FROM GLANCE TO SCRUTINY: PROGRESSIVE DISTORTION REASONING FOR FINE-GRAINED IMAGE QUALITY ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modal large language models (MLLMs) have demonstrated significant potential in image quality assessment (IQA) by bridging visual perception with descriptive evaluations. However, existing approaches mainly focus on holistic quality prediction, often functioning as black boxes that provide limited insight into where distortions occur and how they affect perceived quality, hindering fine-grained analysis of localized and heterogeneous degradations. We propose GS-IQA, a framework that reformulates IQA as a progressive *Where–What–How* diagnosis, emulating the human perceptual process from an initial glance to closer scrutiny. Since a severity judgment is meaningful only for a correctly localized and recognized region, we realize this progression through a two-stage reinforcement learning paradigm that respects such dependencies: the glance stage uses a perception-gated reward to establish where degradations lie and what they are, activating severity feedback only once both are correct, while the scrutiny stage introduces online reward-conditioned degradation generation to synthesize hard examples targeted at the model's perceptual bottlenecks, sharpening its discrimination of subtle severity variations. To enable systematic evaluation, we construct Diag-Bench, a region-level IQA benchmark of about 25K curated samples spanning 12 distortion types and five ordinal severity levels. Extensive experiments show that GS-IQA consistently surpasses state-of-the-art methods in distortion localization, recognition, and severity estimation, and that its diagnostic representations transfer effectively to conventional global quality prediction across diverse external benchmarks. Code and data will be released.

1 INTRODUCTION

Image Quality Assessment (IQA) aims to evaluate visual quality consistently with human perception. Conventional methods estimate quality using handcrafted metrics or learned perceptual representations (Wang et al., 2004; Zhang et al., 2018b; Su et al., 2020; Ke et al., 2021), typically summarizing an image with a scalar score. Despite strong numerical performance, their reliance on holistic score regression often renders the assessment process opaque, offering limited evidence about the distortions underlying a predicted quality score.

Multimodal large language models (MLLMs) has extended IQA beyond scalar prediction toward language-based perception. Score-based methods estimate overall quality (Wu et al., 2023; You et al., 2025a), whereas description-based methods explain perceived artifacts in natural language (Wu et al., 2024a). Through instruction tuning and large-scale quality-description corpora, these methods provide interpretable quality judgments, yet many assessments remain image-level summaries that do not explicitly associate distortion location, category, and severity. This limitation is particularly evident for spatially heterogeneous degradations, where a global assessment may obscure differences in distortion type and severity across regions. As illustrated in Figure 1, Q-Insight (Li et al., 2025) identifies only a dominant degradation whose severity is diluted by pristine regions, whereas VisualQuality-R1 (Wu et al., 2025) assigns a relatively high holistic score despite prominent local defects. Such global assessments may obscure perceptually important artifacts and provide limited guidance for distortion-aware restoration (Zhang et al., 2018a) and region-adaptive enhancement (Moran et al., 2020; Guo et al., 2020).

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



Figure 1: Qualitative comparison of GS-IQA with existing methods. While Q-Insight and VisualQuality-R1 condense spatially heterogeneous degradations into a single distortion description or holistic quality score, GS-IQA employs progressive distortion reasoning to localize individual degraded regions and associate each with its distortion type and ordinal severity, enabling fine-grained and spatially grounded quality assessment.

Human observers, in contrast, rarely assess such complex degradations through an immediate monolithic judgment. Visual quality perception typically unfolds progressively: an initial *glance* establishes a broad awareness of where potential degradations occur and what distortion patterns they resemble, after which closer *scrutiny* examines the relevant local evidence to determine how severely visual quality is affected. This perceptual progression naturally gives rise to three interdependent questions: *where* does the degradation occur, *what* type of distortion is present, and *how* severe is it? Crucially, these questions are neither independent nor equally difficult. Reliable severity estimation presupposes that the model has first attended to the correct region and identified the corresponding distortion pattern; otherwise, the severity signal may be dominated by irrelevant or pristine content.

Motivated by this observation, we propose GS-IQA, a unified framework that performs fine-grained image quality assessment through progressive *Where–What–How* distortion reasoning. Rather than condensing heterogeneous degradations into a single global judgment, GS-IQA associates each degraded region with its distortion category and ordinal severity, thereby providing spatially grounded and verifiable quality interpretations. To learn these fine-grained associations, we develop a two-stage reinforcement learning paradigm based on Group Relative Policy Optimization (GRPO) (Shao et al., 2024), through which the model progresses from broad distortion perception to subtle severity discrimination. During the *glance* stage, a perception-gated reward coordinates distortion localization, recognition, and severity learning according to their prerequisite relationships. In particular, the severity reward is activated only when the predicted region and distortion category are sufficiently reliable, preventing erroneous grounding or recognition from introducing misleading supervision while retaining coarse intensity cues for valid predictions. Having established robust *Where–What* perception, the scrutiny stage shifts the optimization focus toward subtle severity variations that remain difficult to distinguish. Since repeatedly training on a fixed corpus can lead to early saturation, we introduce online reward-conditioned degradation generation, where category-wise severity statistics adapt the sampling probabilities of different distortions. By allocating more synthesized samples to underperforming categories, the training distribution evolves with the model, sustaining informative optimization and progressively sharpening fine-grained severity discrimination. To support the training and systematic evaluation of fine-grained quality assessment, we further construct Diag-Bench, a region-level benchmark comprising 25K carefully curated samples across 12 distortion types and five ordinal severity levels. In summary, our contributions are as follows:

- We introduce GS-IQA, a unified framework that reformulates fine-grained IQA as progressive *Where–What–How* reasoning, explicitly associating each degraded region with its distortion type and ordinal severity.

- A *Glance-to-Scrutiny* reinforcement learning paradigm is developed that respects the perceptual dependencies among distortion localization, recognition, and severity estimation. A perception-gated reward first establishes reliable *Where–What* perception progressively refining fine-grained *How* discrimination.
- We propose online reward-conditioned degradation generation for the scrutiny stage, which tracks category-wise severity performance and adaptively reallocates synthesis toward underperforming distortions, mitigating the saturation of severity learning on static data.
- Diag-Bench provides a region-level quality diagnosis benchmark comprising 25K samples across 12 distortion categories and five ordinal severity levels. Extensive experiments demonstrate the effectiveness of GS-IQA in fine-grained distortion assessment and its transferability to conventional global quality prediction.

2 Related Works

Image Quality Assessment. Full-reference (FR) IQA compares distorted images with pristine references using structural similarity or learned perceptual representations (Wang et al., 2004; Zhang et al., 2018b; Prashnani et al., 2018; Cao et al., 2022), whereas no-reference (NR) IQA predicts quality directly from distorted inputs, evolving from natural scene statistics (Mittal et al., 2012) to deep neural predictors (Talebi & Milanfar, 2018; Su et al., 2020) and multi-scale transformers (Ke et al., 2021). While these methods primarily produce holistic scores, MLLMs extend IQA toward language-based quality understanding through scoring, descriptive assessment, and perceptual reasoning (You et al., 2024; Wu et al., 2023; You et al., 2025a; Wu et al., 2024a;b; Zhang et al., 2025b;a). Recent studies further adopt reinforcement learning for perceptual alignment and distortion understanding (Li et al., 2025; Wu et al., 2025), while spatially grounded methods associate quality judgments with localized visual evidence (Chen et al., 2024a; 2026; Peng et al., 2026). SEAGULL (Chen et al., 2024b) predicts distortion types and ordinal severity for specified regions, and Refine-IQA (Jia et al., 2026) incorporates distortion recognition, severity estimation, and grounding into reinforcement fine-tuning. These advances improve local quality understanding, yet the dependencies among localization, recognition, and severity remain underexploited in joint diagnosis. GS-IQA addresses this gap through progressive *Where–What–How* reasoning, in which reliable localization and recognition constrain ordinal severity supervision, avoiding misleading feedback from incorrect region–type associations. Its reward-conditioned online generation further adapts distortion-category sampling to the model's evolving severity weaknesses, advancing training from coarse degradation perception toward fine-grained severity discrimination.

Reinforcement Learning for MLLMs. Reinforcement learning with verifiable rewards has proven effective for eliciting reasoning without dense annotation (Shao et al., 2024; Guo et al., 2025), and has been extended to multimodal perception such as detection and grounding (Liu et al., 2025b;a; Bai et al., 2025b), where rule-based signals like IoU provide objective supervision. In IQA, recent studies have introduced reward optimization for quality scoring and distortion understanding (Li et al., 2025; Wu et al., 2025; Jia et al., 2026), yet joint region-level diagnosis poses two additional challenges: independently optimized objectives can assign misleading severity credit to incorrectly localized or recognized regions, while fixed training corpora gradually lose informativeness as familiar degradations dominate. Existing curriculum or reweighting strategies (Bai et al., 2025b; Yan et al., 2026), which only reorganize available samples, cannot expand the training distribution. GS-IQA addresses both by gating severity feedback on verified localization and recognition, while using reward-conditioned online synthesis to replenish training with degradations from categories that remain difficult to grade.

3 Methodology

As illustrated in Figure 2, GS-IQA performs fine-grained *Where–What–How* diagnosis by associating each degraded region with its distortion type and ordinal severity. Since severity estimation depends on reliable localization and recognition, we organize training into a two-stage GRPO framework that progresses from region-level perception to fine-grained severity discrimination. In the Glance stage, perception-gated multi-task rewards jointly supervise localization and recognition, releasing ordinal severity credit only for spatially matched, category-correct predictions. Building on

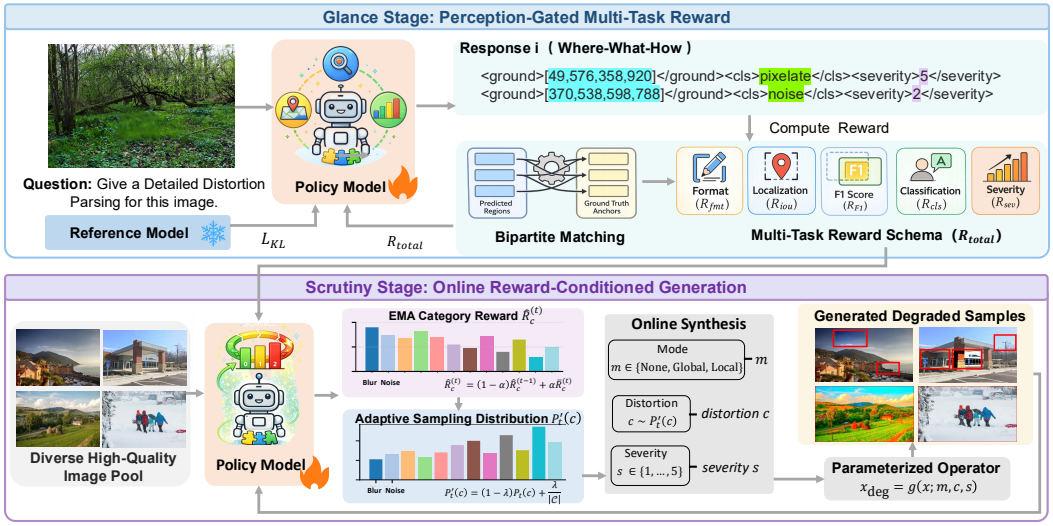


Figure 2: Framework of GS-IQA, which performs quality diagnosis by predicting triplets that localize degraded areas, identify distortion types, and estimate ordinal severity. Training follows a GRPO-based *from-glance-to-scrutiny* scheme: the glance stage learns region-level grounding and recognition under perception-gated multi-task reward, while the scrutiny stage refines severity estimation with online reward-conditioned degradation generation of controllable hard examples.

this foundation, the Scrutiny stage uses category-wise severity feedback to guide online degradation synthesis, allocating additional samples to distortion categories that remain difficult to grade.

3.1 Glance Stage: Perception-Gated Multi-Task Learning

Building on the MLLM's low-level visual priors, the Glance stage learns region-level *Where–What–How* perception under GRPO. A perception-gated multi-task reward jointly supervises distortion localization and category recognition, while activating ordinal severity feedback only for spatially matched regions with correctly recognized distortion types. This design provides reliable supervision for localized degradation perception and establishes a robust perceptual foundation for the subsequent Scrutiny stage.

Format Reward. To enable reliable parsing and automatic reward computation, we adopt a lightweight serialization for the predicted *Where–What–How* associations. The reasoning process is enclosed by `<think>...</think>` and the final prediction by `<answer>...</answer>`. Within the answer, each region–category–severity association $\hat{\mathbf{t}}_i = (\hat{\mathbf{b}}_i, \hat{c}_i, \hat{s}_i)$ is serialized as:

$$\hat{\mathbf{t}}_i \equiv \texttt{<ground>}\hat{\mathbf{b}}_i\texttt{</ground>},\texttt{<cls>}\hat{c}_i\texttt{</cls>},\texttt{<severity>}\hat{s}_i\texttt{</severity>}, \quad (1)$$

where $\hat{\mathbf{b}}_i = [\hat{x}_{i1}, \hat{y}_{i1}, \hat{x}_{i2}, \hat{y}_{i2}]$, $\hat{c}_i$, and $\hat{s}_i$ denote the predicted bounding box, distortion category, and severity level. The extracted associations form $\hat{\mathcal{T}} = \{\hat{\mathbf{t}}_i\}_{i=1}^{N}$, shared by the grounding and severity rewards below. We set $R_{\text{fmt}} = 1$ when both spans are correctly delimited and the answer contains at least one well-formed association, and $R_{\text{fmt}} = 0$ otherwise; semantic correctness is evaluated by subsequent perception rewards. For pristine images, we use the canonical association $([0, 0, 1000, 1000], \texttt{None}, 0)$ to share the same output and reward pipeline as degraded images.

Degradation Grounding Reward. This reward supervises degradation grounding, covering spatial localization (*Where*) and distortion identification (*What*). Since a single image may contain multiple degraded regions and co-occurring artifacts, the predictions form an unordered set rather than an ordered sequence. To handle this permutation ambiguity, we establish an optimal bipartite matching between predicted and ground-truth triplets before reward computation, ensuring each prediction is compared to the best-aligned ground-truth region.

Specifically, we formulate this correspondence as a bipartite assignment problem. Let $\hat{\mathcal{T}} = \{\hat{\mathbf{t}}_i\}_{i=1}^{N}$ and $\mathcal{T} = \{\mathbf{t}_j\}_{j=1}^{M}$ denote the predicted and ground-truth triplet sets, respectively. Each triplet $\mathbf{t} =$

$(\mathbf{b}, c, s)$ comprises a bounding box $\mathbf{b}$, a distortion category c, and an ordinal severity level s. To establish correspondence between the two sets, we define $\Phi(\hat{\mathbf{t}}_i, \mathbf{t}_j)$ as their pairwise perception score and seek the assignment that maximizes the overall compatibility:

$$\hat{\sigma} = \arg \max_{\sigma \in \Omega_{N,M}} \sum_{i=1}^{N} \Phi(\hat{\mathbf{t}}_i, \mathbf{t}_{\sigma(i)}), \tag{2}$$

where $\Omega_{N,M}$ denotes the set of valid one-to-one matchings from predictions to ground-truth, obtained using the Hungarian algorithm (Kuhn, 1955). Based on optimal assignment, we retain the spatially valid pairs whose intersection-over-union exceeds: $\mathcal{M} = \{(i, j) \mid j = \hat{\sigma}(i),\ \mathrm{IoU}(\hat{\mathbf{b}}_i, \mathbf{b}_j) \geq 0.5\}$. The localization and classification rewards are then computed over the matched pairs:

$$R_{\text{iou}} = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} \mathrm{IoU}(\hat{\mathbf{b}}_i, \mathbf{b}_j), \quad R_{\text{cls}} = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} \mathbb{I}(\hat{c}_i = c_j). \tag{3}$$

Both rewards are set to zero when no valid pair is found. To penalize both missed regions and hallucinated predictions, we incorporate the F1-style grounding reward $\frac{2|\mathcal{M}|}{N+M}$, which is equivalent to the harmonic mean of matching precision $|\mathcal{M}|/N$ and recall $|\mathcal{M}|/M$.

Perception-Gated Severity Reward. Severity estimation is inherently category-dependent, as the same ordinal level may exhibit markedly different perceptual salience across distortion types (e.g., Level 1 saturation can be more noticeable than Level 1 blur). A severity prediction is therefore meaningful only relative to a correctly recognized distortion category. For the spatially matched pairs in $\mathcal{M}$, we activate severity supervision only when the predicted category agrees with the ground truth, conditioning *How* discrimination on reliable *Where–What* perception.

To provide denser feedback while preserving the ordinal structure of severity, we assign graded credit according to the distance between the predicted and ground-truth levels. For each matched pair $(i, j) \in \mathcal{M}$, we define the ordinal severity distance as

$$d_{ij}^{\text{sev}} = |\hat{s}_i - s_j|. \tag{4}$$

The tiered ordinal reward assigns progressively lower credit as the severity distance increases, taking $r_{\text{ord}} = 1, \alpha_1, \alpha_2, 0$ for $d = 0, 1, 2$, and $d \geq 3$, respectively, where $\alpha_1 > \alpha_2 > 0$ enforces monotonic decay and $(\alpha_1, \alpha_2) = (0.6, 0.2)$ in practice. Crucially, severity credit is released only when both spatial correspondence and distortion recognition are correct. We therefore define

$$R_{\text{sev}} = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} \mathbb{I}(\hat{c}_i = c_j)\, r_{\text{ord}}\big(d_{ij}^{\text{sev}}\big), \tag{5}$$

Unmatched or misclassified regions contribute zero reward. This normalization prevents selective omission of difficult regions from inflating the severity reward, while preserving the prerequisite relationship that valid *How* supervision requires reliable *Where–What* perception.

Overall Multi-Task Reward. The above components are combined to jointly optimize format validity, degradation grounding, and severity perception:

$$R_{\text{total}} = R_{\text{fmt}} + R_{\text{iou}} + R_{\text{F1}} + R_{\text{cls}} + R_{\text{sev}}. \tag{6}$$

3.2 Scrutiny Stage: Online Reward-Conditioned Generation

After the *Glance* stage establishes reliable region-level degradation perception, further improvement in severity discrimination is often limited by the fixed composition of static training data. A fixed corpus provides finite combinations of distortion patterns and severity appearances, causing optimization to gradually concentrate on already well-learned degradations. To sustain informative training, the Scrutiny stage introduces online reward-conditioned degradation generation, which tracks category-wise severity performance and adaptively reallocates sampling toward distortions that remain difficult to grade. As the policy evolves, the training distribution evolves accordingly, continually refreshing fine-grained supervision around the model's current weaknesses.

Reward-Driven Distribution Evolution. We maintain an online profile of category-wise severity performance to guide degradation sampling. At step t, for each distortion category c, we collect its

spatially matched regions as $\mathcal{B}_c^{(t)} = \{(i,j) \in \mathcal{M} \mid c_j = c\}$. The batch-level score measures severity performance over these regions, where category-mismatched predictions receive zero credit. To reduce batch-wise fluctuations, we further maintain an exponential moving average (EMA):

$$\bar{R}_c^{(t)} = \frac{1}{|\mathcal{B}_c^{(t)}|} \sum_{(i,j)\in\mathcal{B}_c^{(t)}} \mathbb{I}(\hat{c}_i = c_j)\, r_{\mathrm{ord}}\big(d_{ij}^{\mathrm{sev}}\big), \quad \hat{R}_c^{(t)} = (1-\alpha)\hat{R}_c^{(t-1)} + \alpha \bar{R}_c^{(t)}, \tag{7}$$

where $\bar{R}_c^{(t)}$ reflects the current batch-wise severity performance of category c, and $\hat{R}_c^{(t)}$ provides a smoothed estimate of its recent learning status. If category c is absent from the current batch, its EMA statistic remains unchanged. A lower $\hat{R}_c^{(t)}$ therefore indicates a distortion whose severity remains harder to grade. We convert these statistics into an adaptive sampling distribution by applying a temperature-scaled softmax over the negative rewards:

$$P_t(c) = \frac{\exp\Big(-\hat{R}_c^{(t)}/\tau\Big)}{\sum_{c'\in\mathcal{C}} \exp\Big(-\hat{R}_{c'}^{(t)}/\tau\Big)}, \quad P_t'(c) = (1-\lambda)P_t(c) + \frac{\lambda}{|\mathcal{C}|}, \tag{8}$$

where τ controls how strongly sampling concentrates on low-reward categories, while λ preserves exploration through a uniform prior. Consequently, categories with persistently weak severity performance receive more training samples, whereas well-learned categories gradually recede. Unlike reweighting a fixed corpus, this distribution directly governs online degradation synthesis, allowing the training data to evolve with the policy.

Online Degradation Synthesis. Given the adaptive sampling distribution $P_t'(c)$, we synthesize training samples online by sampling a degradation mode $m \in \{\texttt{None}, \texttt{Global}, \texttt{Local}\}$, a distortion category $c \sim P_t'(c)$, and an ordinal severity level $s \in \{1, \ldots, 5\}$. The mode determines whether we keep the image pristine, apply a single global degradation, or inject multiple localized degradations into randomly sampled patches. The generator $g(\cdot)$ applies a degradation operator to a pristine image x to produce a degraded image:

$$x_{\mathrm{deg}} = g(x; m, c, s). \tag{9}$$

Crucially, it emits structured supervision together with the synthesized image. For each degraded region k, we record its bounding box b_k, category c_k, and severity level s_k, forming a set-structured annotation $\mathcal{T} = \{(b_k, c_k, s_k)\}_{k=1}^{K}$. In the local mode, the set of severity $\{s_k\}$ is sampled independently for each injected region, enabling multiple localized degradations with controllable severity within one image. By combining reward-driven category sampling with controllable severity synthesis, this stage continually expands fine-grained supervision while concentrating training capacity on distortions with weak severity estimation.

3.3 Diag-Bench: Benchmark for Fine-grained Quality Diagnosis

We construct Diag-Bench, a 25K-sample benchmark that formulates IQA as a fine-grained *Where–What–How* diagnosis, requiring models to localize degraded regions, identify distortion types, and estimate ordinal severity. Unlike existing benchmarks dominated by image-level scores or coarse descriptions, Diag-Bench combines authentic and synthetic degradations with region-level annotations. Following prior IQA studies (You et al., 2025b), we consider 12 representative distortion categories—noise, blur, compression, overexposure, underexposure, high contrast, low contrast, oversaturate, desaturate, oversharpen, pixelate, quantization—each with five ordinal severity levels.

The authentic subset is built from ViDA-UGC (Liao et al., 2025), a large-scale UGC dataset with region-level distortion annotations, which we map to our predefined categories while discarding unmatched ones. For unambiguous severity annotation, we retain only images with a single distortion category and annotate their ordinal severity, yielding 3,470 training and 868 testing samples. The synthetic subset provides controllable supervision over distortion type, spatial extent, and severity. We sample 8,000 pristine images from KADIS-700K (Lin et al., 2019) and screen their quality using Q-Insight (Li et al., 2025) and VisualQuality-R1 (Wu et al., 2025), followed by manual inspection, retaining 4,275 high-quality sources. Before degradation, these are split into 3,775 training and 500 testing sources to prevent source-level leakage. All static and online samples for Glance and Scrutiny are generated exclusively from the training pool, leaving test sources unseen during optimization. Parameterized OpenCV operators (You et al., 2025b; Bradski, 2000; Lin et al., 2019)

Table 1: Fine-grained image quality diagnosis results on Diag-Bench. We evaluate distortion localization (Where), type recognition (What), and severity estimation (How). † denotes task-aligned variants. The best and second-best results are marked in **bold** and underlined, respectively.

Method	Where				What	How
	Precision	Recall	F1	mIoU	ClsAcc	SevAcc
General-purpose MLLMs						
Gemini-2.5-Pro (Comanici et al., 2025)	0.343	0.428	0.381	0.920	0.466	0.363
Gemini-3-Pro	0.722	0.563	0.633	0.921	0.603	0.563
GPT-4o (Hurst et al., 2024)	0.183	0.227	0.202	0.876	0.600	0.600
GPT-5 (Singh et al., 2025)	0.138	0.269	0.183	0.854	0.677	0.400
Qwen3-VL-4B (Bai et al., 2025a) (Base)	0.544	0.176	0.266	0.905	0.931	0.316
Qwen3-VL-235B (Bai et al., 2025a)	0.624	0.330	0.431	0.954	0.535	0.677
Grounding-oriented Models						
Shikra-7B (Chen et al., 2023)	0.019	0.184	0.034	0.602	0.041	N/A
Ferret-7B (You et al., 2023)	0.032	0.286	0.058	0.687	0.055	N/A
Kosmos-2-1.6B (Peng et al., 2023)	0.012	0.125	0.022	0.583	0.033	N/A
GroundingGPT-7B (Li et al., 2024)	0.024	0.203	0.043	0.653	0.067	N/A
IQA-specific Models						
Q-Insight (Li et al., 2025)	N/A	N/A	N/A	N/A	0.346	0.502
VisualQuality-R1 (Wu et al., 2025)	N/A	N/A	N/A	N/A	0.284	0.606
Fine-grained IQA Models						
Supervised Fine-tuning	0.854	0.861	0.857	0.926	0.467	0.585
Q-Insight† (Li et al., 2025)	0.889	0.814	0.850	0.938	0.892	0.652
VisualQuality-R1† (Wu et al., 2025)	0.910	0.858	0.883	0.946	0.907	0.438
Refine-IQA† (Jia et al., 2026)	0.948	0.832	0.886	0.947	0.869	0.669
GS-IQA-Glance (Ours)	0.948	0.848	0.895	0.953	0.921	0.678
GS-IQA-Scrutiny (Ours)	**0.953**	**0.909**	**0.930**	**0.967**	**0.947**	**0.732**

generate pristine, global, and non-overlapping local degradations in a 1:2:2 ratio, with 1–3 regions per sample. Each image is annotated with $\{(b_k, c_k, s_k)\}_{k=1}^{K}$, where $s_k \in \{1, \ldots, 5\}$ denotes ordinal severity; pristine images use $([0, 0, 1000, 1000], \texttt{None}, 0)$. Human experts discard perceptually ambiguous candidates, resulting in 21,275 samples, with 18,775 for training and 2,500 for testing.

4 Experiments

4.1 Experimental Setup

Implementation Details. We adopt Qwen3-VL-4B (Bai et al., 2025a) as the base model and optimize it through two consecutive reinforcement stages without preceding supervised fine-tuning. The Glance stage is trained on the static training split of Diag-Bench, whereas the Scrutiny stage employs online reward-conditioned degradation generation for severity refinement. GRPO samples a group of $G = 5$ responses per prompt with a global batch size of 128. In the scrutiny stage, the online sampler tracks per-category severity performance with EMA momentum $\alpha = 0.1$, using temperature $\tau = 1.0$ for category reweighting, and mixes the sampling distribution with a uniform prior using $\lambda = 0.05$. All experiments are conducted with NVIDIA H800 GPUs.

Evaluation Protocols and Metrics. We evaluate GS-IQA under three complementary settings covering core diagnosis, cross-dataset generalization, and transfer to holistic IQA. On Diag-Bench, region-level *Where–What–How* diagnosis is evaluated using precision, recall, F1-score, and mIoU for localization, and classification accuracy (ClsAcc) and severity accuracy (SevAcc) for distortion recognition and severity estimation. Cross-dataset generalization is evaluated on external benchmarks without dataset-specific fine-tuning, reporting ClsAcc and SevAcc under global degradation. Finally, transfer to global IQA is measured by Spearman Rank-Order Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC), assessing whether region-aware distortion semantics and ordinal severity cues benefit holistic quality prediction.

4.2 Comparison with State-of-the-Art Methods

Fine-Grained Quality Diagnosis. Table 1 compares GS-IQA with general-purpose MLLMs, grounding models, IQA-specific models, and task-aligned fine-grained baselines on Diag-Bench. General-purpose and grounding-oriented models show limited transfer to low-level degradation localization: GPT-4o and GPT-5 achieve F1 scores of only 0.202 and 0.183, while conventional

Table 3: Cross-dataset generalization of distortion diagnosis on external IQA benchmarks without dataset-specific fine-tuning. Best results are highlighted in **bold**.

Method	Metric	DQ-495K	KADID-10k	TID2013	Waterloo
Q-Insight (Li et al., 2025)	ClsAcc	0.284	0.242	0.443	0.504
	SevAcc	0.398	0.430	0.085	0.254
VisualQuality-R1 (Wu et al., 2025)	ClsAcc	0.157	0.188	0.158	0.332
	SevAcc	0.498	0.232	0.306	0.074
GS-IQA (Ours)	ClsAcc	**0.787**	**0.658**	**0.736**	**0.832**
	SevAcc	**0.532**	**0.514**	**0.354**	**0.399**

Table 4: Holistic quality prediction transferability across five external IQA benchmarks.

Method	BID		CLIVE		SPAQ		KonIQ		AGIQA		AVG.	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Q-Align (Wu et al., 2023)	0.576	0.651	0.554	0.643	0.767	0.779	0.573	0.612	0.682	0.705	0.630	0.678
DeQA-Score (You et al., 2025a)	0.702	0.743	0.743	0.795	0.852	0.858	0.677	0.703	0.738	0.790	0.742	0.778
Q-Insight (Li et al., 2025)	0.806	0.818	0.804	0.837	0.907	0.912	0.812	0.809	0.657	0.705	0.797	0.816
VisualQuality-R1 (Wu et al., 2025)	0.811	0.820	0.811	0.844	**0.913**	**0.917**	0.855	0.870	**0.754**	**0.820**	0.829	0.854
GS-IQA (Ours)	**0.886**	**0.884**	**0.893**	**0.898**	0.886	0.882	**0.885**	**0.901**	0.749	0.780	**0.860**	**0.869**

grounding models remain below 0.06. This reflects the difficulty of localizing degradations, which are often weakly bounded and spatially diffuse rather than semantically well-defined objects.

IQA-specific MLLMs provide stronger quality awareness, but their holistic formulations neither ground individual degraded regions nor reliably disentangle region-wise distortion attributes. Supervised fine-tuning achieves competitive localization but substantially weaker recognition and severity estimation. For fair comparison, we adapt Q-Insight and VisualQuality-R1 to the same prompts, training data, and structured prediction format while preserving their optimization objectives. Both task-aligned variants markedly improve localization and recognition over SFT, but neither explicitly exploits ordinal severity distance, reaching SevAcc scores of 0.652 and 0.438, respectively. GS-IQA-Glance further raises SevAcc to 0.678 through perception-gated ordinal rewards, while GS-IQA-Scrutiny adapts online degradation sampling using category-wise severity feedback, increasing SevAcc to 0.732. These results demonstrate the benefit of explicitly modeling dependencies among localization, recognition, and severity for joint fine-grained quality diagnosis.

Table 2: Global degradation diagnosis.

Method	ClsAcc	SevAcc
AgenticIR	0.426	0.237
Q-Insight	0.434	0.365
VisualQuality-R1	0.482	0.359
GS-IQA-Scrutiny	**0.916**	**0.735**

Global Degradation Diagnosis. We further evaluate the diagnostic capability of GS-IQA beyond localization on globally degraded images, where each image contains a single distortion with uniform severity. This setting removes the grounding requirement and isolates the *What* and *How* capabilities. As shown in Table 2, GS-IQA-Scrutiny substantially outperforms existing IQA methods, improving the best competing ClsAcc from 0.482 to 0.916 and SevAcc from 0.365 to 0.735. These results demonstrate that the advantage of GS-IQA extends beyond spatial grounding to distortion recognition and fine-grained severity discrimination.

4.3 Generalization and Transferability

Cross-Dataset Diagnostic Generalization. We evaluate the transferability of the learned distortion semantics and severity cues beyond Diag-Bench on DQ-495K, KADID-10K, TID2013, and Waterloo, without dataset-specific fine-tuning. Since these benchmarks provide image-level annotations, evaluation focuses on the transferable *What* and *How* dimensions using the distortion categories and five-level severity labels that can be mapped to our taxonomy. As shown in Table 3, GS-IQA consistently outperforms Q-Insight and VisualQuality-R1 across all four datasets, improving ClsAcc over the strongest baseline by 50.3, 41.6, 29.3, and 32.8 points, respectively, while achieving gains of up to 14.5 points in SevAcc. These consistent improvements demonstrate that GS-IQA learns transferable distortion and severity representations that generalize beyond the content distribution and synthesis patterns of Diag-Bench.

Holistic Quality Prediction Transferability. We further evaluate the transferability of the fine-grained representations learned through *Where–What–How* diagnosis to conventional global IQA. Following the training protocol of VisualQuality-R1, we initialize a global quality regressor from

Table 5: Ablation studies of GS-IQA, including (a) the perception-gated severity reward in the initial glance stage, and (b) reward-conditioned online generation in the scrutiny stage.

Configuration	Precision	Recall	F1	mIoU	ClsAcc	SevAcc
(a) Severity Reward in Glance Stage						
Hard Binary Reward	0.955	0.800	0.871	0.960	0.939	0.652
Continuous Linear Reward	0.954	0.879	0.915	0.953	0.926	0.627
Ordinal Ranking Reward	0.966	0.872	0.916	0.960	0.924	0.304
Perception-gated Reward	**0.948**	**0.848**	**0.895**	**0.953**	**0.921**	**0.678**
w/o Perception Gating	0.950	0.816	0.878	0.957	0.927	0.651
(b) Online Generation in Scrutiny Stage						
Static Data Only	0.941	0.849	0.893	0.957	0.945	0.682
Random Online Generation	0.967	0.897	0.931	0.967	0.947	0.694
Reward-conditioned Generation	**0.953**	**0.909**	**0.930**	**0.967**	**0.947**	**0.732**
w/o glance initialization	0.957	0.854	0.902	0.960	0.914	0.665

GS-IQA and fine-tune it using the same training data, dataset splits, and regression objective. As shown in Table 4, GS-IQA achieves the best average performance across five external benchmarks, reaching 0.860 SRCC and 0.869 PLCC. Compared with VisualQuality-R1, the average SRCC and PLCC improve by 0.031 and 0.015, respectively, with GS-IQA ranking first on BID, CLIVE, and KonIQ. These results indicate that Where–What–How learning captures transferable perceptual representations that can be effectively aggregated into calibrated image-level scores. GS-IQA therefore supports both interpretable local diagnosis and conventional global quality prediction.

4.4 Ablation Studies

Impact of Severity Reward. Table 5(a) examines different reward formulations for severity learning in the Glance stage. The proposed perception-gated reward achieves the highest SevAcc of 0.678, outperforming the hard binary reward (0.652) while maintaining competitive localization performance. Continuous linear and ordinal ranking rewards (detailed in Supplementary materials) retain strong F1 scores but yield lower SevAcc of 0.627 and 0.304, respectively, highlighting the importance of reward design for exact severity calibration. Removing perception gating further reduces both F1 (0.895→0.878) and SevAcc (0.678→0.651), confirming that fine-grained severity feedback benefits from reliable *Where–What* associations.

Impact of Online Generation. Table 5(b) evaluates data strategies in the scrutiny stage. Training exclusively on static data leads to a performance bottleneck, while online generation overcomes this limit by introducing dynamic perceptual challenges. Among online strategies, our reward-conditioned generation achieves a higher SevAcc of 0.732 than random generation, indicating that category-wise severity feedback better targets distortions that remain difficult to grade. Removing Glance initialization consistently degrades all three diagnostic tasks, confirming the value of static perception as a stronger initialization for adaptive refinement. Additional analyses of severity-reward evolution and adaptive sampling are provided in Appendix.

Sample Efficiency and Model Scaling. After Glance-stage training, Scrutiny with only 50 pristine source images improves SevAcc from 0.678 to 0.689. Scaling the full GS-IQA model from 4B to 32B further increases SevAcc from 0.732 to 0.754. Detailed source-pool and model-scaling results are provided in Appendix A.3.4.

5 Conclusion

In this paper, we present GS-IQA, a fine-grained image quality diagnosis framework that progressively reasons about degradations from an initial glance to closer scrutiny. Rather than collapsing heterogeneous degradations into a single score, GS-IQA characterizes each along the *Where–What–How* dimensions of region, type, and severity. Recognizing their dependencies, we adopt a two-stage reinforcement learning scheme: perception-gated rewards first consolidate localization and recognition, while reward-conditioned online degradation generation supplies targeted hard cases for severity refinement. Extensive experiments show that GS-IQA outperforms existing MLLM-based IQA methods in fine-grained diagnosis, generalizes to external benchmarks without dataset-specific fine-tuning, and transfers effectively to holistic quality prediction. Fine-grained diagnosis thus proves not a cost paid for interpretability, but a route to quality representations that serve both local diagnosis and generalizable image-level prediction.

6 AI USE STATEMENT

Generative AI tools were used to support manuscript preparation, including drafting, translation, and language polishing, and to facilitate the retrieval and discovery of relevant literature, as well as code implementation and debugging. The authors retain full responsibility for verifying all AI-assisted material and ensuring the accuracy and integrity of the paper's scientific content and citations.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025b.

G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

Yue Cao, Zhaolin Wan, Dongwei Ren, Zifei Yan, and Wangmeng Zuo. Incorporating semi-supervised and positive-unlabeled learning for boosting full reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5851–5861, 2022.

Chaofeng Chen, Sensen Yang, Haoning Wu, Liang Liao, Zicheng Zhang, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Q-Ground: Image quality grounding with large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 486–495, 2024a.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Zewen Chen, Juan Wang, Wen Wang, Sunhan Xu, Hang Xiong, Yun Zeng, Jian Guo, Shuxun Wang, Chunfeng Yuan, Bing Li, et al. Seagull: No-reference image quality assessment for regions of interest via vision-language instruction tuning. *arXiv preprint arXiv:2411.10161*, 2024b.

Zheng Chen, Xun Zhang, Wenbo Li, Renjing Pei, Fenglong Song, Xiongkuo Min, Xiaohong Liu, Xin Yuan, Yong Guo, and Yulun Zhang. Grounding-IQA: Grounding multimodal language model for image quality assessment. In *International Conference on Learning Representations*, volume 2026, pp. 39839–39851, 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1780–1789, 2020.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Ziheng Jia, Jiaying Qian, Zicheng Zhang, Zijian Chen, and Xiongkuo Min. Refine-IQA: Multi-stage reinforcement finetuning for perceptual image quality assessment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-Insight: Understanding image quality via visual reinforcement learning. *Proceedings of the Advances in Neural Information Processing Systems*, 2025.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. GroundingGPT: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6657–6678, 2024.

Wenjie Liao, Jieyu Yuan, Yifang Xu, Chunle Guo, Zilong Zhang, Jihong Li, Jiachen Fu, Haotian Fan, Tao Li, Junhui Cui, et al. ViDA-UGC: Detailed image quality analysis via visual distortion assessment for ugc images. *arXiv preprint arXiv:2508.12605*, 2025.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10k: A large-scale artificially distorted iqa database. In *International Conference on Quality of Multimedia Experience*, pp. 1–3, 2019.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025a.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *2025 IEEE/CVF International Conference on Computer Vision*, pp. 2034–2044, 2025b.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12826–12835, 2020.

Xinge Peng, Yiting Lu, Xin Li, and Zhibo Chen. IQA-Spider: Unifying multi-granularity image quality assessment with reasoning, grounding and referring. *arXiv preprint arXiv:2605.24553*, 2026.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. PIEAPP: Perceptual image-error assessment through pairwise preference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1808–1817, 2018.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3667–3676, 2020.

Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-Align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25490–25500, 2024a.

Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. Towards open-ended visual quality comparison. In *European Conference on Computer Vision*, pp. 360–377, 2024b.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. VisualQuality-R1: Reasoning-induced image quality assessment via reinforcement learning to rank. *Proceedings of the Advances in Neural Information Processing Systems*, 2025.

Qingyang Yan, Guangyao Chen, and Yixiong Zou. Start small, think big: Curriculum-based relative policy optimization for visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11550–11558, 2026.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

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 *European Conference on Computer Vision*, pp. 259–276, 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 Computer Vision and Pattern Recognition Conference*, pp. 14483–14494, 2025a.

Zhiyuan You, Jinjin Gu, Xin Cai, Zheyuan Li, Kaiwen Zhu, Chao Dong, and Tianfan Xue. Enhancing descriptive image quality assessment with a large-scale multi-modal dataset. *IEEE Transactions on Image Processing*, 2025b.

Kai Zhang, Wangmeng Zuo, and Lei Zhang. FFDNet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018b.

Zicheng Zhang, Tengchuan Kou, Shushi Wang, Chunyi Li, Wei Sun, Wei Wang, Xiaoyu Li, Zongyu Wang, Xuezhi Cao, Xiongkuo Min, et al. Q-eval-100k: Evaluating visual quality and alignment level for text-to-vision content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10621–10631, 2025a.

Zicheng Zhang, Haoning Wu, Ziheng Jia, Weisi Lin, and Guangtao Zhai. Teaching lmms for image quality scoring and interpreting. *arXiv preprint arXiv:2503.09197*, 2025b.

Kaiwen Zhu, Jinjin Gu, Zhiyuan You, Yu Qiao, and Chao Dong. An intelligent agentic system for complex image restoration problems. In *International Conference on Learning Representations*, 2025.

A APPENDIX

This supplementary material provides additional details and results to complement the main paper. In particular, it includes:

- Section A.1. Additional Methodology Details.
- Section A.2. Diag-Bench Construction and Analysis.
- Section A.3. Additional Experimental Results.
- Section A.4. Additional Qualitative Results.
- Section A.5. Discussion and Limitations.

A.1 ADDITIONAL METHODOLOGY DETAILS

A.1.1 PROMPT DESIGNS

We adopt a structured prompt design for region-level image quality diagnosis. As summarized in Table 6, the model is instructed to output one or more structured triplets, each corresponding to a degraded region with its location, distortion category, and severity level. Such a design makes the output directly parseable and allows the extracted region-level predictions to be readily used for subsequent reward computation, evaluation, and downstream applications. For models with spatial grounding capability (Comanici et al., 2025; Hurst et al., 2024), we use the same prompt format during evaluation to ensure consistency in structured outputs. For holistic IQA models that only support global image-level assessment (Li et al., 2025; Wu et al., 2025; Zhu et al., 2025), we follow the prompt setting in their original works, while adapting the label space to match the distortion categories and severity levels used in our framework.

Table 6: Prompt design for structured image quality diagnosis. The prompt instructs the model to detect degraded local regions and predict their coordinates, distortion categories, and severity levels in a structured format.

Component	**Description**
Objective	This is an image quality degradation detection task. The input image may contain one or more local regions affected by degradation. The goal is to identify all degraded regions and output their corresponding coordinates, distortion types, and severity levels.
Reasoning Format	Before producing the final answer, the model is required to perform an internal reasoning process enclosed in `<think>...</think>`, followed by the final visible answer enclosed in `<answer>...</answer>`.
Serialized Output	The final answer consists of one or more structured triplets, each written as `<ground>[x1,y1,x2,y2]</ground><cls>type</cls><severity>level</severity>`, where each triplet corresponds to one degraded region and the coordinates are normalized to integers in $[0, 1000]$.
Label Space	The distortion category is selected from the predefined set {`none, noise, blur, compression, overexposure, underexposure, high contrast, low contrast, oversaturate, desaturate, oversharpen, pixelate, quantization`}. The severity is an ordinal label in $\{1, \ldots, 5\}$, with 0 reserved for the `none` category.

A.1.2 DETAILS OF DEGRADATION GROUNDING REWARD

Degradation grounding in GS-IQA operates on unordered triplet sets. Let $\hat{\mathcal{T}} = \{\hat{t}_i\}_{i=1}^{N}$ and $\mathcal{T} = \{t_j\}_{j=1}^{M}$ denote the predicted and ground-truth sets, respectively, with $\hat{t}_i = (\hat{b}_i, \hat{c}_i, \hat{s}_i)$ and $t_j = (b_j, c_j, s_j)$. Here, b, c, and s denote the bounding box, distortion category, and severity level. To establish correspondence between the two sets, we define the pairwise matching score as

$$\Phi(\hat{t}_i, t_j) = \lambda_{\text{giou}} \, \text{GIoU}(\hat{b}_i, b_j) + \lambda_{\text{cls}} \mathbb{I}(\hat{c}_i = c_j) - \lambda_{\text{L1}} \|\hat{b}_i - b_j\|_1, \tag{10}$$

where the three terms measure spatial alignment, category agreement, and coordinate discrepancy, respectively. The coefficients λ_{giou}, λ_{cls}, and λ_{L1} balance their contributions. The optimal bipartite assignment maximizes the total matching score:

$$\hat{\sigma} = \arg \max_{\sigma \in \Omega_{N,M}} \sum_{i=1}^{N} \Phi(\hat{t}_i, t_j), \tag{11}$$

Algorithm 1 Two-stage training of GS-IQA

Require: Training set $\mathcal{D} = \{(x, \mathcal{T})\}$, pristine image pool $\mathcal{X}$, initial policy π_θ, reference policy π_{ref}, group size G, category set $\mathcal{C}$

Glance-Stage Training

for each training step **do**
  Sample a mini-batch $\{(x, \mathcal{T})\}$ from $\mathcal{D}$
  for each sample $(x, \mathcal{T})$ in the mini-batch **do**
    Construct the structured prompt for image x
    Sample a response group $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}$
    for each response o_i **do**
      Parse structured triplets $\hat{\mathcal{T}}_i$ from o_i
      Compute format reward $R_{\text{fmt}}(o_i)$
      Perform bipartite matching between $\hat{\mathcal{T}}_i$ and $\mathcal{T}$
      Compute grounding and severity rewards
      Compute total reward $R(o_i) = R_{\text{fmt}} + R_{\text{iou}} + R_{\text{cls}} + R_{\text{F1}} + R_{\text{sev}}$
    end for
    Normalize rewards within the group to obtain relative advantages
  end for
  Update π_θ with GRPO
end for

Scrutiny-Stage Training

Initialize category-wise performance profile $\{\hat{R}_c\}_{c \in \mathcal{C}}$
for each training step **do**
  Compute adaptive sampling distribution $P(c) = \frac{\exp(-\hat{R}_c/\tau)}{\sum_{c' \in \mathcal{C}} \exp(-\hat{R}'_c/\tau)}$
  Mix with a uniform prior $P'(c) = (1-\lambda)P(c) + \frac{\lambda}{|\mathcal{C}|}$
  Sample pristine image $x \sim \mathcal{X}$
  Sample degradation mode $m \in \{\text{None}, \text{Global}, \text{Local}\}$
  Sample category $c \sim P'(c)$ and severity level $s \in \{1, \ldots, 5\}$
  Synthesize degraded image and structured annotation $(x_{\text{deg}}, \mathcal{T}) \leftarrow g(x; m, c, s)$
  Construct the structured prompt for x_{deg}
  Sample a response group $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}$
  for each response o_i **do**
    Parse structured triplets from o_i
    Compute rewards as in Glance-Stage
  end for
  Normalize rewards within the group and update π_θ with GRPO
  for each category $c \in \mathcal{C}$ appearing in $\mathcal{T}$ **do**
    Compute batch-average severity reward $\bar{R}_c$
    Update EMA profile $\hat{R}_c \leftarrow (1-\alpha)\hat{R}_c + \alpha \bar{R}_c$
  end for
end for

where $\Omega_{N,M}$ contains one-to-one assignments of $\min(N, M)$ pairs, with each prediction and ground-truth triplet participating in at most one pair. The assignment is solved using the Hungarian algorithm (Kuhn, 1955). We then retain spatially valid pairs as $\mathcal{M} = \{(i, j) \in \hat{\sigma} \mid \text{IoU}(\hat{b}_i, b_j) \geq 0.5\}$. These pairs are used to compute the grounding rewards, while severity credit additionally requires category agreement $\hat{c}_i = c_j$.

A.1.3 Training Algorithms

Algorithm 1 summarizes the two-stage training procedure of GS-IQA under the proposed progressive distortion reasoning framework, which proceeds *from Glance to Scrutiny*. The glance stage focuses on learning structured degradation parsing from static supervision, while the scrutiny stage further refines severity perception through online reward-conditioned generation.

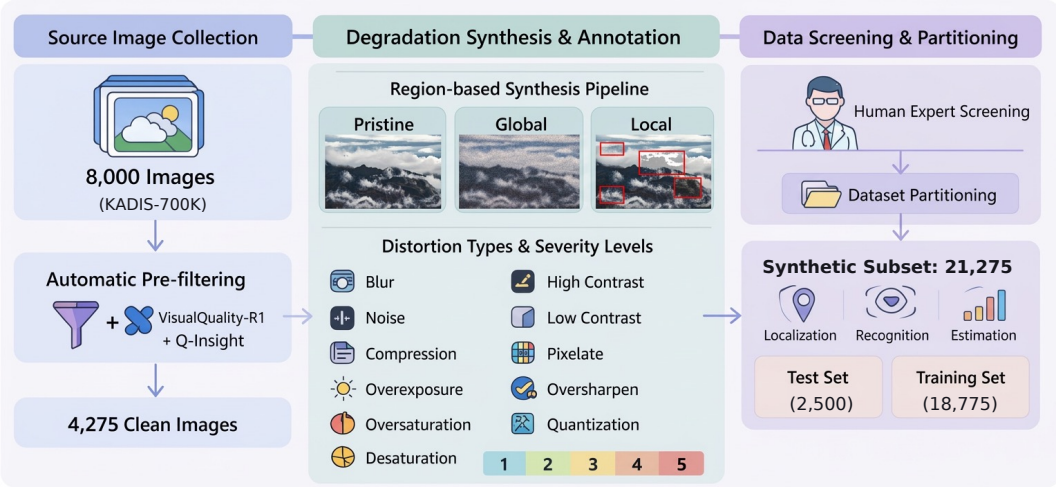


Figure 3: Construction pipeline of the synthetic subset of Diag-Bench. Source images undergo automatic and manual quality screening, followed by source-level partitioning before degradation synthesis. Synthesized samples are subsequently reviewed for perceptual clarity.

A.2 Diag-Bench Construction and Analysis

A.2.1 Distortion Definitions and Parameterization

The overall construction pipeline of Diag-Bench is shown in Figure 3. Within this pipeline, distortion synthesis is implemented with parameterized image processing operators, and each distortion category is calibrated into five ordinal severity levels. For all categories, the severity schedule is designed to increase monotonically with perceptual severity, so that higher levels correspond to progressively stronger degradations. In the local mode, both distortion category and severity are sampled independently for each degraded region, which enables one image to contain multiple heterogeneous artifacts with different severity.

- **Blur** is implemented as lens defocus by convolving the image with a disk kernel. Given an image x, the degraded image is defined as $y = x * k_r$, where k_r denotes a normalized disk kernel with radius r. The radius controls the degree of spatial smoothing, and the five severity levels correspond to $r \in \{2, 3, 4, 6, 8\}$. As r increases, local textures and edges become progressively less distinguishable.
- **Noise** is modeled as additive Gaussian noise. The degraded image is defined as $y = \mathrm{clip}(x + \epsilon, 0, 1), \epsilon \sim \mathcal{N}(0, \sigma^2)$, where σ is the noise standard deviation. The five severity levels are parameterized by $\sigma \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$. This schedule produces a gradual transition from weak grain to strong stochastic corruption.
- **Compression** artifacts are synthesized through JPEG encoding and decoding. Formally, $y = \mathrm{JPEGDecode}(\mathrm{JPEGEncode}(x; q))$, where q denotes the JPEG quality factor. Lower values of q lead to stronger blocking and quantization artifacts. The five severity levels are defined by $q \in \{25, 18, 12, 8, 5\}$.
- **Overexposure** is generated by gamma-based brightening in HSV space. The value channel is transformed as $V' = V^\gamma, y = \mathrm{HSV2RGB}(H, S, V')$, where $\gamma < 1$. Smaller values of γ yield stronger brightening and stronger saturation of high-intensity regions. The five severity levels are given by $\gamma \in \{0.70, 0.58, 0.47, 0.36, 0.25\}$.
- **Underexposure** is generated by gamma-based darkening in HSV space. The value channel is transformed as $V' = V^\gamma, y = \mathrm{HSV2RGB}(H, S, V')$, where $\gamma > 1$. Larger values of γ produce stronger attenuation of luminance and progressively suppress dark-region details. The five severity levels correspond to $\gamma \in \{1.50, 1.80, 2.20, 2.70, 3.50\}$.
- **High contrast** is synthesized by scaling image intensities away from the grayscale mean. Let $\mu = \mathrm{mean}(0.299R + 0.587G + 0.114B)$, where μ is the grayscale mean of the image. The contrast-enhanced image is then defined as $y = \mathrm{clip}(\alpha x + (1-\alpha)\mu, 0, 1)$, where $\alpha > 1$

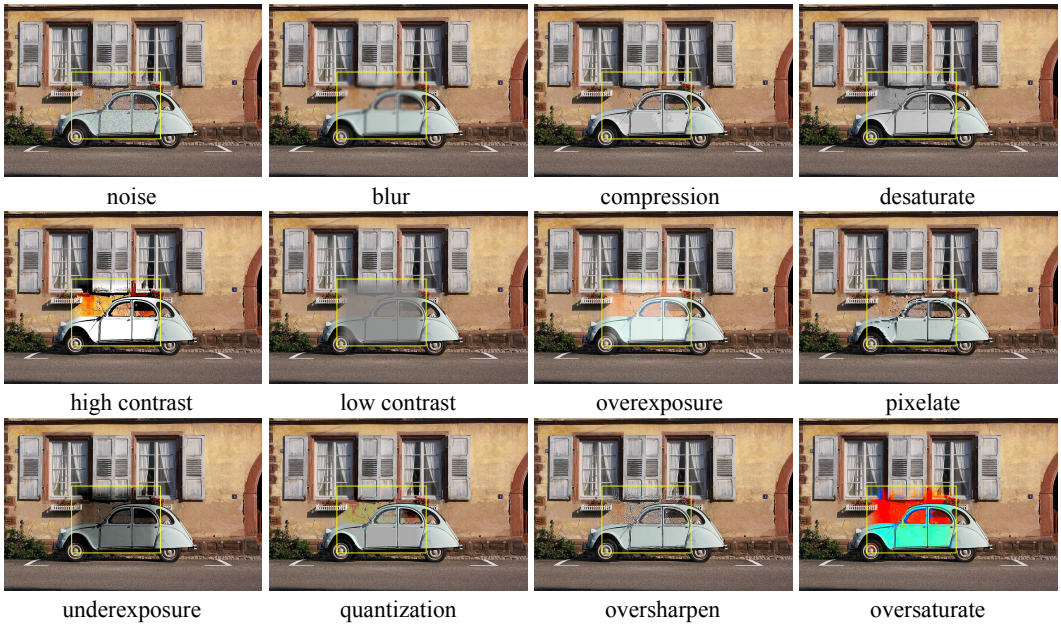


Figure 4: Distortion examples of the 12 distortion categories in Diag-Bench. We showcase one example for each distortion implementation.

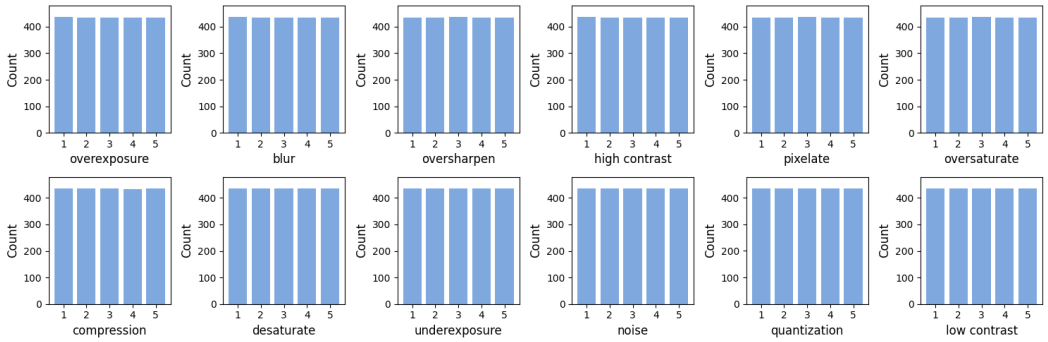


Figure 5: Distribution of severity levels across the 12 distortion categories in Diag-Bench. For each distortion category, the histogram reports the number of annotated degraded regions at severity levels 1 to 5.

is the contrast factor. Larger values of α amplify tonal separation and produce increasingly exaggerated contrast. The five severity levels are $\alpha \in \{1.4, 1.7, 2.1, 2.6, 4.0\}$.

- **Low contrast** is synthesized by compressing image intensities toward the grayscale mean. Using the same grayscale mean μ, the degraded image is defined as $y = \text{clip}(\alpha x + (1 - \alpha)\mu, 0, 1)$, where $0 < \alpha < 1$. Smaller values of α yield stronger flattening of local intensity variation. The five severity levels are $\alpha \in \{0.75, 0.60, 0.45, 0.30, 0.20\}$.
- **Oversaturation** is implemented in YCbCr space by amplifying the deviation of the chroma channels from their neutral values. Let (Y, Cb, Cr) denote the YCbCr representation of the image. The chroma channels are transformed as $Cb' = 0.5 + \lambda(Cb - 0.5), Cr' = 0.5 + \lambda(Cr - 0.5)$, followed by conversion back to RGB space. Here $\lambda > 1$ controls the severity of chroma amplification. The five severity levels are $\lambda \in \{2, 3, 5, 8, 16\}$. Larger values of λ produce increasingly vivid and exaggerated colors.
- **Desaturation** is also implemented in YCbCr space, but the chroma channels are compressed toward their neutral values. The transformation is written as $Cb' = 0.5 + \lambda(Cb - 0.5), Cr' = 0.5 + \lambda(Cr - 0.5)$, where $0 \leq \lambda \leq 1$. Smaller values of λ remove more chromatic information and move the image closer to grayscale. The five severity levels are $\lambda \in \{0.60, 0.40, 0.20, 0.10, 0.00\}$.

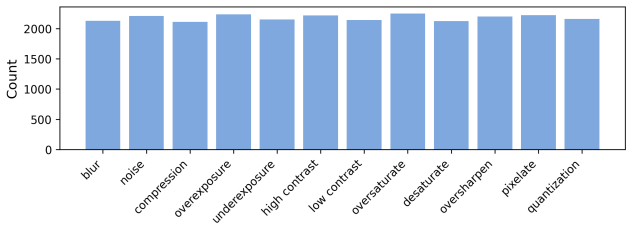


Figure 6: Distribution of degraded regions across distortion categories in Diag-Bench.

- **Oversharpening** is generated by unsharp masking. A blurred version of the image is first obtained as $b = G_{\sigma=1.1}(x)$, where $G_{\sigma=1.1}$ denotes Gaussian smoothing with standard deviation 1.1. The sharpened image is then computed as $y = \mathrm{clip}((1+a)x - ab, 0, 1)$, where a is the sharpening amount. Larger values of a produce stronger edge enhancement and more visible halo artifacts. The five severity levels are $a \in \{2.0, 2.8, 4.0, 6.0, 8.0\}$.
- **Pixelation** is synthesized by first downsampling the image and then resizing it back to the original resolution with nearest-neighbor interpolation. Formally, $y = \mathrm{Up}_{\mathrm{NN}}(\mathrm{Down}_{\mathrm{NN}}(x; pH, pW); H, W)$, where H and W are the original image height and width, and p is the spatial scale ratio. Smaller values of p correspond to coarser block structures and stronger loss of spatial detail. The five severity levels are $p \in \{0.50, 0.40, 0.30, 0.25, 0.20\}$.
- **Quantization** is implemented by discretizing pixel intensities into a limited number of bins. Let n denote the number of quantization bins. The degraded image is defined as $y = \mathrm{clip}\left(\frac{\mathrm{round}(x(n-1))}{n-1}, 0, 1\right)$. Smaller values of n produce stronger banding and color discontinuities. The five severity levels are $n \in \{24, 16, 8, 6, 4\}$.

Overall, these distortion operators cover a broad range of common image quality degradations, including spatial smoothing, additive perturbation, compression artifacts, exposure imbalance, contrast distortion, chromatic distortion, oversharpening, spatial resolution degradation, and color discretization. Representative examples of the 12 distortion categories are shown in Figure 4, which illustrates the characteristic visual patterns associated with each degradation type. The monotonic parameter schedules ensure that Diag-Bench provides consistent supervision for both distortion recognition and ordinal severity estimation under the structured *Where-What-How* formulation.

A.2.2 Dataset Statistics and Distribution Analysis

To further characterize Diag-Bench, we analyze its statistical composition from both the category and severity perspectives. Figure 6 reports the region-level distribution of the 12 distortion categories in the full dataset. The category frequencies are approximately balanced, which reduces the risk of bias toward a small subset of dominant distortions during training and evaluation. Figure 5 further presents the region-level distribution of severity levels within each distortion category. The numbers of annotated degraded regions are generally balanced across the five severity levels, providing comparable supervision for ordinal severity estimation within each category. This design helps prevent the benchmark from being dominated by only mild or severe cases, and instead supports fine-grained evaluation across the full severity spectrum.

A.3 Additional Experimental Results

A.3.1 Alternative Severity Reward Formulations

We provide the alternative severity rewards used in the ablation study, following the notation in the main paper. For each spatially matched pair $(i, j) \in \mathcal{M}$, $d_{ij}^{\mathrm{sev}} = |\hat{s}_i - s_j|$ denotes the ordinal severity distance. The binary and linear variants replace r_{ord} in R_{sev} while retaining the same matching, category gating, and normalization. The ranking variant instead compares relative severity across category-correct matched regions.

Hard Binary Reward. Only an exact severity match receives credit:

$$r_{\mathrm{bin}}\left(d_{ij}^{\mathrm{sev}}\right) = \mathbb{I}\left(d_{ij}^{\mathrm{sev}} = 0\right). \tag{12}$$

Table 7: Ablation on tiered ordinal rewards in the Glance stage. Each tuple lists the values of r_{ord} for $d_{ij}^{\text{sev}} = 0$, 1, 2, and $d_{ij}^{\text{sev}} \geq 3$, respectively. We vary (α_1, α_2) while keeping the remaining reward formulation unchanged. Bold indicates the default setting.

Tiered rewards $(1, \alpha_1, \alpha_2, 0)$	**Precision**	**Recall**	**F1**	**mIoU**	**ClsAcc**	**SevAcc**
$[1.0, 0.5, 0, 0]$	0.955	0.853	0.901	0.953	0.927	0.659
$[1.0, 0.4, 0.1, 0]$	0.952	0.850	0.899	0.953	0.925	0.666
$[\mathbf{1.0}, \mathbf{0.6}, \mathbf{0.2}, \mathbf{0}]$	0.948	0.848	0.895	0.953	0.921	0.678
$[1.0, 0.7, 0.3, 0]$	0.946	0.847	0.894	0.952	0.920	0.671
$[1.0, 0.8, 0.4, 0]$	0.944	0.845	0.892	0.952	0.918	0.660

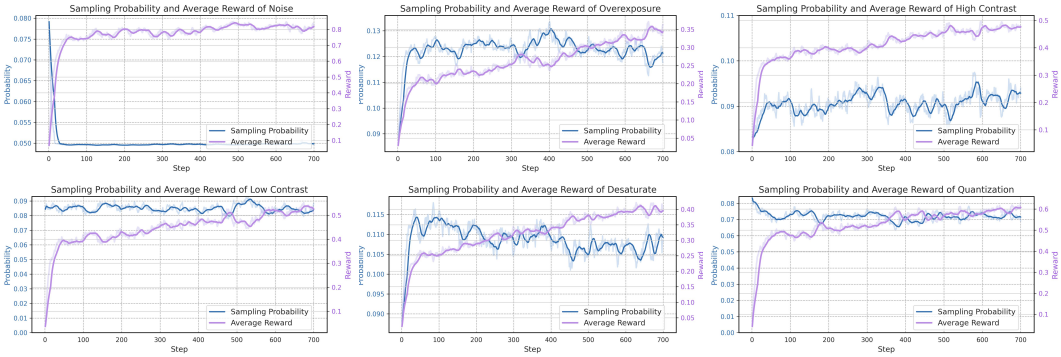


Figure 7: Dynamics of EMA-updated average severity reward and sampling probability for representative categories in the scrutiny stage.

Continuous Linear Reward. Credit decreases linearly with the ordinal distance:

$$r_{\text{lin}}\left(d_{ij}^{\text{sev}}\right) = \max\left(0,\, 1 - \frac{d_{ij}^{\text{sev}}}{4}\right). \tag{13}$$

Ordinal Ranking Reward. For two distinct matched pairs $(i, j), (k, \ell) \in \mathcal{M}$ within the same image, with $\hat{c}_i = c_j$ and $\hat{c}_k = c_\ell$, the pairwise reward is

$$r_{\text{rank}}(\hat{s}_i, s_j, \hat{s}_k, s_\ell) = \mathbb{I}[(s_j - s_\ell)(\hat{s}_i - \hat{s}_k) \geq 0]\,. \tag{14}$$

This reward evaluates relative ordering rather than absolute severity accuracy, assigning unit credit when the predicted and ground-truth differences do not have opposite signs.

A.3.2 Ablation on Tiered Severity Reward Weights

We further analyze the design of the tiered severity reward in the glance stage by varying the weights assigned to near-miss predictions while keeping the overall reward formulation unchanged. Specifically, we adjust the rewards corresponding to $d_{ij}^{\text{sev}} = 1$ and $d_{ij}^{\text{sev}} = 2$ and evaluate how different weight schedules affect both severity estimation and the overall structured diagnosis performance. As shown in Table 7, the default setting used in GS-IQA achieves the best overall trade-off. More sparse designs provide weaker supervision for near-miss severity predictions, while more tolerant designs reduce the discrimination between adjacent severity levels. These results indicate that the default tiered weights offer a balanced severity signal, which benefits not only severity estimation but also the overall optimization of structured quality diagnosis.

A.3.3 Reward-Sampling Dynamics in the Scrutiny Stage

Severity Reward Evolution. Figure 8 shows the evolution of severity rewards during the Scrutiny stage. The overall reward rises steadily, with consistent improvements observed for representative distortions such as overexposure, high contrast, low contrast, desaturation, and quantization. This behavior reflects the intended effect of reward-conditioned sampling: categories with weaker severity estimation continue to receive additional training exposure, enabling progressive refinement rather than early saturation on a fixed training distribution.

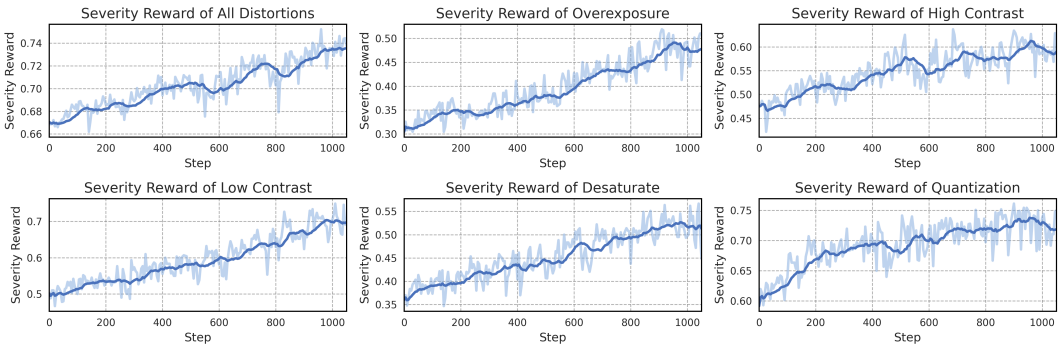


Figure 8: Learning dynamics of severity perceptual refinement in GS-IQA-Scrutiny.

Reward–Sampling Relationship. We further visualize the relationship between category-wise severity reward and sampling probability throughout training. The goal of this analysis is to examine whether the online generator indeed allocates more training samples to distortion categories that remain difficult for the current model. Results in Figure 7 show a clear correspondence between the two quantities. Categories with lower severity reward are assigned higher sampling probabilities, whereas categories with improved reward gradually receive less emphasis. When the reward reaches a stable level, its sampling probability decreases to a floor of 0.004 and remains constant throughout training, ensuring a minimal level of exploration even for categories that are already well learned. This trend indicates that the sampling distribution is not static, but evolves in response to the model's current weaknesses. In this way, the generator forms an implicit hard-case curriculum that continuously shifts attention toward categories with insufficient severity perception. Such dynamics are consistent with the design of the scrutiny stage. Rather than relying on a fixed synthetic distribution, the online generator adapts its sampling policy according to reward feedback from the current model. As a result, training remains focused on those distortion categories for which severity estimation is still unreliable. This behavior provides additional evidence that the gains of GS-IQA in the scrutiny stage arise from targeted reward-conditioned generation, rather than from simply introducing more synthetic data.

A.3.4 Sample Efficiency and Model Scaling

We further analyze GS-IQA under different pristine source-pool sizes and model capacities. The source-pool sizes below refer only to online degradation synthesis during the Scrutiny stage, rather than the static training data used in the Glance stage.

Sample Efficiency of Generative Scrutiny. Table 8(a) investigates the sample efficiency of reward-conditioned generation under different source-pool sizes. Notably, with only 50 source images (1% of the training pool), GS-IQA already achieves an F1 score of 0.902, outperforming the Glance-stage baseline trained on static data. As the source pool expands from 50 to 3,775 images, performance improves steadily from 0.902 to 0.930 in F1 and from 0.689 to 0.732 in SevAcc. These results show that online degradation synthesis can effectively reuse limited pristine content to produce diverse region–category–severity supervision, while broader source diversity further benefits fine-grained severity learning. This makes the Scrutiny stage substantially less dependent on manually constructed training samples and enables scalable refinement from a compact pool of pristine images.

Table 8: Ablation on sample efficiency and model scaling.

Configuration	F1	ClsAcc	SevAcc
(a) Sample Efficiency			
50 (1%)	0.902	0.923	0.689
500 (13%)	0.915	0.932	0.705
1,000 (26%)	0.922	0.939	0.718
2,000 (53%)	0.927	0.943	0.727
3,775 (100%)	**0.930**	**0.947**	**0.732**
(b) Model Capacity			
GS-IQA-4B	0.930	0.947	0.732
GS-IQA-8B	0.944	0.942	0.740
GS-IQA-32B	**0.952**	**0.956**	**0.754**

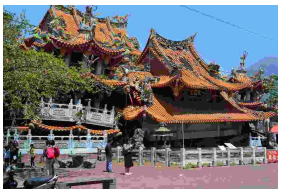
Scaling Analysis of Model Capacity. Table 8(b) shows that increasing model capacity yields diminishing gains in localization but more consistent improvements in severity estimation. As the model scales from 4B to 32B, F1 increases from 0.930 to 0.952, while SevAcc improves from 0.732 to 0.754. This trend suggests that fine-grained severity discrimination benefits more from increased model capacity than spatial localization, further highlighting severity estimation as the more challenging component of the *Where–What–How* diagnosis.

A.4 Additional Qualitative Results

Figure 9 presents additional qualitative comparisons on distortion cases. Compared with competing methods, GS-IQA produces predictions that are more consistent with the ground truth in terms of distortion type, spatial extent, and severity level. In particular, GS-IQA more reliably identifies the dominant degradation, avoids spurious distortion labels, and yields more accurate region-level grounding when localized artifacts are present. These examples further demonstrate that GS-IQA provides more precise and interpretable structured diagnosis, especially in cases that require joint reasoning over distortion category, severity, and spatial location.

A.5 Discussion and Limitations

GS-IQA extends image quality assessment from holistic judgment to structured diagnosis by explicitly predicting degraded regions together with their categories and severity levels. This formulation improves interpretability and makes the output more actionable for downstream applications such as restoration guidance and diagnostic feedback. The progressive training design further enables the model to move from structured degradation parsing to finer-grained severity reasoning. One limitation is that GS-IQA focuses on common and predefined distortion categories. While these categories already cover a broad range of practical image degradation scenarios, the diagnosis setting remains closed-set and does not directly support unknown distortions outside the predefined label space. In future work, we will extend GS-IQA toward open-set or open-vocabulary degradation diagnosis, potentially with more flexible label representations, richer textual supervision, and broader real-world distortion data.



Q-Insight

<think> The image has a slightly pixelated appearance and a low level of detail. It's possible that some of these effects could be due to file compression or file corruption, but the image isn't noticeably noisy, blurry, overexposed, or underexposed. The colors are saturated, which might suggest oversaturate, but the image also shows a slight lack of fine detail and potentially slightly reduced contrast or detail. </think>
<answer>{distortion class: compression, severity: serious}</answer>

GPT-4o

<think> The image contains one or more degradation artifacts, and region-wise analysis is therefore required. From a global perspective, the degraded areas are mainly concentrated in specific local regions. Visual inspection further identifies the degradation types as noise and compression artifacts. The degradation level is then assessed based on the affected region, degradation type, and severity. </think><answer>
<ground>[200, 100, 600,500]</ground><cls>compression</cls><severity>3</severity>
<ground>[50, 450, 300, 700]</ground><cls>noise</cls><severity>2</severity>
</answer>

Gemini-3-Pro

<think> I examined the image to identify potential quality degradations. I observed a uniform presence of block-like artifacts and a loss of fine detail across the entire visual field, which are characteristic of heavy lossy compression. Based on the severity of these artifacts, I classified the degradation as compression with a high intensity level and defined the region as covering the full image. </think> <answer>
<ground>[0, 0, 1000, 1000]</ground><cls>compression</cls><severity>5</severity>
</answer>

GS-IQA (ours)

<think> Significant blocky artifacts and detail breakage are visible in textured regions such as rooftops, railings, and crowds. Edges show noticeable ringing and jaggedness, while fine textures are heavily smoothed and nearly unrecognizable. These severe compression artifacts substantially degrade the main structures and scene details, supporting compression, level 5. </think>
<answer> <ground>[0, 0, 1000, 1000]</ground><cls>compression</cls><severity>5</severity> </answer>

GT: <ground>[0, 0, 1000,1000]</ground><cls>compression</cls><severity>5</severity>



Q-Insight

<think> The image quality is generally clear and detailed across most regions. However, the central dark area shows slight blurriness and minor pixelation, which may indicate a small loss of detail caused by compression. Aside from this localized effect, there is no clear evidence of noise, overexposure, underexposure, or significant contrast issues in the image. </think>
<answer>{distortion class: compression, severity: moderate}</answer>

GPT-4o

<think> 1. Observe the degradation signs in the image: a prominent black rectangular region is visible against the sky-blue background, forming a strong contrast with the surrounding normal areas.
2. Determine the degradation type: the black rectangle may indicate overexposure.
3. Identify the coordinates of the degraded region: the black rectangle is located in the upper part of the image. When mapped to the [0,1000] coordinate system, the estimated region is [600, 300, 900, 500].
4. Analyze the degradation severity: due to the obvious difference between the color regions, it is considered severity level 4.
5. Prepare to output all extracted information.</think>
<answer><ground>[600, 300, 900, 500]</ground><cls>overexposure</cls><severity>4</severity></answer>

Gemini-3-Pro

<think> Observe the degradation signs in the image: A prominent black rectangular region is visible in the sky-blue background, forming a strong contrast with the surrounding normal areas.
Determine the degradation type: The black rectangle may indicate overexposure.
Identify the coordinates of the degraded region: The black rectangle is located in the upper part of the image. When mapped to the [0,1000]coordinate system, the estimated region is [600,300,900,500].
Analyze the degradation severity: Due to the obvious difference between the color regions, it is considered severity level 4.</think>
<answer> <ground> [306, 250, 636, 513] <ground><cls> underexposure </cls><severity>5</severity> </answer>

GS-IQA (ours)

<think> The image shows clear signs of degradation, with a prominent black rectangular region appearing against the sky-blue background and forming a strong contrast with the surrounding normal areas. This abnormal region may indicate underexposure. The degradation severity is assessed as level 3, indicating a relatively severe but not extreme level. </think>
<answer> <ground>[294, 238, 653, 539]</ground><cls>underexposure</cls><severity>3</severity> </answer>

GT: <ground>[290, 238, 653, 538]</ground><cls>underexposure</cls><severity>3</severity>

Figure 9: Qualitative comparison of GS-IQA against competing methods.