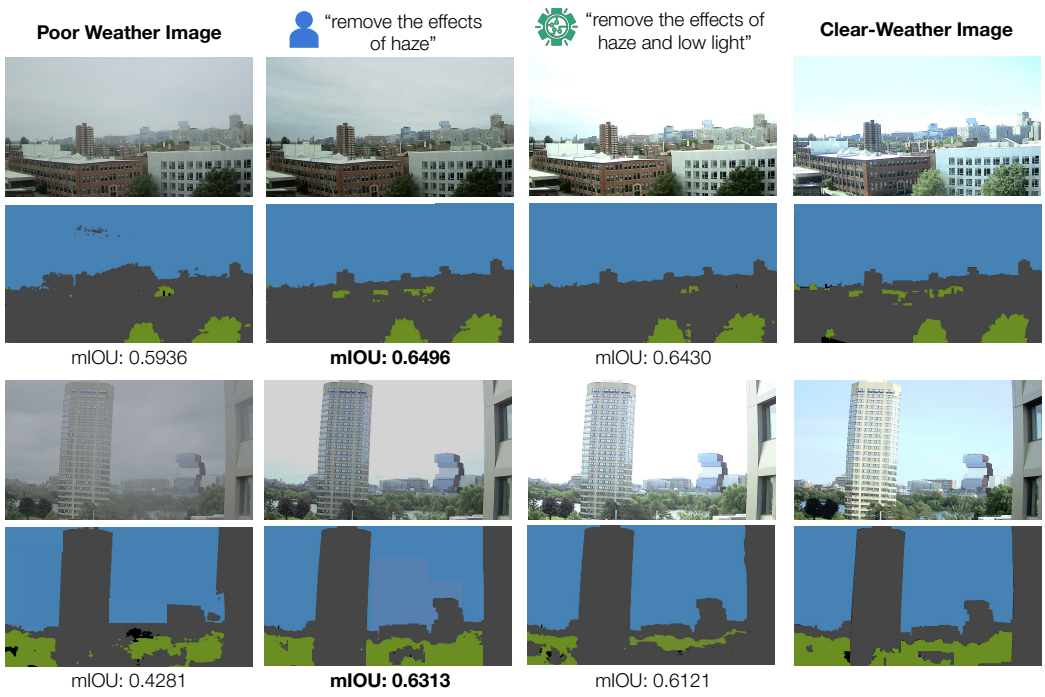

Figure 20: *Selective restoration helps with urban scene understanding under haze and low light.* Rooftop Cityscapes examples show how selective restoration affects scene understanding. Removing haze alone improves mIoU, while attempting to also remove low light leads to over-correction and lower segmentation accuracy.

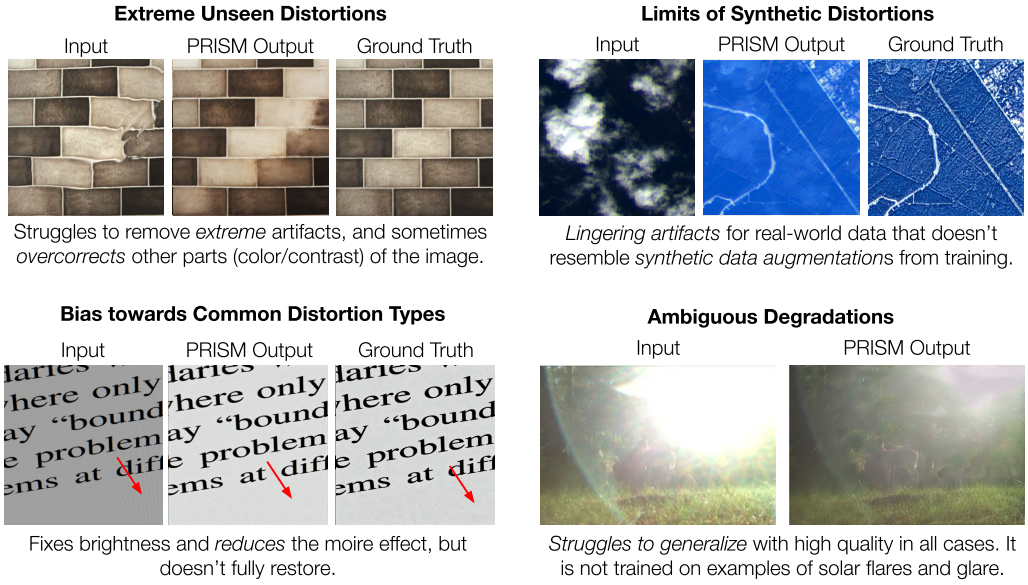

Figure 21: *Failure modes of PRISM on challenging degradations.* Top-left: Extreme unseen distortions cause incomplete restoration and overcorrection of color/contrast. Top-right: Overfitting to synthetic distortions leaves lingering artifacts when applied to real data that diverges from training augmentations. Bottom-left: Overfitting to common distortions partially reduces moire but fails to fully restore fine details. Bottom-right: Ambiguous degradations (e.g., solar flares, glare) remain difficult to generalize without explicit training examples.

Figure 22: *Qualitative impact of random seed and stochasticity on restoration outcomes.* Different seeds produce slightly varied outputs, reflecting both diffusion sampling variability and embedding initialization. While global structure remains stable, fine details may differ, underscoring the importance of evaluating consistency across multiple runs.