# LEARNING A SHARED REPRESENTATION FOR LOW-LEVEL VISION VIA DEGRADATION UNDERSTANDING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Low-level vision tasks such as image quality assessment (IQA) and restoration rely on common evidence about image perturbations, yet learned degradation representations are usually architecture- or task-specific. We propose Degradation Understanding as a supervisory task for learning a shared low-level representation. It predicts a hierarchical state over degradation type, composition, parameter keys, and continuous values, exposing a model-agnostic interface without assuming recovery of the true imaging history. To support this task, we construct DU-200K from 160,000 synthetic samples with structured supervision and 40,000 alignment samples for restoration-grounded feedback. We design a Degradation Understanding vision-language model (DU-VLM) that learns structured prediction via supervised fine-tuning (SFT) and outcome alignment via Group We evaluate the resulting representation on multiple downstream tasks. For IQA, we evaluate supervised target-dataset adaptation by training a separate approximately 1.05M-parameter prediction head with the training labels of each target dataset. Under this protocol, DU-GRPO increases the mean SRCC from 0.623 for Base Qwen to 0.809, a 29.9% relative improvement, and exceeds DeQA-Score by 0.033 absolute SRCC points, or 4.3% relatively, on average. On real-captured data, explicit degradation prompts predicted by our method improve PSNR by up to 1.74 dB over generic text prompts across state-of-the-art image-editing models. On entirely unseen and challenging real-world images, our system remains highly competitive among restoration systems based on diverse degradation-representation paradigms, supporting transfer across restoration interfaces, backbones, and real scenes.

## 1 INTRODUCTION

Image degradations arise from perturbations throughout the observation process, including optical transfer, environmental propagation, sensor response, compression, and compositions of these effects. While IQA estimates their perceptual consequences and restoration seeks an approximate inverse, both tasks depend on evidence about which perturbations are present and how strongly they affect the image. A representation that exposes this shared evidence could provide a common interface for both tasks.

Existing learning-based systems typically expose degradation information through interfaces optimized for a particular consumer. Degradation-aware restoration methods often learn conditioning embeddings that are coupled to one restoration architecture and are difficult to inspect or reuse Luo et al. (2024); Hu et al. (2026); Yang et al. (2025). Multimodal quality and restoration systems produce free-form descriptions, class labels, or scalar quality judgments Wu et al. (2024a); You et al. (2024); Li et al. (2025); Liu et al. (2026). Although effective in their intended settings, these task-specific outputs do not expose degradation evidence through a unified representation reusable across IQA and restoration systems.

We address this gap by reformulating Degradation Understanding as a structured supervisory task for shared low-level representation learning. The state specifies degradation type, composition, parameter keys, and continuous values in a tree-structured output. Predicting this state encourages the model to retain low-level factors shared by IQA and restoration, while the explicit output remains interpretable and model-agnostic. Our structured benchmark shows that both closed and open-weight



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

multimodal large language models (MLLMs) frequently violate the output schema or predict incompatible parameter keys, indicating that this state is not reliably elicited by prompting alone.

Real-world degradation states are difficult to annotate at scale because the underlying processes are not directly observable in ordinary image collections. We therefore use a sim-to-real route. Synthetic data provides explicit degradation-state supervision, while 40,000 alignment samples provide restoration-quality feedback for outcome alignment without requiring physical parameter labels. On this basis, we construct **DU-200K**. DU-200K combines a total of 160,000 synthetic samples with structured state supervision and 40,000 alignment samples for outcome alignment. The synthetic component is split by clean-image identity into 140,000 training, 4,000 validation, and 16,000 test examples. Only the 140,000-example training split is used for SFT. Validation and test examples are held out for model selection and final evaluation, respectively. DU-VLM learns structured state prediction through SFT and then uses GRPO on the alignment component for restoration-feedback alignment.

The predicted parameters provide an explicit interface for inverse operations and state-guided restoration, while the hidden states provide transferable conditioning features for IQA and restoration backbones. The downstream gains, rather than exact recovery of unobserved real imaging histories, are the principal evidence for the learned shared representation. Our contributions are summarized below:

(1) We formulate Degradation Understanding as a supervisory task for learning a shared low-level representation, using a tree-structured state over degradation type, composition, parameter keys, and continuous values.

(2) We introduce DU-200K, a benchmark with 160,000 synthetic samples for structured supervision and 40,000 alignment samples for outcome alignment.

(3) We show that a frozen DU-VLM supports supervised transfer to five IQA datasets through dataset-specific prediction heads, transfers through latent features to restoration conditioners, and transfers through explicit states to frozen image generators and unseen real-world scenes on CleanBench.

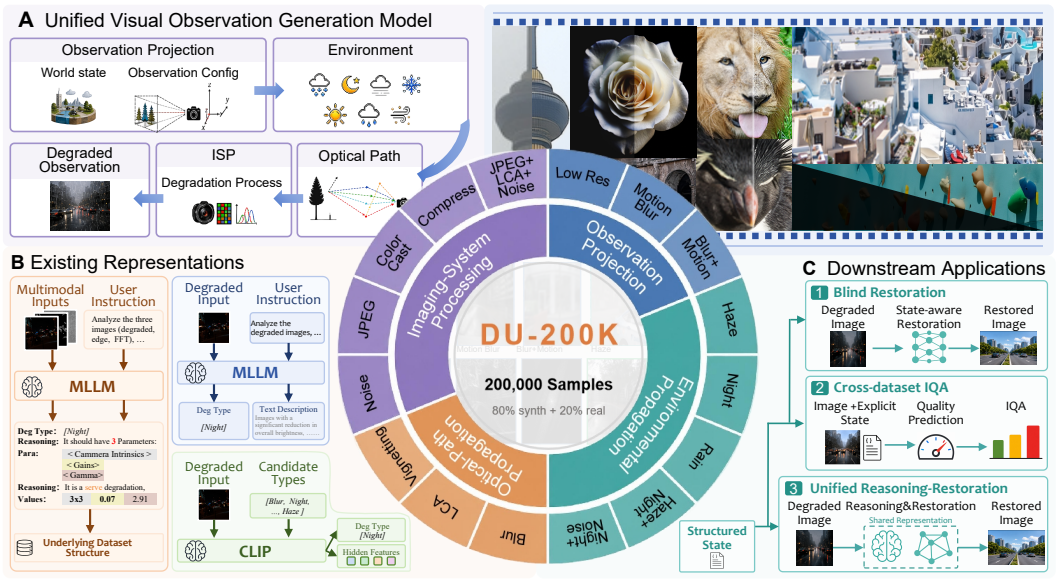



Figure 1: **Degradation Understanding vision-language model (DU-VLM) overview.** (A) Degradation Understanding 200K (DU-200K) construction. (B) Explicit hierarchical degradation-state prediction. (C) Transfer through explicit prompts and frozen features to restoration and image quality assessment.

## 2 RELATED WORK

**Degradation Understanding with Vision-Language Models.** VLMs mark a shift from pixel-level analysis to semantic perception (Wu et al., 2024b; Zhang et al., 2024). DA-CLIP (Luo et al., 2024) pioneered this direction by leveraging degradation-aware embeddings as restoration priors. Subsequent works, such as Q-bench (Zhang et al., 2024), Q-Instruct (Wu et al., 2024a), and DepictQA (You et al., 2024), established benchmarks and frameworks for perceptual quality assessment and free-form description. Recently, JarvisIR (Lin et al., 2025) further utilized VLMs as high-level planners to orchestrate restoration experts. These methods provide useful task-specific embeddings, quality judgments, or textual descriptions, but they do not train an explicit quantitative state whose internal representation is evaluated across both IQA and restoration. Our work treats structured degradation prediction as the common supervisory task for that transfer.

**Shared Representations for Low-Level Vision.** Unified restoration methods learn task-adaptive prompts or degradation-aware features within a single restoration architecture (Potlapalli et al., 2023; Luo et al., 2024), while large-scale pre-training has been used to learn common restoration representations across diverse corruptions (Li et al., 2024). UARE further unifies quality assessment and restoration in one vision-language framework (Li et al., 2025). Complementary to these end-to-end unified models, we ask whether Degradation Understanding can serve as a standalone supervisory task whose frozen representation transfers to separately instantiated IQA and restoration consumers.

**Universal Image Restoration.** While Single Degradation Image Restoration (SDIR) methods (Guo et al., 2017; Conde et al., 2024; Lan et al., 2025) excel in specialized domains, they struggle with complex, co-occurring degradations. This has necessitated All-in-One (AiOIR) frameworks (Jiang et al., 2024; Tian et al., 2025). Evolution in this field ranges from task-specific encoders (Li et al., 2020) to unified architectures (Li et al., 2022; Potlapalli et al., 2023) and Mixture-of-Experts (MoE) designs (Valanarasu et al., 2022; Jiang et al., 2024), which dynamically route inputs to specialized modules. Recent prompt-based (Potlapalli et al., 2023) and VLM-guided methods (Lin et al., 2025) introduce semantic control to enhance adaptability. Despite these advances, current methods often struggle with unseen degradation types and exhibit limited zero-shot generalization. “‘latex

## 3 PROBLEM FORMULATION AND BENCHMARK

### 3.1 HIERARCHICAL STRUCTURED PREDICTION

Degradation Understanding maps a degraded observation to a structured degradation state. Given a degraded image $I_d \in \mathbb{R}^{H\times W\times 3}$, the model predicts a state $z_{\text{deg}}$ that records which perturbation processes are present, how they are composed, which parameter keys govern each process, and what continuous values instantiate those keys. This gives downstream modules a model-agnostic output interface. IQA can estimate the perceptual consequence of the state, while restoration can condition or select an inverse process.

We represent the state as an ordered hierarchy rather than a flat label. A degradation description contains one or more process nodes,

$$\mathcal{D} = (d_1, \ldots, d_n), \qquad d_j = \left(t_j, \{(k_{j,\ell}, v_{j,\ell})\}_{\ell=1}^{m_j}\right), \tag{1}$$

where $t_j$ denotes the degradation type of process $j$, $k_{j,\ell}$ is a type-specific parameter key, and $v_{j,\ell}$ is its continuous value. The composition $c \in \mathcal{C}$ is encoded by the number and order of process nodes $(d_1, \ldots, d_n)$ rather than by a separate field. The hierarchy is defined below.

**Level 1. Degradation Type** ($t \in \mathcal{T}$). The type identifies the physical or signal-processing process, such as defocus blur, haze, compression, rain, or low-light corruption.

**Level 2. Parameter Keys** ($k \in \mathcal{K}_t$). Each type exposes a set of valid parameter names, such as defocus-blur kernel width, JPEG quality, haze transmission, noise scale, or a color-space transformation matrix.

**Level 3. Parameter Values** ($v \in \mathcal{V}_{t,k}$). Each key receives a numerical scalar, matrix, or spatial map that quantifies the degradation under the predefined parameterization.

**Level 4. Composition** ($c \in \mathcal{C}$). Composition is represented implicitly by the combination and order of the $t$–$k$–$v$ process groups. One group denotes a single degradation, whereas multiple ordered groups define a compound chain and its application order.

Accordingly, the model $f_\theta$ maps the visual input to the structured output below.

$$\hat{\mathcal{D}} = f_\theta(I_d) = (\hat{d}_1, \ldots, \hat{d}_{\hat{n}}), \hat{d}_j = \left(\hat{t}_j, \{(\hat{k}_{j,\ell}, \hat{v}_{j,\ell})\}_{\ell=1}^{\hat{m}_j}\right). \tag{2}$$

### 3.2 DU-200K Dataset

DU-200K is the benchmark used for training and evaluating DU-VLM. It contains 160,000 synthetic samples with structured degradation-state supervision and 40,000 alignment samples used for outcome alignment. We use the term *alignment samples* because the source datasets do not share one provenance: RESIDE contains both synthetic and real subsets, whereas DIV2K includes derived low-resolution counterparts. Thus, the 40,000 count denotes processed samples selected for alignment and does not imply that every item is an independently captured photograph.

The synthetic component is generated from clean images drawn from ImageNet, CelebA, nuScenes, and BDD100K (Deng et al., 2009; Liu et al., 2015; Caesar et al., 2020; Yu et al., 2020). It is split by clean-image identity into 140,000 training, 4,000 validation, and 16,000 test samples. SFT uses only the 140,000 training samples. All degraded variants of one source image remain in one partition. Validation is used only for model and checkpoint selection, and test data only for reporting. The component covers 16 named degradation classes derived from 12 base degradation operators grouped by the image-formation chain stage they affect. These stages are observation projection (low resolution, motion blur), environmental propagation (night degradation, haze, rain), optical-path physical propagation (defocus blur, lateral chromatic aberration, longitudinal chromatic aberration), and imaging-system processing (noise, JPEG compression, color cast, vignetting). The state annotation contains 20 physical parameter keys and includes both single degradations and compound compositions. Each sample provides the degraded RGB image together with edge and frequency-domain views, giving the model spatial and spectral evidence for the same degradation state.

The alignment component is drawn from selected subsets of RESIDE, GoPro, a deduplicated union of LOL-v1 and LOL-v2, and DIV2K (Li et al., 2019; Nah et al., 2017; Wei et al., 2018; Yang et al., 2021; Agustsson & Timofte, 2017). It mixes paired and unpaired samples and provides no degradation-parameter annotations. The counting unit, source roles, deduplication rule, split usage, and reward-branch definition are specified in Appendix A.1; the restoration reward is defined in Appendix A.5.

## 4 Methodology

### 4.1 Autoregressive Degradation Understanding

Degradation understanding decomposes into coupled predictions over an ordered sequence of degradation types, parameter keys, and continuous values. These targets live in different output spaces, so a direct multi-head formulation creates optimization conflicts and does not expose a single downstream interface.

We instead use next-token prediction (NTP) as a single sequential objective, following work that represents numerical prediction through token generation and analyzes the statistical behavior of next-token learning (Gruver et al., 2023; Song & Bahri, 2025; Yüksel & Flammarion, 2025). Under the quantized formulation used here, the model emits a token sequence that encodes the same structured state, thereby turning heterogeneous supervision into a homogeneous sequence modeling problem.

**Approximate NTP Decomposition.** **Given** the structured discrete state $s = (\boldsymbol{t}, c, \boldsymbol{k})$ defined above, a value vector $v \in \mathbb{R}^d$, and a coordinate-wise quantizer with bin widths $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_d)$.

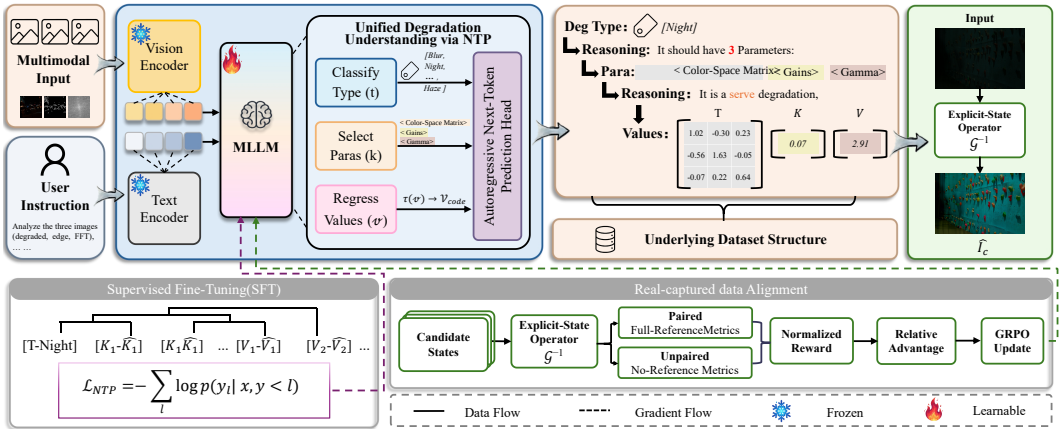



Figure 2: **DU-VLM training and inference.** RGB, edge, and frequency views predict a hierarchical degradation state. SFT uses synthetic state labels, and GRPO uses fixed restoration feedback. The $3 \times 3$ matrix in the low-light example represents a color-space transformation rather than geometric camera intrinsics.

**Assume** (i) the conditional value density is locally $\mathcal{N}(\mu_\theta(s, \boldsymbol{x}), \sigma^2 \mathbf{I})$, and (ii) each quantization cell is sufficiently small for the density to be approximately constant within it. Writing $\mu_\theta = \mu_\theta(s, \boldsymbol{x})$, we obtain

$$\mathcal{L}_{\text{NTP}} \approx \underbrace{-\log p_\theta(s \mid \boldsymbol{x})}_{\mathcal{L}_{\text{struct}}} + \underbrace{\frac{\|v - \mu_\theta\|_2^2}{2\sigma^2}}_{\mathcal{L}_{\text{value}}} + C_{\sigma,d} - \sum_{j=1}^{d} \log \Delta_j. \tag{3}$$

where the final two terms are independent of $\theta$.

This result gives a common token-level objective for a discrete structure and an arbitrary-dimensional continuous representation without claiming exact equality between token cross-entropy and continuous regression.

**Synthetic SFT.** We implement the structured predictor with Qwen3-VL-8B. The degraded image, auxiliary edge view, auxiliary frequency view, and user instruction are tokenised into the same context. For a degraded image $I_d$, the model uses NTP to autoregressively predict the ordered degradation state $\mathcal{D} = (d_1, \ldots, d_n)$ defined in Eq. equation 1. Composition is implicit in $\mathcal{D}$. The value of $n$ distinguishes single from compound degradation, and the order of its process nodes specifies the composition order.

### 4.2 Explicit-State Operator Composition

The explicit state predicted by DU-VLM deterministically compiles into a sequence of fixed restoration operators. Its type and composition select and order the operators, while its keys and values instantiate their parameters. The operator library and composition rules remain fixed throughout learning.

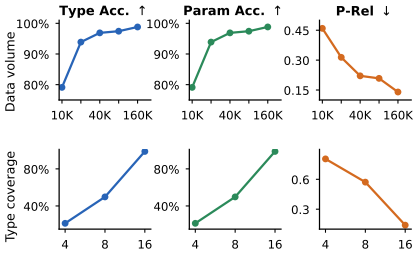
For the low-light restoration pipeline, the operations are applied in the following order: color-space mapping, gamma transformation, RGB gain, inverse gamma transformation, and inverse color-space mapping. Let $\mathcal{M}_{\text{cs}}$ denote the color-space mapping, $\mathcal{G}_\gamma$ the gamma transformation, and $\mathcal{R}_g$ the RGB-gain operation. The resulting transformation is

$$I_{\text{out}} = \mathcal{M}_{\text{cs}}^{-1}\left(\mathcal{G}_\gamma^{-1}\left(\mathcal{R}_g\left(\mathcal{G}_\gamma\left(\mathcal{M}_{\text{cs}}(I_{\text{in}})\right)\right)\right)\right). \tag{4}$$

These fixed routines are approximate restoration heuristics rather than exact inverses of all degradation processes. Their implementation details are provided in Appendix A.3.




Figure 3: **Failure modes in structured degradation understanding.** (A–C) Task-level hallucination, semantic bias, and parameter error. (D–E) Class-wise recall and row-normalized confusion matrices for generalist and DU-supervised models.




(a) DU-SFT scaling.

| Method | T-Acc ↑ | M-F1 ↑ | P-Acc ↑ | J-Acc ↑ | P-Abs ↓ | P-Rel ↓ |
|---|---|---|---|---|---|---|
| Qwen3-VL-8B | 3.46 | 3.98 | 4.02 | 0.14 | 9.19 | 1.00 |
| InternVL1.5-25.5B | 9.04 | 2.33 | 9.06 | 9.04 | 9.02 | 0.95 |
| InternVL3.5-30B-A3B | 13.69 | 10.72 | 13.75 | 13.69 | 9.07 | 1.02 |
| Qwen3-VL-30B-A3B | 28.52 | 24.00 | 28.90 | 28.52 | 8.56 | 0.93 |
| Gemini-3.5-Flash | 56.12 | 53.82 | 56.19 | 56.12 | 8.41 | 0.75 |
| GLM-5V-Turbo | 25.87 | 20.26 | 25.99 | 25.87 | 8.78 | 0.98 |
| GPT-5.6-sol | 42.90 | 37.92 | 42.86 | 42.86 | 9.02 | 0.84 |
| Grok-4.5 | 36.90 | 31.84 | 36.88 | 36.88 | 9.36 | 0.91 |
| DU-SFT | **98.81** | **98.83** | **98.82** | **98.81** | **1.15** | **0.14** |
| DU-GRPO | 91.08 | 91.13 | 86.69 | 86.66 | 1.34 | 0.28 |

(b) Synthetic schema-compliance diagnostic; this panel motivates the structured supervision used in DU-200K.

Figure 4: **Structured prediction and DU-SFT scaling.** Panel (a) varies data volume (top) and taxonomy coverage (bottom). Its 160K/16-type endpoint is the DU-SFT row in (b). Bold and underline in (b) mark the best and second-best results.

### 4.3 Alignment-Data Optimization via GRPO

Synthetic SFT teaches the vocabulary and structure of the degradation state, but the alignment samples do not provide degradation-parameter labels. We therefore optimize DU-VLM with the standard GRPO procedure, using restoration quality as the sole reward. Each prompt produces a group of $G = 8$ completions with rewards $R_i = r_{\mathrm{rec},i}$. The implementation uses sample-standard-deviation reward scaling, clipped token ratios, and a clipped k3-style reference-policy penalty. The complete objective and the 24-GPU continuation configuration are specified in Appendices A.6 and A.7.

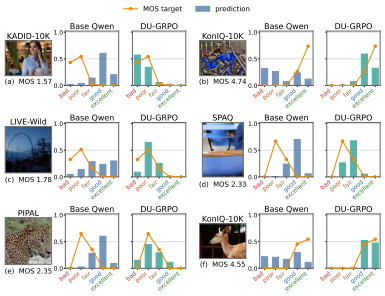
The reward $r_{\mathrm{rec}}$ combines aesthetic quality with MSE when a clean target is available and uses only the aesthetic score otherwise. GRPO updates only DU-VLM. Exact metric definitions, normalizations, and sampling settings are provided in the Technical Supplement.

## 5 Experiments

### 5.1 Experimental Protocol

**Baselines.** *Structured Degradation Understanding.* We compare with open-weight Qwen3-VL-8B, Qwen3-VL-30B-A3B, InternVL1.5-25.5B, and InternVL3.5-30B-A3B, and with commercial models Gemini-3.5-Flash, GLM-5V-Turbo, GPT-5.6-sol, and Grok-4.5. Each model receives the same multimodal inputs and is evaluated with the same DU output schema and parser. Generalist MLLMs receive the schema as a prompt template, whereas DU-trained models emit that schema natively. The input views and scored output fields are otherwise identical.

*Quality assessment.* We compare with the traditional no-reference methods NIQE and BRISQUE, the deep IQA methods NIMA, HyperIQA, DBCNN, MUSIQ, and MANIQA, and the vision-language methods CLIP-IQA+, Compare2Score, Q-Align, and DeQA-Score.




(a)

| Method | KADID-10K | KonIQ-10K | LIVE-Wild | PIPAL | SPAQ | Avg |
|---|---|---|---|---|---|---|
| NIQE | 0.468 / 0.405 | 0.533 / 0.530 | 0.493 / 0.449 | 0.195 / 0.161 | 0.679 / 0.664 | 0.474 / 0.442 |
| BRISQUE | 0.429 / 0.356 | 0.225 / 0.226 | 0.361 / 0.313 | 0.267 / 0.232 | 0.490 / 0.406 | 0.354 / 0.307 |
| NIMA | 0.532 / 0.535 | 0.896 / 0.859 | 0.814 / 0.771 | 0.390 / 0.399 | 0.838 / 0.856 | 0.694 / 0.684 |
| HyperIQA | 0.506 / 0.468 | 0.917 / 0.906 | 0.772 / 0.749 | 0.410 / 0.403 | 0.791 / 0.788 | 0.679 / 0.663 |
| DBCNN | 0.497 / 0.484 | 0.884 / 0.875 | 0.773 / 0.755 | 0.384 / 0.381 | 0.812 / 0.806 | 0.670 / 0.660 |
| MUSIQ | 0.575 / 0.556 | 0.924 / 0.929 | 0.789 / 0.830 | 0.431 / 0.431 | 0.868 / 0.863 | 0.717 / 0.722 |
| CLIP-IQA+ | 0.653 / 0.654 | 0.909 / 0.895 | 0.832 / 0.805 | 0.427 / 0.419 | 0.866 / 0.864 | 0.737 / 0.727 |
| MANIQA | 0.499 / 0.465 | 0.849 / 0.834 | 0.849 / 0.832 | 0.457 / 0.452 | 0.768 / 0.758 | 0.684 / 0.668 |
| Compare2Score | 0.500 / 0.453 | 0.923 / 0.910 | 0.786 / 0.772 | 0.354 / 0.342 | 0.867 / 0.860 | 0.686 / 0.667 |
| Q-Align | 0.674 / 0.684 | 0.941 / 0.940 | 0.853 / 0.860 | 0.403 / 0.419 | 0.886 / 0.887 | 0.751 / 0.758 |
| DeQA-Score | 0.694 / 0.687 | **0.953 / 0.941** | **0.892 / 0.879** | 0.472 / 0.478 | 0.895 / 0.896 | 0.781 / 0.776 |
| Base Qwen | 0.393 / 0.365 | 0.743 / 0.640 | 0.803 / 0.670 | **0.525 / 0.540** | 0.909 / 0.900 | 0.674 / 0.623 |
| DU-SFT | 0.565 / 0.564 | 0.813 / 0.745 | 0.822 / 0.734 | 0.492 / 0.477 | 0.917 / 0.911 | 0.722 / 0.686 |
| DU-GRPO | **0.883 / 0.882** | 0.919 / 0.899 | 0.886 / 0.845 | 0.503 / 0.494 | **0.931 / 0.926** | **0.824 / 0.809** |

(b)

Figure 5: (a) **IQA transfer.** Representative quality predictions. (b) PLCC/SRCC under target-dataset supervised adaptation on five IQA datasets. Avg is the unweighted mean. Bold and underline mark the best and second-best values under the same evaluation protocol.

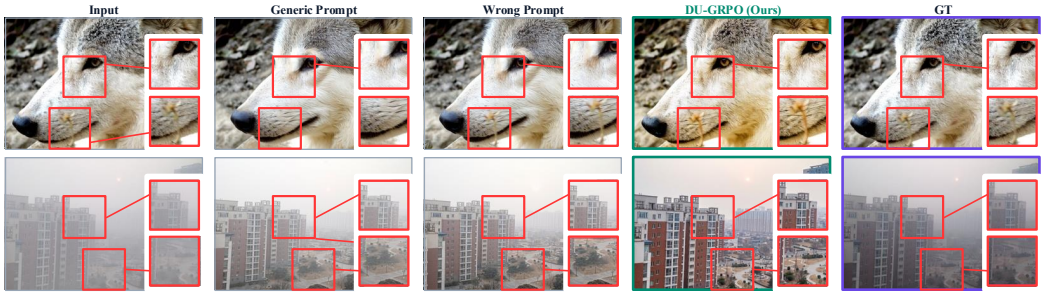



Figure 6: **Explicit degradation-state prompting.** The top and bottom rows use Qwen-Image-Edit and HunyuanImage-3 as the restoration backbones. The columns show the degraded input, results obtained with a generic prompt, a mismatched DU state, and the predicted DU-GRPO state, followed by the reference image from the paired test set. Insets highlight local details.

*Restoration.* We test two interfaces to frozen restoration models. For hidden-state transfer, DiffBIR provides a matched comparison across feature conditioners. For explicit-state transfer, Qwen-Image-Edit and HunyuanImage-3 compare a generic restoration prompt with a DU-derived state prompt while keeping the generator weights fixed.

**Metrics.** For direct understanding, exact-format validity, type accuracy (T-Acc), macro F1 (M-F1), parameter accuracy (P-Acc), and joint accuracy (J-Acc) measure structural correctness. Penalized absolute and relative errors (P-Abs/P-Rel) assign the zero-prediction penalty to missing or type-incompatible values. IQA uses the Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SRCC). Restoration uses PSNR, SSIM, LPIPS, Fréchet inception distance (FID), and CLIP-IQA. Comparisons are made only within a shared protocol and backbone.

### 5.2 Degradation Understanding Analysis

**Overall structured prediction.** Figure 4b reports a synthetic schema-compliance diagnostic on the 16-class common test. All models receive the same output schema and parser. General-purpose MLLMs frequently produce invalid structures or incompatible parameter keys, whereas DU-SFT learns to follow the prescribed representation. These results justify constructing DU-200K with explicit structured supervision. They are not used as evidence of real-world degradation recovery, nor as the main evidence for downstream transfer.

**Synthetic-to-restoration trade-off.** DU-GRPO trades synthetic-label accuracy for restoration-grounded alignment, whereas DU-SFT preserves the high-precision structured vocabulary. Downstream stage attribution therefore compares the corresponding checkpoints. CleanBench separately evaluates the final DU-GRPO system.

**SFT scaling.** Figure 4a varies data volume and taxonomy coverage using the same train/validation/test proportions as the full 140,000/4,000/16,000 split. Both improve structural ac-




Figure 7: **Unseen real-world qualitative comparison.** Representative outputs from the same testset in Table 3 are shown for the input, DA-CLIP, JarvisIR, and DU-GRPO.

curacy and reduce parameter error, with taxonomy coverage producing the larger controlled change at a fixed 140K training budget.

**Input-view ablation.** Table 1 shows that removing either auxiliary view degrades both metrics, and RGB-only input is weakest. This trend supports complementary spatial and spectral evidence without treating token serialization as a reasoning mechanism.

Table 1: **Input-view ablation.** Bold and underline mark the best and second-best results.

| Metric | Full | w/o FFT | w/o Edge | RGB only |
|---|---|---|---|---|
| J-Acc (↑) | **98.81%** | <u>97.6%</u> | 96.2% | 91.4% |
| P-Rel (↓) | **0.14** | <u>0.183</u> | 0.191 | 0.207 |

### 5.3 Transfer to IQA

We train a separate prediction head on the labeled training partition of each target IQA dataset while keeping the representation model frozen. The datasets are KADID-10K, KonIQ-10K, LIVE-Wild, PIPAL, and SPAQ. The splits and score-distribution processing follow DeQA-Score (You et al., 2025).

This protocol evaluates supervised adaptation to each target dataset. Target dataset labels are used to train the prediction head; therefore, these results do not represent cross-dataset generalization. Cross-dataset generalization requires training on a specified source dataset and evaluating on target datasets without using their training labels. We do not interpret the results in Figure 5b as cross-dataset generalization.

**SFT-to-GRPO transfer.** GRPO improves DU-SFT on all five datasets, with the largest gain on KADID-10K. Gains on the four authentic-distortion datasets show that restoration-grounded alignment is not limited to synthetic supervision.

### 5.4 Shared Low-Level Representation Transfer

The final study tests two interfaces to frozen generation models. Hidden-state conditioning asks whether a DU-supervised internal representation can support a trainable restoration adapter, whereas text-state prompting asks whether the explicit predicted state can guide a multimodal image generator without changing its weights.

The hand-crafted restoration operators used to form the alignment reward are not applied to, or provided as inputs to, the downstream multimodal generators. For the DU-GRPO conditions in Table 2 and Table 3, the DU-GRPO output is passed to a frozen multimodal generator as a hidden-state conditioner or an explicit text prompt; the reported scores evaluate the final generator outputs. Because some image-quality metrics overlap with the reward, these results should be interpreted as transfer evidence under this protocol rather than as a metric-independent proof of degradation understanding.

**Hidden-state conditioning.** Under the matched DiffBIR protocol in Table 2, DU-GRPO hidden states give the strongest fidelity results, while Base Qwen leads CLIP-IQA. Because only the conditioner changes, this comparison isolates the transferred representation.

**Text-state prompting.** On the same 200 paired held-out inputs, which are disjoint from the alignment data by source identifier and image hash, we evaluate explicit-state prompting with fixed Qwen-Image-Edit and HunyuanImage-3 backbones. The Mismatched DU condition follows a three-level hierarchical construction: the degradation type is incorrect at the upper level; at the next level, the degradation type is correct but the parameters are incorrect; the complete DU-GRPO state has both the correct degradation type and parameters. We retain this construction as a single control condition in Table 2.

**State-component ablation.** To further examine the contribution of quantitative degradation information, we derive all ablations from the same DU-GRPO prediction for each input. The Type-only condition retains only the predicted degradation type. The coarse-severity condition replaces exact values with three severity levels. The random-legal-parameter condition retains the same degradation type and parameter keys but replaces the predicted values. Shuffled-value conditions preserve the type, parameter keys, and process order while permuting the values. Shuffled-order conditions preserve each complete degradation process while changing the order of compound processes. All conditions use the same inputs, backbone, prompt template, decoding settings, and random seeds.

Table 2: **Restoration transfer through hidden states and explicit text.** Rows are comparable only within each backbone group. Best and second-best results within each group are marked in **bold** and underline, respectively.

| Condition | PSNR ↑ | SSIM ↑ | LPIPS ↓ | FID ↓ | CLIP-IQA ↑ |
|---|---|---|---|---|---|
| *DiffBIR (hidden-state conditioner)* | | | | | |
| Base Qwen | 20.906 | 0.688 | 0.201 | 59.117 | **0.412** |
| DepictQA | 18.506 | 0.590 | 0.269 | 65.052 | 0.397 |
| DU-SFT | 21.206 | 0.697 | 0.212 | 58.112 | 0.365 |
| DU-GRPO | **22.249** | **0.723** | **0.196** | **57.623** | 0.384 |
| *Qwen-Image-Edit (text prompt)* | | | | | |
| Generic | 17.214 | 0.584 | 0.354 | 63.149 | **0.509** |
| Mismatched DU | 16.195 | 0.585 | 0.418 | 83.428 | 0.364 |
| DU-SFT | 16.230 | 0.604 | 0.402 | 90.291 | 0.345 |
| DU-GRPO | **18.949** | **0.649** | **0.292** | **53.633** | 0.464 |
| *HunyuanImage-3 (text prompt)* | | | | | |
| Generic | 15.632 | 0.637 | 0.285 | 51.847 | **0.516** |
| Mismatched DU | 16.200 | 0.571 | 0.307 | 49.673 | 0.504 |
| DU-SFT | 15.641 | 0.563 | 0.325 | 54.350 | 0.489 |
| DU-GRPO | **17.143** | **0.654** | **0.259** | **46.053** | 0.481 |

### 5.5 Unseen Real-World Evaluation

We evaluate system-level generalization on CleanBench (Lin et al., 2025), whose real-world images are excluded from structured supervision, alignment data, and all other training stages. We exclude variants derived from these images and remove source-level and image-hash overlap before sampling 200 official test images. We evaluate every method on the same $512 \times 512$ inputs and rerun all baselines. DU-GRPO conditions frozen HunyuanImage-3 on its predicted degradation parameters.

Table 3: **Unseen real-world evaluation.** Best and second-best results are marked in **bold** and underline.

| Method | NIQE ↓ | BRISQUE ↓ | CLIP-IQA ↑ | MUSIQ ↑ |
|---|---|---|---|---|
| *Input* | 3.3288 | 15.9226 | 0.3233 | 61.5763 |
| DA-CLIP | 5.6650 | 37.7601 | 0.3501 | 43.3463 |
| TAO | 5.3316 | 34.0366 | 0.2950 | 35.0166 |
| GenDeg | 9.4595 | 54.1586 | 0.3290 | 35.2143 |
| JarvisIR VLM | 6.6577 | 40.5421 | **0.4474** | 53.4338 |
| DFPIR | 5.6899 | 37.3546 | 0.3668 | 43.0313 |
| **DU-SFT** | 3.4237 | 14.0985 | 0.3922 | 62.8714 |
| **DU-GRPO** | **3.0814** | **12.6701** | 0.4219 | **69.1243** |

Table 3 and Figure 7 use the identical subset. DU-GRPO leads three of four no-reference metrics and ranks second on CLIP-IQA, supporting generalization to unseen real-world scenes. This evaluates the final system rather than attributing the gain specifically to GRPO.

## 6 Limitations

Without real-world labels, we use fixed handcrafted operators' restoration metrics as physical proxies. Multimodal transfer (§5.4–5.5) shows useful observation–degradation correlations, not physical identity. We will label real-camera degradations to study physical intelligence.

## 7 Conclusion

DU-200K and DU-VLM learn shared degradation representations through structured supervision and restoration-grounded alignment. Target-dataset IQA adaptation and transfer to frozen multimodal generators demonstrate utility beyond handcrafted operators, supporting Degradation Understanding as a transferable interface for low-level vision.

## 8 REPRODUCIBILITY STATEMENT

The supplement specifies the composition and partitioning of DU-200K, including the 140,000/4,000/16,000 split for synthetic data and the 40,000 real-captured samples used for restoration-grounded alignment. It also provides the structured output schema, degradation parameter definitions, fixed restoration operator routines, GRPO objective, training configuration, feature-extraction procedure, IQA-head architecture, and dataset-specific evaluation protocol. The supplement records the checkpoint roles, source identifiers, image-hash checks, held-out evaluation subsets, and aggregation rules used for the reported results. All IQA and restoration comparisons use the stated data partitions, frozen backbones, and evaluation metrics. The authors are responsible for maintaining the corresponding code, configuration files, and evaluation records required to reproduce the reported tables and figures.

## 9 ETHICS STATEMENT

This work uses public image datasets and established restoration benchmarks, including ImageNet, CelebA, nuScenes, BDD100K, RESIDE, GoPro, LOL-v1, LOL-v2, DIV2K, and CleanBench. These datasets may contain people, identifying visual details, or copyrighted content. No new human participants, personal identifiers, or human-subject data were collected for this study. Dataset access and use follow the corresponding licenses and source agreements. Source media, game assets, and pretrained models remain subject to their respective terms, and this statement does not grant unrestricted redistribution rights. The reported quality metrics should not be interpreted as a universal definition of perceptual quality, and deployment in safety-critical or privacy-sensitive settings requires additional safeguards.

## 10 AI USE STATEMENT

AI assistance was used for manuscript language editing, LaTeX organization, and consistency checks across the text, tables, figures, and references. AI assistance did not generate experimental data, numerical results, model checkpoints, or human judgments. All experimental protocols, analyses, and reported claims were reviewed and verified by the authors. The authors are responsible for the manuscript, its implementation, and all reported results.

## REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 126–135, 2017.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

Marcos V Conde, Gregor Geigle, and Radu Timofte. InstructIR: High-quality image restoration following human instructions. In *Proceedings of the European Conference on Computer Vision*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems*, volume 36, pp. 19622–19635, 2023.

Xiaojie Guo, Yu Li, and Haibin Ling. LIME: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, 2017. doi: 10.1109/TIP.2016.2639450.

JiaKui Hu, Zhengjian Yao, Lujia Jin, and Yanye Lu. Bridging degradation discrimination and generation for universal image restoration. *arXiv preprint arXiv:2602.00579*, 2026.

Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu. AutoDIR: Automatic all-in-one image restoration with latent diffusion. In *Proceedings of the European Conference on Computer Vision*, pp. 340–359. Springer, 2024.

Guanzhou Lan, Qianli Ma, Yuqi Yang, Zhigang Wang, Dong Wang, Xuelong Li, and Bin Zhao. Efficient diffusion as low light enhancer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21277–21286, June 2025.

Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28 (1):492–505, 2019.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17452–17462, June 2022.

Hao Li, Xiang Chen, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. FoundIR: Unleashing million-scale training data to advance foundation models for image restoration. *arXiv preprint arXiv:2412.01427*, 2024.

R. Li, R. T. Tan, and L. F. Cheong. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3172–3182, 2020. doi: 10.1109/CVPR42600.2020.00324.

Weiqi Li, Xuanyu Zhang, Bin Chen, Jingfen Xie, Yan Wang, Kexin Zhang, Junlin Li, Li Zhang, Jian Zhang, and Shijie Zhao. UARE: A unified vision-language model for image quality assessment, restoration, and enhancement. *arXiv preprint arXiv:2512.06750*, 2025.

Yunlong Lin, Zixu Lin, Haoyu Chen, Panwang Pan, Chenxin Li, Sixiang Chen, Kairun Wen, Yeying Jin, Wenbo Li, and Xinghao Ding. JarvisIR: Elevating autonomous driving perception with intelligent image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22369–22380, 2025.

Jiyao Liu, Junzhi Ning, Wanying Qu, Lihao Liu, Chenglong Ma, Junjun He, and Ningsheng Xu. MedQ-UNI: Toward unified medical image quality assessment and restoration via vision-language modeling. *arXiv preprint arXiv:2603.18465*, 2026.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for multi-task image restoration. In *International Conference on Learning Representations*, 2024.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3883–3891, 2017.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36:71275–71293, 2023.

Xingyou Song and Dara Bahri. Decoding-based regression. *Transactions on Machine Learning Research*, 2025. URL `https://openreview.net/forum?id=avUQ8jguxg`.

Xiangpeng Tian, Xiangyu Liao, Xiao Liu, Meng Li, and Chao Ren. Degradation-aware feature perturbation for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28165–28175, 2025.

Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. TransWeather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2353–2363, 2022.

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-Instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25490–25500, 2024a.

Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang. A comprehensive study of multimodal large language models for image quality assessment. In *Proceedings of the European Conference on Computer Vision*, pp. 143–160. Springer, 2024b.

Cuixin Yang, Rongkang Dong, and Kin-Man Lam. Vision-language model guided image restoration. *arXiv preprint arXiv:2512.17292*, 2025.

Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021.

Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *Proceedings of the European Conference on Computer Vision*, pp. 259–276. Springer, 2024.

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14483–14494, 2025.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2636–2645, 2020.

Oğuz Kaan Yüksel and Nicolas Flammarion. On the sample complexity of next-token prediction. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 694–702. PMLR, 2025.

Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. Q-Bench: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

# A Reproducibility and Technical Details

## A.1 Dataset Composition and Partitioning

DU-200K contains 160,000 synthetic samples for structured degradation supervision and 40,000 alignment samples for restoration-grounded feedback. These components serve different training objectives and provide different forms of supervision. We call the second component an alignment set because its sources have mixed provenance: RESIDE contains both synthetic and real subsets, and DIV2K includes low-resolution images derived from higher-resolution images. The 40,000 count therefore refers to processed alignment samples rather than to 40,000 independently captured photographs.

The synthetic component uses clean images from ImageNet, CelebA, nuScenes, and BDD100K (Deng et al., 2009; Liu et al., 2015; Caesar et al., 2020; Yu et al., 2020). It contains 140,000 training samples, 4,000 validation samples, and 16,000 test samples. Partitioning is performed by clean-image identity: all degraded variants derived from the same source image belong to the same partition. Only the training partition is used for SFT. The validation partition is used for model and checkpoint selection, and the test partition is reserved for reporting structured prediction performance.

The alignment component contains selected subsets of RESIDE, GoPro, a deduplicated union of LOL-v1 and LOL-v2, and DIV2K Li et al. (2019); Nah et al. (2017); Wei et al. (2018); Yang et al. (2021); Agustsson & Timofte (2017). The source datasets have mixed provenance: RESIDE contains both synthetic and real subsets, while DIV2K includes low-resolution images derived from higher-resolution images. Duplicate low-light inputs occurring across LOL-v1 and LOL-v2 are retained only once.

The counting unit is a degraded input sample rather than an image pair. One paired input and its clean target count as one sample, and one input without a clean target also counts as one sample. For every sample, the provenance manifest records the source dataset and subset, source image or scene/video identifier, clean-target availability, and whether the observation is captured, synthetically generated, or derived from a low-resolution construction. It also records the crop and resampling operation and the duplicate-check result. None of the 40,000 samples has a degradation-parameter annotation. The alignment component is partitioned at the source-image, scene, or video level before cropping or repeated sampling, so all variants from one source unit stay in one partition. The training and validation splits are kept separate; the GRPO job does not create an additional internal split or run online evaluation. The supplement reports the per-source and per-subset counts and the paired-to-unpaired ratio.

Target availability determines the reward branch independently of the source name. Paired samples with a clean target use an equal-weight combination of aesthetic quality and MSE, whereas samples without a clean target use only the aesthetic score. Both branches are defined in Appendix A.5. Thus, the alignment data ground the learned state in restoration consequences without requiring a degradation-parameter label or a clean target for every sample.

## A.2 Input Representation and Structured Output

DU-VLM uses Qwen3-VL-8B as the structured predictor. The degraded RGB image, an auxiliary Sobel edge view, an auxiliary log-amplitude FFT view, and the user instruction are included in the input context. The auxiliary views provide spatial and spectral evidence about the same degraded observation.

The prediction target is the ordered degradation state defined in Eq. (1). Each process node specifies a degradation type together with its parameter keys and values. Composition is encoded by the number and order of process nodes rather than by an independent field. A single node represents one degradation process, whereas multiple ordered nodes represent a compound degradation chain.

The synthetic benchmark contains 16 named degradation classes derived from 12 base operators and uses 20 degradation parameter keys. The operators cover observation projection, environmental propagation, optical-path physical propagation, and imaging-system processing. The model gener-

ates a token sequence encoding the structured state under the quantized formulation described in Section 4.

For low-light degradation, the structured state includes a color-space transformation matrix $\mathbf{M}_{\mathrm{cs}} \in \mathbb{R}^{3\times 3}$. Using column-vector notation, the color-space mapping is

$$\mathbf{c}_{\mathrm{out}} = \mathbf{M}_{\mathrm{cs}}\mathbf{c}_{\mathrm{in}}, \tag{5}$$

where $\mathbf{c}_{\mathrm{in}}$ and $\mathbf{c}_{\mathrm{out}}$ are the three-channel color vectors before and after the mapping. This matrix describes a linear transformation of color channels rather than geometric camera intrinsics. Gain and gamma are represented by separate parameter keys. The matrix constitutes one parameter key with nine scalar entries.

For the low-light restoration pipeline, the operations are applied in the following order: color-space mapping, gamma transformation, RGB gain, inverse gamma transformation, and inverse color-space mapping. Let $\mathcal{M}_{\mathrm{cs}}$ denote the color-space mapping, $\mathcal{G}_\gamma$ the gamma transformation, and $\mathcal{R}_g$ the RGB-gain operation. The complete sequence is

$$I_{\mathrm{out}} = \mathcal{M}_{\mathrm{cs}}^{-1}\big(\mathcal{G}_\gamma^{-1}(\mathcal{R}_g(\mathcal{G}_\gamma(\mathcal{M}_{\mathrm{cs}}(I_{\mathrm{in}}))))\big)\,. \tag{6}$$

The preceding equation specifies the execution order used by the fixed low-light restoration routine.

### A.3 Fixed Restoration Operator Composition

Let the predicted state be

$$\hat{\mathcal{D}} = (\hat{d}_1, \ldots, \hat{d}_{\hat{n}}), \qquad \hat{d}_j = \Big(\hat{t}_j, \{(\hat{k}_{j,\ell}, \hat{v}_{j,\ell})\}_{\ell=1}^{\hat{m}_j}\Big)\,. \tag{7}$$

The compiler first checks every key against the schema of its predicted type. It then instantiates the registered restoration operator $\mathcal{O}_{\hat{t}_j}$ with the corresponding values. The process nodes retain their predicted order. Because the forward chain applies node 1 before node $\hat{n}$, the restoration routines are evaluated in reverse order,

$$\hat{I}_c = \mathcal{O}_{\hat{t}_1} \circ \cdots \circ \mathcal{O}_{\hat{t}_{\hat{n}}}(I_d). \tag{8}$$

The composition acts from right to left, so the routine associated with the last forward process is applied first. The compiler contains no trainable parameters. Table 4 lists the registered routines for individual and compound classes. These fixed heuristics are not assumed to be exact inverses of the corresponding image-formation processes.

The same compiler is used for every GRPO completion. It is independent of the VLM parameters and of all downstream IQA and restoration backbones. The reward evaluates the restored image using the fixed quality metrics defined in Appendix A.5. Some of these metrics use pretrained networks; their parameters are not updated during GRPO.

### A.4 Training Stages

Training consists of synthetic SFT followed by GRPO alignment on the alignment component. SFT uses next-token prediction to learn the structured output from synthetic state annotations. The resulting checkpoint is denoted DU-SFT. The checkpoint obtained after restoration-grounded alignment is denoted DU-GRPO. Comparisons between these checkpoints distinguish the effects of structured supervision and subsequent outcome alignment.

During GRPO, each prompt produces a group of $G = 8$ completions. A completion is compiled into a restored image using the fixed routines in Appendix A.3. Restoration quality provides the sole reward, $R_i = r_{\mathrm{rec},i}$. GRPO updates DU-VLM while the restoration routines and quality evaluators remain fixed. The reward and objective are specified in Appendices A.5 and A.6, respectively.

The DU-GRPO results reported in this version use the final checkpoint obtained after the 24-GPU continuation stage. The preceding GRPO checkpoint 7500 is used only to initialize that continuation stage; it is not the checkpoint used to report the current DU-GRPO results. When DU-SFT is shown as an ablation, it denotes the checkpoint before restoration-grounded GRPO.

Table 4: Fixed restoration routines for the predicted degradation classes. The listed operations are deterministic and contain no learned module; they are not generally exact physical inverses.

| Predicted class | Executed fixed pipeline |
|---|---|
| `night` | color-space mapping, gamma transformation, RGB gain, inverse gamma transformation, and inverse color-space mapping |
| `noise` | Gaussian unsharp masking followed by OpenCV color non-local means (NLM) denoising |
| `night_noise` | gamma brightening followed by unsharp masking and color NLM denoising |
| `blur` | Gaussian unsharp masking |
| `motion_blur` | Gaussian unsharp masking |
| `blur_motion_blur` | Gaussian unsharp masking |
| `low_res` | `INTER_AREA` downsampling, `INTER_CUBIC` upsampling, and detail residual |
| `haze` | analytic atmospheric-scattering inversion $(I - A(1-t))/t$ |
| `haze_night` | gamma brightening followed by analytic atmospheric-scattering inversion |
| `color_cast` | normalized RGB channel-coefficient scaling |
| `lateral_chromatic_aberration` | RGB channel scaling |
| `longitudinal_chromatic_aberration` | RGB channel scaling |
| `jpeg_compression` | Gaussian unsharp masking |
| `jpeg_lca_noise` | channel scaling followed by unsharp masking and color NLM denoising |
| `additive_rain` | gamma correction followed by unsharp masking and color NLM denoising |
| `non_uniform_vignetting` | RGB gain |

### A.5 Restoration Reward

Exactly one reward branch is active for each sample. A sample with a clean target uses the paired reward below. Only a sample without a clean target uses the no-reference fallback.

#### A.5.1 Paired Examples

The degraded input and clean target are bicubically resized to $256 \times 256$. The predicted state is compiled by the fixed restoration routines in Appendix A.3. PSNR, SSIM, and LPIPS compare the restored RGB image with the resized clean target. Aesthetic and CLIP-IQA evaluate the restored image without using the clean target. The paired reward combines these full-reference and no-reference terms:

$$r_{\mathrm{rec}}^{\mathrm{paired}} = \frac{1}{2}\left(s_{\mathrm{Aes}} + s_{\mathrm{MSE}}\right) \tag{9}$$

The five normalized terms have equal weight $0.2$, and their final mean is clipped to $[0, 1]$. The normalized scores are

$$s_{\mathrm{PSNR}} = \mathrm{clip}_{[0,1]}\left((\mathrm{PSNR} - 10)/40\right), \tag{10}$$

$$s_{\mathrm{SSIM}} = \mathrm{clip}_{[0,1]}(\mathrm{SSIM}), \tag{11}$$

$$s_{\mathrm{LPIPS}} = \mathrm{clip}_{[0,1]}(1 - \mathrm{LPIPS}), \tag{12}$$

$$s_{\mathrm{Aes}} = \mathrm{clip}_{[0,1]}(\mathrm{Aesthetic}/10), \tag{13}$$

$$s_{\mathrm{CLIP-IQA}} = \mathrm{clip}_{[0,1]}(\mathrm{CLIP\text{-}IQA}). \tag{14}$$

#### A.5.2 Examples without a Clean Target

Let $\mathcal{M}_{\mathrm{avail}}$ denote the available no-reference metrics. The reward is

$$r_{\mathrm{rec}}^{\mathrm{NR}} = s_{\mathrm{Aes}} \tag{15}$$

Table 5: No-reference reward normalization.

| Metric | Normalized score $s_m$ |
|---|---|
| NIQE | $\mathrm{clip}_{[0,1]}((12-v)/11)$ |
| MUSIQ | $\mathrm{clip}_{[0,1]}(v/100)$ |
| MANIQA, $v \le 1.5$ | $\mathrm{clip}_{[0,1]}(v)$ |
| MANIQA, $v > 1.5$ | $\mathrm{clip}_{[0,1]}(v/10)$ |
| Aesthetic | $\mathrm{clip}_{[0,1]}(v/10)$ |
| CLIP-IQA | $\mathrm{clip}_{[0,1]}(v)$ |

The metric-specific transformations are listed in Table 5. Only successfully evaluated metrics enter $\mathcal{M}_{\mathrm{avail}}$, and each completion must provide at least one valid score. Each available metric receives weight $1/|\mathcal{M}_{\mathrm{avail}}|$. Consequently, omitting a metric changes the weights of the remaining terms and can change the reward scale and ranking.

### A.6 GRPO Objective

For each prompt $q$, we sample $G = 8$ completions. The implementation uses a single reward function, so $R_i = r_{\mathrm{rec},i}$. With reward scaling enabled, the group statistics and advantages are

$$\mu_q = \frac{1}{G}\sum_{i=1}^{G} R_i, \qquad s_q = \mathrm{Std}_{\mathrm{sample}}(R_1, \ldots, R_G), \tag{16}$$

$$A_i = \frac{R_i - \mu_q}{s_q + 10^{-4}}. \tag{17}$$

The standard deviation is the default sample standard deviation used by `torch.std`. If all rewards in a group are equal, every advantage is zero.

For completion token $y_{i,t}$, write $\ell^a_{i,t} = \log \pi_a(y_{i,t} \mid q, y_{i,<t})$ for $a \in \{\theta, \mathrm{old}, \mathrm{ref}\}$. The importance ratio is

$$\rho_{i,t} = \exp(\ell^{\theta}_{i,t} - \ell^{\mathrm{old}}_{i,t}), \tag{18}$$

and $\bar{\rho}_{i,t} = \mathrm{clip}(\rho_{i,t}, 0.8, 1.2)$. The reference-policy penalty uses a clipped k3-style expression:

$$\Delta_{i,t} = \min(\ell^{\mathrm{ref}}_{i,t} - \ell^{\theta}_{i,t},\, 5), \tag{19}$$

$$D^{\mathrm{k3}}_{i,t} = e^{\Delta_{i,t}} - \Delta_{i,t} - 1. \tag{20}$$

The upper clamp on $\Delta_{i,t}$ is part of the implementation. Accordingly, this is a clipped reference-policy penalty rather than the unmodified k3 KL estimator. The objective averages tokens within each completion and then averages the completion losses:

$$\begin{aligned} \mathcal{L}_{\mathrm{GRPO}} &= \frac{1}{B}\sum_{i}\frac{1}{|C_i|}\sum_{t\in C_i}\mathcal{L}_{i,t}, \\ \mathcal{L}_{i,t} &= -\min(\rho_{i,t}A_i, \bar{\rho}_{i,t}A_i) + \beta D^{\mathrm{k3}}_{i,t}. \end{aligned} \tag{21}$$

Here, $C_i$ contains the completion tokens and $\beta$ is the reference-policy KL coefficient. The outer batch contains completion sequences. The 24-GPU configuration in Appendix A.7 uses $B = 384$ sequences from 48 prompts and eight completions per prompt.

### A.7 GRPO Training Configuration

Table 6 specifies the 24-GPU GRPO continuation stage. The policy starts from checkpoint 7500 of the preceding GRPO stage, whereas all DU-GRPO results reported in this version use the final checkpoint obtained after this continuation stage.

The training job does not perform online evaluation or create an additional internal split. Held-out validation is performed separately for checkpoint selection, as described in Appendix A.1.

Table 6: Configuration of the 24-GPU GRPO continuation stage.

| Category | Setting | Category | Setting |
|---|---|---|---|
| Backbone | Qwen3-VL-8B-Instruct | Adaptation | LoRA, rank 64, alpha 256, dropout 0.05, all-linear |
| Policy initialization | preceding GRPO checkpoint 7500 | Reference policy | `ref_adapter` stored with checkpoint 7500 |
| Hardware | 3 nodes × 8 GPUs | Per-GPU batch | 16 completion sequences |
| Gradient accumulation | 1 | Group size | `num_generations`=8 |
| Per optimizer step | 48 prompts, 384 completions | Sampling | temperature 1.0, top-$p$ 0.95, top-$k$ 50 |
| Lengths | context 8192, completion 128 | Padding | left padding and left truncation |
| Learning rate | $3 \times 10^{-8}$ | KL coefficient | $\beta = 0.5$ |
| GRPO clipping | $\epsilon_{\text{clip}} = 0.2$ | Reward scaling | `scale_rewards=true` |
| Gradient clipping | maximum norm 0.3 | Optimizer | fused AdamW |
| Adam coefficients | $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$ | Weight decay | 0.1 |
| Scheduler | cosine, warmup ratio 0.05 | Precision | BF16, FP16 disabled |
| Distributed training | DeepSpeed ZeRO-2 with optimizer offload | Rollout backend | `use_vllm=false` |
| Training length | 20,000 optimizer steps | Checkpointing | every 500 steps, retain at most 10 |
| In-job evaluation | disabled; `split_dataset_ratio=0` | Seed | 42 |
| Reward functions | one, `vlm_enhance_quality` | Additional reward weights | none |

All DU-GRPO results reported in this version use the final checkpoint from this 24-GPU continuation stage. The preceding checkpoint 7500 is the initialization point only.

### A.8 Structured Prediction Evaluation

Structured prediction is evaluated on the 16-class synthetic common test. Generalist MLLMs and DU-trained models receive the same multimodal inputs and are evaluated using the same output schema and parser. Generalist models receive the schema through the prompt, whereas DU-trained models learn to produce it during training. The input views and scored output fields are otherwise identical.

Exact-format validity, type accuracy (T-Acc), macro F1 (M-F1), parameter accuracy (P-Acc), and joint accuracy (J-Acc) measure structural correctness. Penalized absolute and relative errors (P-Abs and P-Rel) measure parameter prediction error. Missing or type-incompatible values receive the zero-prediction penalty rather than being excluded from error measurement.

The SFT scaling experiments vary training-data volume and taxonomy coverage while following the train/validation/test proportions of the full synthetic component. The input-view ablation compares the full input with variants that remove the frequency view, remove the edge view, or retain only RGB. These experiments assess the supervision and input evidence used for structured prediction.

### A.9 Frozen-Feature Extraction and Projection

We extract the final decoder hidden states `hidden_states[-1]` from DU-VLM. Each token is represented by a 4096-dimensional vector. For IQA, the selected token set contains only the body tokens inside the first `<reason>...</reason>` span in the assistant output. For DiffBIR, we select five disjoint token groups. They comprise the image tokens for the degraded RGB, EDGE, and FFT inputs and the body tokens inside the `<reason>...</reason>` and `<answer>...</answer>` spans.

For each selected group $\mathcal{S}$, arithmetic mean pooling gives

$$z_{\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} h_t \in \mathbb{R}^{4096}. \tag{22}$$

IQA uses the pooled reason vector directly as its 4096-dimensional feature. For DiffBIR, mean pooling first produces five 4096-dimensional slots $z_{\mathrm{RGB}}, z_{\mathrm{EDGE}}, z_{\mathrm{FFT}}, z_{\mathrm{reason}}, z_{\mathrm{answer}}$. Their equal-weight average $\bar{z} = \frac{1}{5} \sum_s z_s$ is projected as

$$u = \mathrm{Linear}_{4096 \to 1024} \left( \mathrm{LayerNorm}(\bar{z}) \right), \tag{23}$$

$$C = \mathrm{reshape}_{77 \times 1024} \left( \mathrm{Linear}_{1024 \to 77 \times 1024} \left( \mathrm{SiLU}(u) \right) \right). \tag{24}$$

Thus, $C$ contains 77 conditioning tokens of width 1024. The reason and answer spans specify token-selection boundaries for feature extraction; their use does not by itself establish a reasoning mechanism.

### A.10 IQA Head and Dataset-Specific Transfer

The IQA prediction head applies a non-affine 4096-dimensional LayerNorm, followed by

$$z_{\mathrm{IQA}} = \mathrm{LayerNorm}_{\text{no affine}}(z_{\mathrm{reason}}), \tag{25}$$

$$u_{\mathrm{IQA}} = \mathrm{GELU} \left( \mathrm{Linear}_{4096 \to 256}(z_{\mathrm{IQA}}) \right), \tag{26}$$

$$\hat{q} = \mathrm{Linear}_{256 \to 5} \left( \mathrm{Dropout}_{0.1}(u_{\mathrm{IQA}}) \right). \tag{27}$$

It contains 1,050,117 trainable parameters.

For KADID-10K, KonIQ-10K, LIVE-Wild, PIPAL, and SPAQ, a separate IQA head is trained on the labeled training partition of each target dataset. The five output logits parameterize a five-level quality-score distribution, which is decoded into a continuous quality score using the processing and decoding procedure of DeQA-Score You et al. (2025). No training, validation, or test images are pooled across datasets.

The representation model remains frozen, and only the IQA head is trained. Base Qwen, DU-SFT, and DU-GRPO are compared under the same target-dataset supervision protocol. PLCC and SRCC are reported for each dataset, and the average is the unweighted mean across the five datasets for each metric. The SFT-to-GRPO comparison measures transfer from the corresponding checkpoints separately from synthetic structured prediction accuracy.

This protocol evaluates supervised adaptation to each target dataset. Cross-dataset generalization instead requires training on a specified source dataset and evaluating on target datasets without using their training labels. Results from these protocols must be reported separately; freezing the representation model does not remove target-dataset supervision.

### A.11 Hidden-State Transfer to DiffBIR

For DiffBIR, the two-layer projection in Appendix A.9 is the only trainable conditioner. Its $77 \times 1024$ output is added residually to the 77 CLIP context tokens of DiffBIR, and its final linear layer is initialized to zero. DU-VLM, the DiffBIR backbone, and the CLIP context encoder remain frozen. Features are extracted offline or read from a precomputed cache.

Base Qwen, DepictQA, DU-SFT, and DU-GRPO supply the conditioning representations. The conditioner architecture is held fixed across these comparisons while the frozen feature source is changed. Comparisons under this protocol assess the representation supplied to the restoration model.

### A.12 Explicit-State Prompting and Restoration Evaluation

For explicit-state transfer, Qwen-Image-Edit and HunyuanImage-3 receive restoration instructions with different conditioning information while their weights remain fixed. The comparison includes

a generic prompt, a deliberately mismatched DU state, a DU-SFT prediction, and a DU-GRPO prediction. Evaluation uses 200 paired held-out inputs balanced across four source subsets. These inputs and their source scenes are excluded from the 40,000-sample alignment manifest and from model or checkpoint selection; source identifiers and image hashes are checked before evaluation. The mismatched-state condition tests whether the result depends on image-specific degradation information rather than the presence of structured degradation text alone.

Restoration measurements include PSNR, SSIM, LPIPS, FID, and CLIP-IQA. Higher PSNR, SSIM, and CLIP-IQA values are preferred, whereas lower LPIPS and FID values are preferred. Results are compared within each backbone group. Improvements in fidelity metrics are interpreted together with changes in CLIP-IQA rather than as uniform improvements in every aspect of image quality.

### A.13 Unseen Real-World Evaluation

CleanBench (Lin et al., 2025) evaluates generalization of the complete restoration system. Its test images are excluded from structured supervision, alignment data, and the other training stages of the pipeline. We also exclude all variants derived from these images and remove any source-level or hash-level overlap before sampling 200 official test images. All methods are evaluated on the same inputs at $512 \times 512$ resolution. All baselines are rerun on this subset.

DU-GRPO supplies predicted degradation parameters to frozen HunyuanImage-3. Table 3 reports NIQE, BRISQUE, CLIP-IQA, and MUSIQ on the shared subset. Lower NIQE and BRISQUE values are preferred, whereas higher CLIP-IQA and MUSIQ values are preferred. Figure 7 presents representative outputs from the same test subset.

This evaluation measures generalization of the complete system, including the degradation predictor and restoration backbone. It does not isolate the contribution of GRPO. Stage-specific comparisons instead use the DU-SFT and DU-GRPO checkpoints under the corresponding matched transfer protocols.

### A.14 Derivation of the Approximate NTP Decomposition

This section derives the approximation stated in Section 4.1. The analysis concerns the negative log-likelihood of a canonical state encoding, with a fixed one-to-one mapping between structured states and their serialized sequences. It does not assert an exact identity between an arbitrary token-averaged training loss and a continuous regression loss. Matrix-valued parameters contribute their scalar entries to the value vector in a fixed order.

Let $s = (\boldsymbol{t}, c, \boldsymbol{k})$ denote the structured discrete state, let $v \in \mathbb{R}^d$ denote its continuous values, and let $q_{\boldsymbol{\Delta}}$ be a coordinate-wise quantizer with fixed bin widths $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_d)$. Assume that the conditional value density is locally $\mathcal{N}(\mu_\theta(s, \boldsymbol{x}), \sigma^2 \mathbf{I})$ for fixed $\sigma > 0$ and that each quantization cell is sufficiently small for the density to be approximately constant within it. Writing $\mu_\theta = \mu_\theta(s, \boldsymbol{x})$, the NTP loss satisfies

$$\mathcal{L}_{\mathrm{NTP}} \approx \underbrace{-\log p_\theta(s \mid \boldsymbol{x})}_{\mathcal{L}_{\mathrm{struct}}} + \underbrace{\frac{\|v - \mu_\theta\|_2^2}{2\sigma^2}}_{\mathcal{L}_{\mathrm{value}}} + C_{\sigma,d} - \sum_{j=1}^{d} \log \Delta_j, \tag{28}$$

where $C_{\sigma,d} = \frac{d}{2} \log(2\pi\sigma^2)$ and the final two terms are independent of $\theta$.

**Derivation.** Let $z = q_{\boldsymbol{\Delta}}(v)$ denote the quantized value state. The chain rule separates the joint negative log-likelihood into its discrete structure and quantized-value terms:

$$\mathcal{L}_{\mathrm{NTP}} = -\log p_\theta(s, z \mid \boldsymbol{x}) \tag{29}$$

$$= -\log p_\theta(s \mid \boldsymbol{x}) - \log p_\theta(z \mid s, \boldsymbol{x}). \tag{30}$$

The first term is the cross-entropy loss for the complete discrete structure, including the degradation types, parameter keys, and their composition.

Let $\mathcal{B}_z = \prod_{j=1}^{d} \mathcal{B}_{z_j}$ be the rectangular cell associated with $z$. Its volume is $\mathrm{vol}(\mathcal{B}_z) = \prod_{j=1}^{d} \Delta_j$. The probability assigned to the quantized value is

$$p_\theta(z \mid s, \boldsymbol{x}) = \int_{\mathcal{B}_z} p_\theta(u \mid s, \boldsymbol{x})\, \mathrm{d}u. \tag{31}$$

The integral mean-value theorem gives a point $\xi_z \in \mathcal{B}_z$ for which this mass equals $p_\theta(\xi_z \mid s, \boldsymbol{x}) \prod_j \Delta_j$. Local constancy of the density within a sufficiently small cell gives

$$p_\theta(z \mid s, \boldsymbol{x}) \approx p_\theta(v \mid s, \boldsymbol{x}) \prod_{j=1}^{d} \Delta_j. \tag{32}$$

Consequently,

$$-\log p_\theta(z \mid s, \boldsymbol{x}) \approx -\log p_\theta(v \mid s, \boldsymbol{x}) - \sum_{j=1}^{d} \log \Delta_j. \tag{33}$$

For a positive density continuous at $v$, the error in this cell-mass approximation tends to zero as $\max_j \Delta_j \to 0$. This limit does not remove the separate Gaussian modeling assumption.

Under the local Gaussian assumption,

$$p_\theta(v \mid s, \boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|v - \mu_\theta\|_2^2}{2\sigma^2}\right). \tag{34}$$

Taking the negative logarithm yields

$$-\log p_\theta(v \mid s, \boldsymbol{x}) = \frac{\|v - \mu_\theta\|_2^2}{2\sigma^2} + \frac{d}{2} \log(2\pi\sigma^2). \tag{35}$$

Substituting Eqs. equation 33 and equation 35 into Eq. equation 30 gives the stated decomposition. Since $\sigma$ and $\boldsymbol{\Delta}$ are fixed, $C_{\sigma,d}$ and $-\sum_j \log \Delta_j$ do not affect optimization over $\theta$. Thus, under the stated approximation, NTP jointly minimizes structure cross-entropy and a variance-weighted squared value error.

### A.15 Conditional Bounds on State and Parameter-Value Errors

Let $s$ collect the discrete type, key, and composition variables, and let $v \in \mathbb{R}^d$ be their continuous representation in a fixed canonical parameter space. A coordinate-wise quantizer with bin widths $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_d)$ defines the discrete state $z_{\boldsymbol{\Delta}} = (s, q_{\boldsymbol{\Delta}}(v))$.

Assume that $z_{\boldsymbol{\Delta}}$ is deterministic for each input $\boldsymbol{x}$, that

$$\mathbb{E}_{\boldsymbol{x}} D_{\mathrm{KL}}\left(\delta_{z_{\boldsymbol{\Delta}}} \| P_\theta(\cdot \mid \boldsymbol{x})\right) \le \epsilon,$$

that prediction uses MAP decoding over the quantized state, and that the true values and decoded cell centers lie in a common canonical value domain with Euclidean diameter $D_{\max}$.

Let $\rho_{\boldsymbol{\Delta}} = \frac{1}{2}\|\boldsymbol{\Delta}\|_2$ be the maximum center-decoding radius of a rectangular quantization cell. Then the exact-state classification error and parameter-value MSE satisfy

$$R_{\mathrm{cls}} = \Pr[\hat{z}_{\boldsymbol{\Delta}} \neq z_{\boldsymbol{\Delta}}] \le \sqrt{2\epsilon}, \tag{36}$$

$$R_{\mathrm{reg}} = \mathbb{E}\|\hat{v} - v\|_2^2 \le \rho_{\boldsymbol{\Delta}}^2 + D_{\max}^2 \sqrt{2\epsilon}. \tag{37}$$

**Derivation.** The Pinsker inequality gives

$$\mathrm{TV}\left(\delta_{z_{\boldsymbol{\Delta}}}, P_\theta\right) \le \sqrt{\frac{D_{\mathrm{KL}}\left(\delta_{z_{\boldsymbol{\Delta}}} \| P_\theta\right)}{2}}.$$

For a deterministic target, MAP error is at most twice this total-variation distance. Taking the expectation and applying the Jensen inequality gives $R_{\mathrm{cls}} \le \sqrt{2\epsilon}$. Correct state decoding incurs at

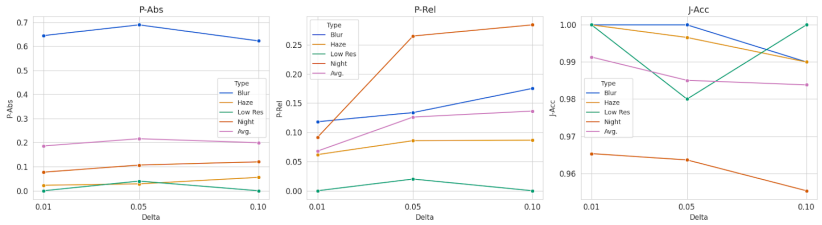



Figure 8: **Quantization-grid ablation.** Results for the four displayed degradation categories across bin widths $\Delta \in \{0.01, 0.05, 0.1\}$ in terms of penalized absolute error (P-Abs), penalized relative error (P-Rel), and joint accuracy (J-Acc).

most $\rho_{\mathbf{\Delta}}^2$ center-quantization error, whereas an incorrect state contributes at most $D_{\max}^2$. Combining the two events gives the stated parameter-value error bound. Type-only and key-only classification errors are subsets of the exact-state error event and therefore obey the same conservative upper bound.

These conditional bounds concern the structured state and its parameter values. They do not bound restored-image error or perceptual quality. The MAP assumption refers to global decoding over the complete quantized state, not necessarily greedy token decoding or sampled completions. The deterministic-target assumption also excludes ambiguity in identifying a physical degradation history from an observed image. The bounds therefore do not establish a guarantee for GRPO restoration or downstream transfer.

### A.16 Quantization-Grid Ablation

Figure 8 reports the quantization-grid ablation for the four displayed categories: blur, haze, low resolution, and night degradation. The tested bin widths are $\Delta \in \{0.01, 0.05, 0.1\}$. Among these settings, $\Delta = 0.01$ has the lowest plotted average P-Abs and P-Rel and the highest plotted average J-Acc. Individual class curves are not uniformly monotonic. This four-category diagnostic is distinct from the 16-class common test and does not by itself demonstrate downstream representation transfer or validate the assumptions of Appendix A.15.