# ROSA-IQA: LEARNING WITH REASONING FOR IMAGE QUALITY ASSESSMENT VIA ON-POLICY SELF-DISTILLATION

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

## ABSTRACT

Image quality assessment (IQA) takes two complementary forms: quality scoring quantifies perceptual quality, while quality description reveals the attributes and defects underlying human judgments. Although multimodal large language models (MLLMs) show promising reasoning capabilities, adapting them for quality description remains constrained by the fact that widely used IQA datasets typically provide only a scalar score per image. On-policy reinforcement learning directly learns from quality scores but provides only coarse sequence-level credit, whereas off-policy dense distillation from external MLLMs introduces a distribution shift between teacher demonstrations and student rollouts. We introduce a **R**easoning-based **O**n-policy **S**elf-distillation **A**pproach for IQA (ROSA-IQA). Under a teacher-student distillation paradigm, ROSA-IQA uses a frozen snapshot of the MLLM being optimized as the self-teacher, conditioned on curated privileged quality scores and segmentation-derived region guidance, to provide token-level supervision for student-sampled reasoning rollouts. We further use factual–counterfactual disagreement to identify tokens sensitive to quality conditions, thereby emphasizing quality-relevant descriptions in the student's reasoning. A scoring head then maps the combined visual and textual representations to continuous quality predictions. Extensive experiments show that ROSA-IQA outperforms leading no-reference IQA (NR-IQA) models in quality scoring while generating human-aligned quality descriptions. Importantly, we show that learning with reasoning improves cross-dataset generalization, suggesting that it acts as a regularizer that encourages the model to learn transferable quality features.

## 1 INTRODUCTION

Image quality assessment (IQA) is commonly cast as *quality scoring*, which represents the perceived quality of images with scalar scores. Such scalar scores provide a direct basis for evaluating and developing a wide range of vision systems, spanning image acquisition (Fang et al., 2020), enhancement (Gu et al., 2020), and generation (Li et al., 2024). Early no-reference IQA (NR-IQA) methods are largely knowledge-driven, typically relying on handcrafted features derived from natural scene statistics (Mittal et al., 2012; 2013). With the advent of deep neural networks (DNNs), feature representation and quality prediction can be learned jointly end-to-end (Su et al., 2020; Zhang et al., 2020; Ke et al., 2021; Yang et al., 2022). Despite fitting training data well and achieving high correlation with human ratings, DNN-based IQA models still struggle to generalize across diverse distortion scenarios, motivating research on quality-aware pretraining (Madhusudana et al., 2022), multi-dataset training (Zhang et al., 2021), and continual adaptation (Zhang et al., 2023a).

In contrast, *quality description* provides fine-grained and interpretable insights into perceptual defects, their semantic context, and spatial locations. Such descriptions provide diagnostic cues for identifying failures and optimizing image processing pipelines (Zhai et al., 2026), ultimately improving end-user quality of experience. However, quality description remains far less explored than quality scoring, largely because existing IQA datasets lack human-written descriptions. Collecting such descriptions at scale through well-controlled formal annotation procedures is laborious and expensive. Moreover, unlike scalar ratings that can be aggregated into a mean opinion score (MOS), free-form natural language descriptions are difficult to consolidate into a unified supervision signal.



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

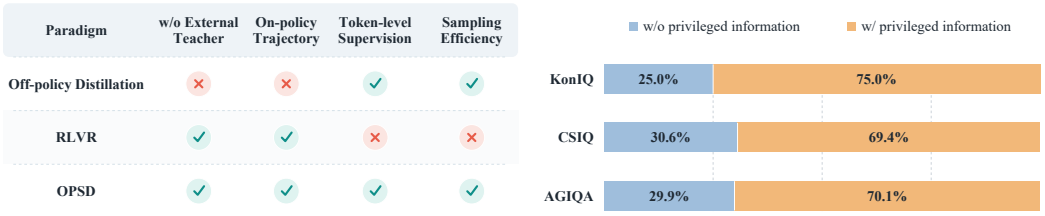



Figure 1: Left: Comparison of three paradigms for training reasoning-induced MLLMs. Right: Paired-comparison results from sample-efficient psychophysical testing (Feng et al., 2025) across KonIQ (Hosu et al., 2020), CSIQ (Larson & Chandler, 2010), and AGIQA (Li et al., 2024). The results show that incorporating privileged information substantially improves the alignment between MLLM-generated quality descriptions and human judgments. Further experimental details are provided in Appendix E.

Multimodal large language models (MLLMs) (Liu et al., 2023; Bai et al., 2025) offer a promising avenue by articulating perceptual quality in language and serving as proxy annotators. Q-Instruct constructs instruction-tuning data from human feedback to enhance MLLMs' visual quality understanding (Wu et al., 2024a); DepictQA uses LLM-assisted annotation to generate detailed quality descriptions (You et al., 2024); and DepictQA-Wild further scales descriptive supervision through MOS-informed MLLM annotation for diverse real-world IQA tasks (You et al., 2025b).

Despite the progress in quality scoring and quality description along their respective lines, reliably integrating both capabilities into a single model remains challenging (You et al., 2025a). One of the most straightforward approaches is off-policy distillation (Cai et al., 2025), where a potentially stronger MLLM (*e.g.*, a proprietary model such as GPT-4V) generates quality reasoning traces for supervised fine-tuning (SFT) of the student model. While providing dense token-level supervision, off-policy distillation suffers from exposure bias, potentially impairing generalization (Agarwal et al., 2024). Alternatively, reinforcement learning with verifiable rewards (RLVR) elicits reasoning directly from answer-level supervision. In NR-IQA, Q-Insight (Li et al., 2025) and VisualQuality-R1 (Wu et al., 2025) frame quality scoring as absolute score regression and relative quality ranking, respectively, and employ group relative policy optimization (GRPO) (Guo et al., 2025) to explore plausible reasoning paths that naturally serve as quality descriptions. However, RLVR provides only coarse sequence-level credit for reasoning and requires multiple rollouts for reward computation, incurring substantial computational overhead.

In this paper, we introduce ROSA-IQA, a **R**easoning-based **O**n-policy **S**elf-distillation **A**pproach for NR-IQA that combines on-policy learning with dense token-level supervision via on-policy self-distillation (OPSD) (Zhao et al., 2026). As summarized in Figure 1, OPSD samples a single on-policy reasoning trajectory, while a frozen copy of the model, equipped with curated privileged information, serves as a self-teacher that provides dense supervision along the trajectory. We use MOS and image regions extracted using SAM-2 (Ravi et al., 2025) as complementary privileged information. MOS provides a reliable global anchor for perceived quality, while the extracted regions provide local visual context for the self-teacher (Moorthy & Bovik, 2009; Siahaan et al., 2018). We further introduce *factual–counterfactual disagreement* to identify quality-sensitive tokens and focus distillation on quality-relevant reasoning. ROSA-IQA applies a scoring head to the final-layer visual-textual hidden representation to compute quality scores. Extensive experiments across diverse distortion scenarios demonstrate competitive quality scoring performance against other NR-IQA methods. Notably, reasoning-based training improves cross-dataset generalization, suggesting that reasoning trajectories serve as an effective regularizer for model training. We also show that ROSA-IQA generates human-aligned quality descriptions, as evaluated by an MLLM-as-a-judge (Lee et al., 2024) and a sample-efficient subjective testing method based on MAximum Discrepancy (MAD) Competition (Feng et al., 2025).

## 2 Related Work

This section briefly reviews IQA, focusing on recent advances in MLLM-based quality scoring and description, as well as the techniques used to train such models.

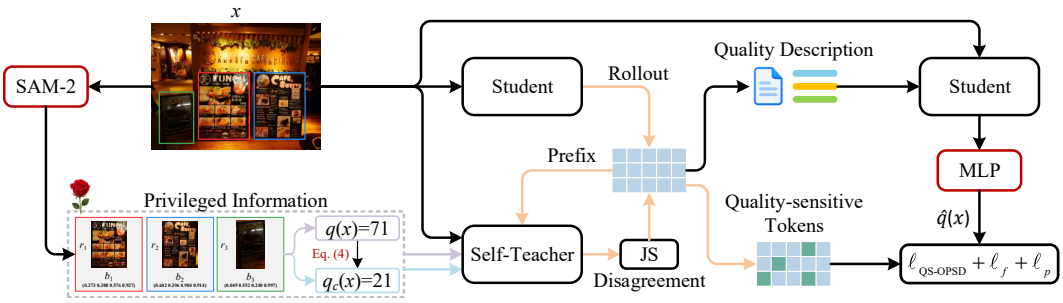



Figure 2: System diagram of the proposed ROSA-IQA. The student and self-teacher policies are instantiated from the same MLLM. Factual and counterfactual MOS values, together with local regions derived from SAM-2, are employed as privileged information, enabling the self-teacher to perform quality-sensitive distillation. Joint visual and textual representations are used for quality scoring.

**Image Quality Assessment.** IQA is broadly categorized into full-reference IQA (FR-IQA) and NR-IQA. FR-IQA has evolved from structure-aware comparisons (Wang et al., 2004) to deep perceptual metrics (Zhang et al., 2018; Ding et al., 2022), while NR-IQA has progressed from handcrafted natural-scene statistics (Mittal et al., 2012; 2013) to learned quality representations (Zhang et al., 2020; Su et al., 2020; Ke et al., 2021; Madhusudana et al., 2022). More recently, LIQE (Zhang et al., 2023b) and CLIP-IQA (Wang et al., 2023) leverage vision-language models as stronger backbones for improved generalization.

**MLLM-Based IQA.** MLLMs support both quantitative scoring and qualitative description through autoregressive generation. Scoring-oriented methods derive scalar estimates from text-defined levels (Wu et al., 2024b), score distributions (You et al., 2025a), or comparative judgments (Zhu et al., 2024), while description-oriented methods generate textual analyses of perceptual defects and quality comparisons (Wu et al., 2024a;c; You et al., 2024; 2025b). Training has evolved from off-policy imitation toward on-policy RLVR. Q-Ponder (Cai et al., 2025) uses off-policy teacher trajectories for cold-start SFT before GRPO, whereas Q-Insight (Li et al., 2025) and VisualQuality-R1 (Wu et al., 2025) directly optimize reasoning-based NR-IQA models with GRPO. Zoom-IQA (Liang et al., 2026) and IQA-T1 (Wu et al., 2026) further extend this paradigm with tool calling.

**On-Policy Self-Distillation.** On-policy distillation (Agarwal et al., 2024) trains on student-generated trajectories to reduce the train–inference distribution shift of off-policy distillation, while using an external teacher to provide dense token-level supervision unavailable in RLVR. OPSD (Zhao et al., 2026) instantiates teacher and student from the same model, conditioning only the teacher on privileged information. Subsequent work improves this paradigm by refining privileged contexts (Ye et al., 2026), increasing teacher supervision reliability (Stein et al., 2026), and focusing distillation on informative tokens (Xu et al., 2026). Building on this paradigm, we provide the self-teacher with MOS and semantically meaningful regions as privileged information and use factual–counterfactual disagreement to selectively distill quality-sensitive tokens.

## 3 METHOD

We propose ROSA-IQA, an NR-IQA model for both quality scoring and description. We specify two forms of privileged information within the OPSD paradigm and introduce a quality-sensitive distillation approach. Figure 2 illustrates the overall training pipeline of ROSA-IQA.

### 3.1 PRELIMINARIES

**Problem formulation.** Given an image $x$, a quality scoring function $\hat{q}(\cdot)$ is learned to produce a scalar estimate $\hat{q}(x)$ that agrees with its underlying perceived quality $q(x)$, which typically takes the form of MOS. By contrast, a quality description function aims to generate a language trajectory that qualitatively identifies and describes the attributes underlying the perceptual quality of $x$, which can

be formulated in the context of an MLLM as

$$o = (o_1, \ldots, o_T) \sim \pi_\theta(\cdot \mid x, c), \tag{1}$$

where $T$ is the trajectory length, $c$ is a text prompt, and $\pi_\theta$ denotes the autoregressive policy parameterized by $\theta$.

**On-policy self-distillation.** Following a teacher-student distillation paradigm, OPSD (Zhao et al., 2026) instantiates a learnable student policy $\pi_{\theta_S}$ and a teacher policy $\pi_{\theta_T}$ initialized from the same underlying model. Given $x$ and $c$, the student samples an on-policy trajectory $o$. At each position $t$, both policies compute next-token distributions conditioned on the same prefix $o_{<t}$. Compared to the student, the teacher is provided with additional privileged information $z(x)$, leading to the following distillation objective:

$$\ell_{\text{OPSD}} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{D}(\text{sg}[\pi_{\theta_T}(\cdot \mid x, c, z(x), o_{<t})] \, , \pi_{\theta_S}(\cdot \mid x, c, o_{<t})) \, , \tag{2}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator and $\mathcal{D}$ is a distribution-matching loss. Following OPSD (Zhao et al., 2026), the teacher parameters $\theta_T$ are preferably frozen to improve training stability.

### 3.2 Privileged Information Construction

IQA is inherently a hierarchical process, in which a global rating summarizes the holistic perceived quality, while localized regions reveal the spatially varying quality cues underlying the overall judgment (Ying et al., 2020). Following the rationalization principle of OPSD (Zhao et al., 2026), we provide the self-teacher with two complementary forms of privileged information, *i.e.*, MOS as an overall perceptual anchor and spatially grounded region crops as localized cues.

An MOS value is perceptually meaningful to both human observers and MLLMs only when interpreted within a specified range. Given an image $x$ sampled from an IQA dataset, we apply min–max normalization to rescale its original MOS $q(x)$ to the range $[0, 100]$, where larger values indicate higher perceptual quality.

Compared with arbitrary local crops, semantically meaningful regions are more likely to influence human judgments of perceptual quality (Liu et al., 2024). We therefore employ a foundational visual segmentation model as a region proposer to identify local crops that may help an MLLM better describe image quality. We instantiate it with SAM-2 (Ravi et al., 2025) for its strong zero-shot performance in rendering semantically coherent masks across diverse visual content. We represent each extracted region and its bounding box by $r_k$ and $b_k$, yielding the grounded-region set

$$\mathcal{G}(x) = \{(b_k, r_k)\}_{k=1}^{K} \, , \tag{3}$$

where $K$ is the number of extracted regions for $x$. In summary, the MOS $q(x)$ and the grounded-region set $\mathcal{G}(x)$ together constitute the privileged information $z(x) = \{q(x), \mathcal{G}(x)\}$ for $x$.

### 3.3 Quality-Sensitive Distillation

The literature suggests that vanilla full-trajectory distillation can be suboptimal because not all tokens provide equally informative supervision (Xu et al., 2026; Liu et al., 2026; Kim & Baek, 2026). For quality description generation, we seek to concentrate distillation on quality-sensitive tokens rather than quality-invariant ones, such as linguistic connectives. We therefore view the privileged information construction illustrated in Sec. 3.2 as describing *what to distill* from the self-teacher. It remains to specify *where to distill*, for which we design a counterfactual condition by perturbing the MOS $q(x)$ toward an alternative quality hypothesis within the range of $[0, 100]$:

$$q_c(x) = \begin{cases} \max(0, q(x) - \Delta), & q(x) \geq 50, \\ \min(100, q(x) + \Delta), & q(x) < 50. \end{cases} \tag{4}$$

Here, $\Delta$ is a predefined positive offset. Intuitively, if a token is sensitive to perceptual quality, changing the quality condition should induce a larger change in the teacher's next-token distribution at that position. Accordingly, the disagreement between the factual and counterfactual distributions

conditioned on $q(x)$ and $q_c(x)$, respectively, is expected to provide a useful signal for locating the positions of quality-sensitive tokens in student-generated reasoning trajectories. Specifically, we quantify the distributional discrepancy at the $t$-th token using the Jensen–Shannon (JS) divergence:

$$\begin{aligned} \bar{\pi}_t &= \frac{1}{2}\left(\bar{\pi}_F^{(t)} + \bar{\pi}_{CF}^{(t)}\right), \\ \mathrm{JS}_t &= \frac{1}{2}\left[D_{\mathrm{KL}}\left(\bar{\pi}_F^{(t)} \| \bar{\pi}_t\right) + D_{\mathrm{KL}}\left(\bar{\pi}_{CF}^{(t)} \| \bar{\pi}_t\right)\right], \end{aligned} \tag{5}$$

where $\bar{\pi}_F^{(t)} = \pi_{\theta_T}(\cdot \mid x, c, q(x), \mathcal{G}(x), o_{<t})$, $\bar{\pi}_{CF}^{(t)} = \pi_{\theta_T}(\cdot \mid x, c, q_c(x), \mathcal{G}(x), o_{<t})$, and $D_{\mathrm{KL}}$ denotes the Kullback–Leibler (KL) divergence. Let $\mathcal{T}$ denote the set of candidate positions in the generated trajectory. We select the $M = \lfloor \rho |\mathcal{T}| \rfloor$ positions with the highest JS divergence scores, where $\rho \in [0, 1]$ is a predefined selection ratio. This yields a binary mask

$$m_t = \begin{cases} 1, & t \text{ is among the top-}M \text{ positions in } \mathcal{T}, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Finally, we align the student policy with the factual teacher distribution only at the selected positions. Using reverse KL divergence, we define the quality-sensitive distillation loss, averaged over a minibatch $\mathcal{B}$, as

$$\ell_{\text{QS-OPSD}} = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \frac{1}{M} \sum_{t \in \mathcal{T}} m_t D_{\mathrm{KL}} \left(\pi_{\theta_S}(\cdot \mid x, c, o_{<t}) \,\middle\|\, \mathrm{sg}\left[\bar{\pi}_F^{(t)}\right]\right). \tag{7}$$

### 3.4 Quality Scoring

Reasoning-induced NR-IQA methods commonly formulate quality scoring as autoregressive text generation, employing either reinforcement learning or SFT on human-annotated or teacher-generated trajectories to encourage the MLLM to produce a quality score after reasoning (Cai et al., 2025; Li et al., 2025; Wu et al., 2025). Both paradigms present distinct limitations, as illustrated in Figure 1. Within the OPSD framework, we instead decouple quality reasoning from numerical score prediction. Specifically, given an input image $x$ and its student-generated quality description $o$, we extract a joint representation from the student MLLM underlying the policy $\pi_{\theta_S}$. Let $F_{\theta_S}$ denote the hidden-state mapping of the student MLLM. We take the hidden state corresponding to the last token as the representation[1]:

$$\mathbf{h} = [F_{\theta_S}(x, o)]_{\text{last}}. \tag{8}$$

Based on $\mathbf{h}$, we instantiate a scoring head $H_\phi$ as a multilayer perceptron (MLP), parameterized by $\phi$, to predict the quality score:

$$\hat{q}(x) = H_\phi(\mathbf{h}). \tag{9}$$

We optimize the model using a combination of a fidelity loss (Tsai et al., 2007) and a Pearson linear correlation coefficient (PLCC)-induced loss. Specifically, given a minibatch $\mathcal{B} = \{(x_i, q(x_i))\}_{i=1}^{|\mathcal{B}|}$, we construct a ranking label for each image pair $(x_i, x_j)$, $i \neq j$, as

$$g_{ij} = \begin{cases} 1, & q(x_i) > q(x_j), \\ 0.5, & q(x_i) = q(x_j), \\ 0, & q(x_i) < q(x_j). \end{cases} \tag{10}$$

Following Thurstone's Case V model (Thurstone, 1927), we estimate the probability that $x_i$ has higher perceptual quality than $x_j$ as

$$\hat{g}_{ij} = \Phi\left(\frac{\hat{q}(x_i) - \hat{q}(x_j)}{\sqrt{2}}\right), \tag{11}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard Gaussian distribution. The fidelity loss then measures the statistical discrepancy between the ranking label $g_{ij}$ and its predicted probability $\hat{g}_{ij}$ within the minibatch:

$$\ell_f = \frac{1}{|\mathcal{B}|(|\mathcal{B}| - 1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{|\mathcal{B}|} \left[1 - \sqrt{g_{ij}\hat{g}_{ij}} - \sqrt{(1 - g_{ij})(1 - \hat{g}_{ij})}\right]. \tag{12}$$

[1]For notational simplicity, we omit the quality-scoring prompt from the input.

Table 1: SRCC / PLCC comparison between our method and competing methods on the score regression task. All methods are trained on the KonIQ dataset.

| Method | KonIQ | SPAQ | LIVE-W | KADID | PIPAL | CSIQ | AGIQA | AVG. |
|---|---|---|---|---|---|---|---|---|
| **SRCC** | | | | | | | | |
| *Non-MLLM Methods* | | | | | | | | |
| NIMA (Talebi & Milanfar, 2018) | 0.859 | 0.856 | 0.771 | 0.535 | 0.399 | 0.649 | 0.654 | 0.675 |
| HyperIQA (Su et al., 2020) | 0.906 | 0.788 | 0.749 | 0.468 | 0.403 | 0.717 | 0.640 | 0.667 |
| DBCNN (Zhang et al., 2020) | 0.875 | 0.806 | 0.755 | 0.484 | 0.381 | 0.572 | 0.641 | 0.645 |
| MUSIQ (Ke et al., 2021) | 0.929 | 0.863 | 0.830 | 0.556 | 0.431 | 0.710 | 0.630 | 0.707 |
| MANIQA (Yang et al., 2022) | 0.834 | 0.758 | 0.832 | 0.465 | 0.452 | 0.627 | 0.636 | 0.658 |
| CLIP-IQA+ (Wang et al., 2023) | 0.895 | 0.864 | 0.805 | 0.654 | 0.419 | 0.719 | 0.685 | 0.720 |
| LIQE (Zhang et al., 2023b) | 0.928 | 0.833 | 0.870 | 0.662 | 0.501 | 0.808 | 0.708 | 0.759 |
| TOPIQ (Chen et al., 2024) | 0.927 | 0.873 | 0.811 | 0.538 | 0.430 | 0.748 | 0.669 | 0.714 |
| *MLLM-based Methods* | | | | | | | | |
| Compare2Score (Zhu et al., 2024) | 0.910 | 0.860 | 0.772 | 0.453 | 0.342 | 0.705 | 0.671 | 0.673 |
| Q-Align (Wu et al., 2024b) | 0.940 | 0.887 | 0.860 | 0.684 | 0.419 | 0.737 | 0.735 | 0.752 |
| DeQA-Score (You et al., 2025a) | 0.941 | 0.896 | 0.879 | 0.687 | 0.478 | 0.744 | 0.729 | 0.765 |
| Q-Insight (Li et al., 2025) | 0.895 | 0.903 | 0.839 | 0.702 | 0.435 | 0.640 | 0.766 | 0.740 |
| VisualQuality-R1 (Wu et al., 2025) | 0.896 | 0.892 | 0.827 | **0.712** | 0.441 | 0.707 | 0.760 | 0.748 |
| Zoom-IQA (Liang et al., 2026) | 0.922 | 0.900 | 0.870 | 0.700 | 0.465 | 0.754 | 0.765 | 0.768 |
| IQA-T1 (Wu et al., 2026) | 0.925 | **0.908** | 0.860 | 0.708 | 0.490 | **0.840** | 0.754 | 0.784 |
| ROSA-IQA | **0.946** | 0.900 | **0.885** | 0.709 | **0.504** | 0.799 | **0.768** | **0.787** |
| **PLCC** | | | | | | | | |
| *Non-MLLM Methods* | | | | | | | | |
| NIMA (Talebi & Milanfar, 2018) | 0.896 | 0.838 | 0.814 | 0.532 | 0.390 | 0.695 | 0.715 | 0.697 |
| HyperIQA (Su et al., 2020) | 0.917 | 0.791 | 0.772 | 0.506 | 0.410 | 0.752 | 0.702 | 0.693 |
| DBCNN (Zhang et al., 2020) | 0.884 | 0.812 | 0.773 | 0.497 | 0.384 | 0.586 | 0.730 | 0.667 |
| MUSIQ (Ke et al., 2021) | 0.924 | 0.868 | 0.789 | 0.575 | 0.431 | 0.771 | 0.722 | 0.726 |
| MANIQA (Yang et al., 2022) | 0.849 | 0.768 | 0.849 | 0.499 | 0.457 | 0.623 | 0.723 | 0.681 |
| CLIP-IQA+ (Wang et al., 2023) | 0.909 | 0.866 | 0.832 | 0.653 | 0.427 | 0.772 | 0.736 | 0.742 |
| LIQE (Zhang et al., 2023b) | 0.912 | 0.846 | 0.830 | 0.667 | 0.487 | 0.827 | 0.772 | 0.763 |
| TOPIQ (Chen et al., 2024) | 0.942 | 0.877 | 0.840 | 0.569 | 0.465 | 0.788 | 0.776 | 0.751 |
| *MLLM-based Methods* | | | | | | | | |
| Compare2Score (Zhu et al., 2024) | 0.923 | 0.867 | 0.786 | 0.500 | 0.354 | 0.735 | 0.777 | 0.706 |
| Q-Align (Wu et al., 2024b) | 0.941 | 0.886 | 0.853 | 0.674 | 0.403 | 0.671 | 0.772 | 0.743 |
| DeQA-Score (You et al., 2025a) | 0.953 | 0.895 | 0.892 | 0.694 | 0.472 | 0.787 | 0.809 | 0.786 |
| Q-Insight (Li et al., 2025) | 0.918 | 0.903 | 0.870 | 0.702 | 0.458 | 0.685 | 0.816 | 0.765 |
| VisualQuality-R1 (Wu et al., 2025) | 0.910 | 0.889 | 0.856 | 0.703 | 0.451 | 0.768 | 0.817 | 0.771 |
| Zoom-IQA (Liang et al., 2026) | 0.938 | 0.902 | 0.887 | 0.701 | 0.468 | 0.797 | 0.816 | 0.787 |
| IQA-T1 (Wu et al., 2026) | 0.942 | **0.905** | 0.888 | 0.686 | 0.480 | **0.850** | 0.816 | 0.795 |
| ROSA-IQA | **0.956** | 0.900 | **0.908** | **0.709** | **0.532** | 0.834 | **0.826** | **0.809** |

In addition to pairwise ranking consistency, we employ a PLCC-induced loss to encourage linear correlation between the predicted quality scores and the ground-truth MOSs:

$$\ell_p = 1 - \frac{\sum_{i=1}^{|\mathcal{B}|}(\hat{q}(x_i) - \bar{\mu})(q(x_i) - \mu)}{\sqrt{\sum_{i=1}^{|\mathcal{B}|}(\hat{q}(x_i) - \bar{\mu})^2 + \varepsilon}\sqrt{\sum_{i=1}^{|\mathcal{B}|}(q(x_i) - \mu)^2 + \varepsilon}}, \tag{13}$$

where $\bar{\mu}$ and $\mu$ denote the minibatch means of the predicted scores and ground-truth MOSs, respectively, and $\varepsilon > 0$ is a small constant introduced for numerical stability.

Finally, we derive the overall objective for jointly learning quality description and quality scoring:

$$\ell(\theta_S, \phi) = \lambda_d \ell_{\text{QS-OPSD}}(\theta_S) + \lambda_f \ell_f(\theta_S, \phi) + \lambda_p \ell_p(\theta_S, \phi), \tag{14}$$

where $\lambda_d$, $\lambda_f$, and $\lambda_p$ are the weights for the corresponding loss terms.

# 4 Experiments

We conduct comprehensive experiments across various distortion scenarios, along with ablation studies, to evaluate ROSA-IQA from both quality scoring and quality description perspectives.

## 4.1 Experimental Setups

**Implementation Details.** We perform parameter-efficient fine-tuning of Qwen3-VL-4B (Bai et al., 2025) as the backbone for ROSA-IQA, using LoRA (Hu et al., 2022) with a rank of 128 and a scaling factor of 256. The AdamW (Loshchilov & Hutter, 2019) optimizer is employed with an

| Image | ROSA-IQA | Q-Insight |
|---|---|---|
|  | The image presents a serene garden scene with a wooden bridge arching over a stream, surrounded by lush greenery and vibrant flowers. The composition is balanced, with natural textures and colors enhancing the tranquil atmosphere. **However, the image exhibits noticeable digital artifacts, particularly in the form of unnatural magenta and purple overlays that distort the foliage and bridge, creating a surreal, almost glitched effect.** These aberrations disrupt the visual harmony, making the scene feel manipulated rather than authentic. The lighting appears ... | For this image, the bridge and surrounding scenery appear to have decent sharpness and detail, providing **a clear view of textures and colors**. Exposure seems reasonable, and there are **no clear noise or artifacts**. The coloration, however, is somewhat unnatural, suggesting it might be slightly altered or filtered, possibly impacting the final rating. |
|  | **The image exhibits noticeable visual artifacts, including smeared textures and inconsistent lighting**, particularly around the building's facade and roofline. The greenery appears overly saturated and lacks natural detail, while the sky shows a flat, washed-out tone. **Structural elements, like the window and cornice, appear distorted, suggesting digital corruption or compression**... The visual perception is compromised by noise and unnatural rendering, reducing the sense of realism ... | The photo is **relatively clear and focused**, yet the color appears somewhat muted, perhaps due to the quality and light conditions. The image seems well-composed, and the architectural features and details on the building **are fairly well-defined**. The trees in the foreground also add some depth and color contrast, which is good. There is a slight lack of fine detail resolution and a hint of pixelation, yet the subject is still legible and recognizable. |
|  | The image radiates warmth and vibrancy, with a soft, golden light enhancing the subject's joyful expression. The blonde hair glows naturally, catching the ambient glow, while the textured orange coat adds depth and comfort ... Background bokeh creates a dreamy, urban atmosphere... The lighting is flattering, highlighting facial features with gentle shadows and highlights. The image's color palette is harmonious... **Overall, the visual perception is polished and emotionally engaging, capturing a candid, happy moment with excellent focus and atmospheric**… | The image appears clear and sharp, showing vibrant colors and good contrast. The subject, a person with a smiling expression, is prominently featured. **The lighting seems adequate, and the focus is slightly blurred around the subject, which creates a slightly artistic effect**. However, the lack of fine details, which may usually indicate high image quality, could suggest **an average quality rating** in that respect. The depth of field adds a pleasing softness to the image, yet the overall focus is not as crisp as one may expect… |

Figure 3: Visual examples of quality descriptions generated by ROSA-IQA and Q-Insight, with their quality-relevant judgments highlighted in blue and red, respectively.

Table 2: Comparison of quality description performance. Higher values indicate better performance.

| | | Q-Insight | | | ROSA-IQA | | |
|---|---|---|---|---|---|---|---|
| Evaluator | Criterion | KonIQ | AGIQA | CSIQ | KonIQ | AGIQA | CSIQ |
| GPT-5.6 | Completeness | 6.24 | 5.35 | 6.27 | **6.49** | **5.52** | **6.82** |
| | Accuracy | 6.68 | 6.00 | 6.47 | **6.94** | **6.51** | **6.90** |
| | Reasonableness | 6.77 | 5.99 | 6.58 | **6.92** | **5.99** | **6.98** |
| | Confidence | 6.27 | 6.14 | 5.50 | **8.62** | **8.77** | **8.57** |

initial learning rate of $10^{-4}$ and a cosine annealing decay schedule. Training runs for three epochs with a mini-batch size of 8. We retain the top $\rho = 0.15$ fraction of quality-sensitive tokens and set the counterfactual score offset to $\Delta = 50$. We set $\lambda_d = \lambda_f = \lambda_p = 1$. Additional implementation details are provided in Appendix C.

**Datasets and Metrics.** We conduct experiments on multiple IQA datasets across diverse distortion scenarios, including KonIQ (Hosu et al., 2020), SPAQ (Fang et al., 2020), LIVE Challenge (LIVE-W) (Ghadiyaram & Bovik, 2016), KADID (Lin et al., 2019), PIPAL (Gu et al., 2020), CSIQ (Larson & Chandler, 2010), and AGIQA (Li et al., 2024). The same data splits and settings as previous work (You et al., 2025a) are adopted, with ROSA-IQA trained on KonIQ and evaluated on both its test split and the remaining datasets. We use the SRCC and PLCC to evaluate the quality rating performance. We follow the MLLM-as-a-judge paradigm (Lee et al., 2024) to evaluate quality descriptions with GPT-5.6 (OpenAI, 2026). More details are provided in Appendix B.

### 4.2 Main Results

**Quality Scoring.** Table 1 reports the SRCC and PLCC results of competing NR-IQA methods and reveals three key observations. First, MLLM-based methods do not necessarily outperform their non-MLLM counterparts (*e.g.*, LIQE versus Compare2Score). Meanwhile, SFT-based methods remain competitive with reasoning-based models (*e.g.*, DeQA-Score versus Q-Insight), despite the generalization benefits often attributed to RLVR-based training (Chu et al., 2025). These two slightly counterintuitive results suggest that MLLM architectures and explicit reasoning alone do not guarantee better quality prediction. Second, by augmenting quality reasoning with explicit distortion evidence, IQA-T1 achieves highly competitive performance, as these cues may help identify distortion characteristics. Third, ROSA-IQA achieves the best average performance across all seven datasets, with a particularly clear advantage in PLCC. We attribute these gains to the effective quality

Table 3: SRCC and PLCC results of ablation studies. "Score" trains with MOS only, while "+ On/Off" adds on-/off-policy distillation. "R" denotes privileged region guidance; "Rand." and "QS" denote random and quality-sensitive token selection, respectively.

| Variants | R | Token | KonIQ | | SPAQ | | LIVE-W | | KADID | | PIPAL | | CSIQ | | AGIQA | | AVG. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC |
| Score | – | – | **0.952** | **0.961** | 0.895 | 0.898 | 0.892 | 0.912 | 0.691 | 0.691 | 0.502 | 0.516 | 0.682 | 0.731 | 0.744 | 0.804 | 0.765 | 0.787 |
| + Off | ✓ | All | 0.951 | **0.961** | 0.900 | 0.902 | **0.895** | **0.912** | 0.684 | 0.684 | **0.514** | 0.535 | 0.735 | 0.771 | 0.758 | 0.816 | 0.777 | 0.797 |
| + On | – | All | 0.947 | 0.957 | 0.901 | 0.902 | 0.892 | 0.911 | 0.693 | 0.694 | 0.503 | 0.532 | 0.766 | 0.796 | 0.762 | 0.822 | 0.781 | 0.802 |
| | ✓ | All | 0.948 | 0.958 | **0.902** | **0.904** | 0.890 | 0.911 | 0.700 | 0.702 | 0.507 | **0.539** | 0.773 | 0.809 | 0.764 | 0.819 | 0.783 | 0.806 |
| | ✓ | Rand. | 0.947 | 0.958 | 0.901 | 0.902 | 0.888 | 0.910 | 0.706 | 0.705 | 0.510 | 0.534 | 0.758 | 0.796 | 0.764 | 0.823 | 0.782 | 0.804 |
| | ✓ | QS | 0.946 | 0.956 | 0.900 | 0.900 | 0.885 | 0.908 | **0.709** | **0.709** | 0.504 | 0.532 | **0.799** | **0.834** | **0.768** | **0.826** | **0.787** | **0.809** |



The **image captures** a **striking portrait** with **vivid visual** contrast. Bright yellow sunflowers frame the subject**'s** face, their petals sharply defined against her bold, electric-blue hair. Her lips, painted a vibrant magenta, stand out against the neutral tones of her skin and the white background. Natural sunlight creates **strong,** directional shadows that sculpt her features, adding depth **and** drama. The composition balances organic elements-flowers, foliage, and hair-with a **clean**, architectural backdrop. **The shallow** depth of field keeps the focus on her expressive gaze, subtly hinting at emotion. **Color** saturation is **rich but** not overwhelming, **lending** the image a **dynamic**, almost surreal quality. Overall, the visual **perceptual** quality is **high-engaging, balanced,** and evocative.

Figure 4: Visualization of the quality-sensitive token selection. Boldface indicates selection.

representations learned through OPSD and the scoring head's ability to directly produce continuous quality estimates.

**Quality Description.** To evaluate ROSA-IQA's quality description capability, we compare it with Q-Insight in Table 2 and Figure 3. ROSA-IQA matches or outperforms Q-Insight across all criteria. We also conduct a sample-efficient and de-biased subjective preference study (Feng et al., 2025) to corroborate these findings (see Appendix E for more details). These results suggest that the student effectively learns to generate meaningful quality descriptions from the self-teacher, which benefits from a more comprehensive context afforded by privileged information.

### 4.3 Ablation Studies

Table 3 compares model variants that differ in their learning signals, use of privileged information, and token selection strategies for distillation.

**Effect of Self-Distillation.** The "Score" baseline achieves strong in-domain performance on KonIQ but exhibits inferior generalization in cross-dataset evaluations, suggesting a tendency toward overfitting. Following Cai et al. (2025), off-policy distillation of quality reasoning before scoring, from a teacher with all privileged information, yields gains across several datasets, suggesting that training-time quality reasoning improves generalization. These gains are more pronounced with on-policy self-distillation, attributable to reduced training-inference exposure bias.

**Effect of Semantic Regions.** Augmenting the privileged MOS with segmentation-derived semantic regions yields marginal performance gains, suggesting that regional cues complement the global quality anchor by providing spatially grounded information. This effect is more evident in quality descriptions, as illustrated in Appendix G.

**Effect of Quality-Sensitive Distillation.** We compare quality-sensitive distillation with full-token distillation and a baseline that randomly selects the same number of tokens. Its higher average performance indicates that the gains arise from perceptually meaningful token selection rather than distillation sparsity. As shown in Figure 4, the proposed factual–counterfactual disagreement mechanism primarily selects tokens associated with quality-relevant descriptions.

### 4.4 Further Analysis

**Suitability of OPSD for IQA.** Recent studies suggest that OPSD may reduce response diversity and suppress exploration in open-ended tasks (Kaur et al., 2026; Wang et al., 2026). However,

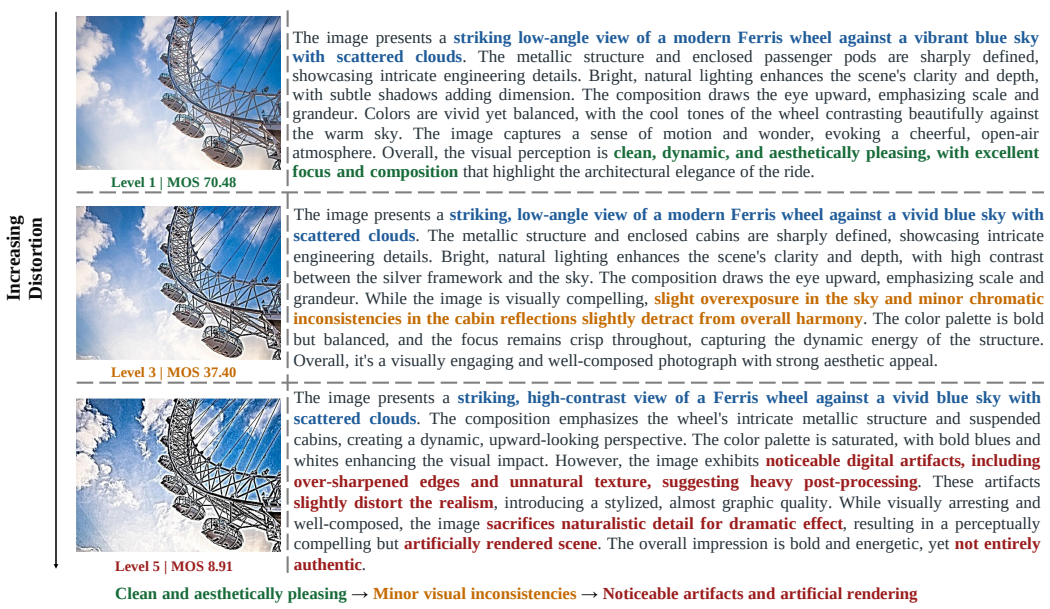



Figure 5: Quality descriptions produced by ROSA-IQA under increasing synthetic distortion.

Table 4: Comparison of inference with and without quality reasoning.

| Setting | KonIQ | | SPAQ | | LIVE-W | | KADID | | PIPAL | | CSIQ | | AGIQA | | AVG. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC |
| Direct | 0.942 | 0.953 | **0.904** | **0.905** | 0.880 | 0.904 | **0.731** | **0.730** | **0.506** | **0.537** | **0.802** | **0.837** | **0.775** | **0.834** | **0.792** | **0.814** |
| Reasoning | **0.946** | **0.956** | 0.900 | 0.900 | **0.885** | **0.908** | 0.709 | 0.709 | 0.504 | 0.532 | 0.799 | 0.834 | 0.768 | 0.826 | 0.787 | 0.809 |

IQA favors consistent perceptual judgments aligned with human quality ratings (Hosu et al., 2020). Figure 5 shows that as the same reference image undergoes more severe distortions, ROSA-IQA maintains consistent scene understanding while strengthening its criticism of quality degradation.

**Effect of Quality Reasoning.** Although quality reasoning benefits training, it remains unclear whether it must precede quality scoring at inference. Table 4 shows that direct quality scoring slightly outperforms the reasoning-then-scoring pipeline. This suggests that explicit quality description can be decoupled from quality scoring at inference, while serving as structured supervision that encourages perceptually grounded quality representations during training.

## 5 Conclusion

We have introduced ROSA-IQA based on OPSD, where the teacher and student are instantiated from the same MLLM. The self-teacher leverages MOS and segmentation-derived local regions as privileged context to provide more informative targets for student training. We further employ factual–counterfactual disagreement as a distillation-worthiness mechanism to identify quality-sensitive tokens, and a scoring head to produce continuous quality scores. ROSA-IQA achieves favorable performance across diverse distortion scenarios, particularly in cross-dataset evaluation, demonstrating its effectiveness in learning generalizable representations. We also show quantitatively and qualitatively that ROSA-IQA generates human-aligned quality descriptions.

Future work will pursue a more principled understanding of the reasoning–scoring relationship and extend ROSA-IQA to fine-grained scoring and descriptions across multiple perceptual attributes. We also envision a dual-feedback mechanism that combines quality reasoning and scoring to support real-world applications such as perceptual optimization (Zhang et al., 2024), computational photography (You et al., 2026), and image generation (Jaiswal et al., 2026).

## AI USE STATEMENT

This work studies generative AI models as the object of research and investigates their image quality assessment capabilities. In the experiments, we further used generative AI tools as automatic judges to evaluate the generated quality reasoning. We also used generative AI tools to provide grammar suggestions for the manuscript. All remaining aspects of this work, including the research idea, method, experimental design, implementation, data analysis, and figure generation, were designed and carried out by the authors. We have reviewed all AI-assisted outputs and take full responsibility for the final content of this work.

## REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Zhuoxuan Cai, Jian Zhang, Xinbin Yuan, Peng-Tao Jiang, Wenxiang Chen, Bowen Tang, Lujian Yao, Qiyuan Wang, Jinwen Chen, and Bo Li. Q-Ponder: A unified training pipeline for reasoning-based visual quality assessment. *arXiv preprint arXiv:2506.05384*, 2025.

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

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *International Conference on Machine Learning*, pp. 10818–10838, 2025.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3677–3686, 2020.

Kehua Feng, Keyan Ding, Hongzhi Tan, Kede Ma, Zhihua Wang, Shuangquan Guo, Yuzhou Cheng, Ge Sun, Guozhou Zheng, Qiang Zhang, and Huajun Chen. Sample-efficient human evaluation of large language models via maximum discrepancy competition. In *Annual Meeting of the Association for Computational Linguistics*, pp. 10913–10947, 2025.

Deepti Ghadiyaram and Alan C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2016.

Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy S. Ren, and Chao Dong. PIPAL: A large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, pp. 633–651, 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.

Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Shantanu Jaiswal, Mihir Prabhudesai, Nikash Bhardwaj, Zheyang Qin, Amir Zadeh, Chuan Li, Katerina Fragkiadaki, and Deepak Pathak. Iterative refinement improves compositional image generation. *arXiv preprint arXiv:2601.15286*, 2026.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. *arXiv preprint arXiv:2607.05184*, 2026.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021.

Minsang Kim and Seung Jun Baek. Explain in your own words: Improving reasoning via token-selective dual knowledge distillation. In *International Conference on Learning Representations*, 2026.

Eric C. Larson and Damon M. Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. Prometheus-Vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11286–11315, 2024.

Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. AGIQA-3K: An open database for AI-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):6833–6846, 2024.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-Insight: Understanding image quality via visual reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 36802–36827, 2025.

Guoqiang Liang, Jianyi Wang, Zhonghua Wu, Shangchen Zhou, and Chen Change Loy. Zoom-IQA: Image quality assessment with reliable region-aware reasoning. In *European Conference on Computer Vision*, 2026.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023.

Kai Liu, Ziqing Zhang, Wenbo Li, Renjing Pei, Fenglong Song, Xiaohong Liu, Linghe Kong, and Yulun Zhang. Dog-IQA: Standard-guided zero-shot MLLM for mix-grained image quality assessment. *arXiv preprint arXiv:2410.02505*, 2024.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning. *arXiv preprint arXiv:2605.21606*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022.

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.

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

Anush Krishna Moorthy and Alan C. Bovik. Visual importance pooling for image quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):193–201, 2009.

OpenAI. GPT-5.6. `https://openai.com/index/gpt-5-6/`, 2026.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

Ernestasia Siahaan, Alan Hanjalic, and Judith A. Redi. Semantic-aware blind image quality assessment. *Signal Processing: Image Communication*, 60:237–252, 2018.

Alex Stein, Furong Huang, and Tom Goldstein. GATES: Self-distillation under privileged context with consensus gating. *arXiv preprint arXiv:2602.20574*, 2026.

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 *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.

Louis L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927.

Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. FRank: A ranking method with fidelity loss. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 383–390, 2007.

Hao Wang, Guozhi Wang, Han Xiao, Yufeng Zhou, Yue Pan, Jichao Wang, Ke Xu, Yafei Wen, Xiaohu Ruan, Xiaoxin Chen, and Honggang Qi. Skill-SD: Skill-conditioned self-distillation for multi-turn LLM agents. *arXiv preprint arXiv:2604.10674*, 2026.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023.

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, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, Qiong Yan, and Weisi Lin. 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, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029, 2024b.

Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, Xiaohong Liu, Guangtao Zhai, Shiqi Wang, and Weisi Lin. Towards open-ended visual quality comparison. In *European Conference on Computer Vision*, pp. 360–377, 2024c.

Jinjian Wu, Jiaqi Tang, Wei Wei, Yingying Yan, Jianmin Chen, Botong Geng, Lei Zhang, and Qifeng Chen. IQA-T1: Tool-based visual evidence reasoning for image quality assessment. In *European Conference on Computer Vision*, 2026.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. VisualQuality-R1: Reasoning-induced image quality assessment via reinforcement learning to rank. In *Advances in Neural Information Processing Systems*, volume 38, pp. 88167–88190, 2025.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1191–1200, 2022.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan C. Bovik. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3572–3582, 2020.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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*, 34:8201–8215, 2025b.

Zhiyuan You, Ke Wang, He Zhang, Xin Cai, Jinjin Gu, Tianfan Xue, Chao Dong, and Zhoutong Zhang. PhotoFramer: Multi-modal image composition instruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10197–10207, 2026.

Shuyan Zhai, Jiaqi He, Weixia Zhang, Liang Wang, Zhenjie Lee, Zufeng Zhang, and Kede Ma. PrISM-IQA: Image quality assessment made practical for smartphone photography. *arXiv preprint arXiv:2606.31626*, 2026.

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, 2018.

Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2020.

Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021.

Weixia Zhang, Dingquan Li, Chao Ma, Guangtao Zhai, Xiaokang Yang, and Kede Ma. Continual learning for blind image quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2864–2878, 2023a.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14071–14081, 2023b.

Weixia Zhang, Dingquan Li, Guangtao Zhai, Xiaokang Yang, and Kede Ma. When no-reference image quality models meet MAP estimation in diffusion latents. *arXiv preprint arXiv:2403.06406*, 2024.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang. Adaptive image quality assessment via teaching large multimodal model to compare. In *Advances in Neural Information Processing Systems*, volume 37, pp. 32611–32629, 2024.

# APPENDIX

## A DESIGNED PROMPTS

The prompts for quality description and quality scoring are listed in Table 5. The student and teacher use the same task instruction and output constraints, while the teacher additionally receives the ground-truth MOS and semantic-region information. The factual and counterfactual conditions differ only in the supplied MOS. Semantic regions provide local references for reasoning about subtle degradations. For score prediction, the generated description and original image are jointly encoded, and the final hidden state is mapped by the scoring head to a continuous quality score. The description prompt prohibits numeric scores. To reduce computation, we limit responses to 160 words, which has little effect on the retained information.

Table 5: Prompts used by ROSA-IQA. Bracketed content denotes a sample-dependent field. The semantic-region line is repeated for each selected crop.

| Prompt | Content |
|---|---|
| Student prompt | Task: Assess the visual perceptual quality of the image 1.<br><br>Inputs:<br>- Image 1: the original image to assess.<br><br>Rules:<br>- Do not output a numeric quality score or ranking.<br>- Keep the description within 160 words. |
| Privileged teacher prompt | Task: Assess the visual perceptual quality of the image 1.<br><br>Inputs:<br>- Image 1: the original image to assess.<br>- Image $k$: a semantic-region crop from Image 1; bbox [left=$l_k\%$, top=$t_k\%$, right=$r_k\%$, bottom=$b_k\%$] in Image 1.<br><br>Privileged information:<br>- The true human quality score (MOS) is $s$ on a 0–100 scale (higher = better).<br><br>Rules:<br>- Do not output a numeric quality score or ranking.<br>- Keep the description within 160 words.<br>- Use each semantic-region crop and its bbox only as a local reference; assess only Image 1. |
| Quality scoring prompt | Based on the following quality description:<br>[generated quality description]<br><br>Assign an overall visual perceptual quality score for this image. |

## B QUALITY DESCRIPTION EVALUATION CRITERIA

As summarized in Table 6, we use MLLM-as-a-judge with four complementary criteria to evaluate the quality description. Specifically, completeness measures the coverage of perceptually relevant attributes, accuracy evaluates consistency with visual evidence, reasonableness assesses the coherence between observations and the overall quality judgment, and confidence characterizes the decisiveness of expression. Together, these criteria capture both the perceptual validity and the explanatory quality of the generated reasoning. During evaluation, the correspondence between each generated description and the input image serves as the primary basis for judgment, while the corresponding reference MOS provides supplementary guidance on overall quality. To support consistent comparison across models, the evaluator assigns an integer score from 1 to 9 for each criterion.

Table 6: Criteria used to evaluate image-quality descriptions.

| Criterion | Definition |
|---|---|
| Completeness | Whether the description covers the most important perceptual qualities and defects a human would notice. |
| Accuracy | Whether the description is faithful to the visual evidence. Human visual perception takes priority over blindly matching the reference MOS. |
| Reasonableness | Whether the reasoning logically connects the visual evidence to a holistically appropriate quality assessment. |
| Confidence | Whether the description uses clear and decisive language rather than uncertain or hedged language, regardless of whether it is correct. |

Table 7: Effect of the quality-sensitive token-retention ratio $\rho$. All distilled variants use reverse KL divergence.

| $\rho$ | KonIQ | | SPAQ | | LIVE-W | | KADID | | PIPAL | | CSIQ | | AGIQA | | AVG. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC |
| 0 | **0.952** | **0.961** | 0.895 | 0.898 | **0.892** | **0.912** | 0.691 | 0.691 | 0.502 | 0.516 | 0.682 | 0.731 | 0.744 | 0.804 | 0.765 | 0.787 |
| 0.15 | 0.946 | 0.956 | 0.900 | 0.900 | 0.885 | 0.908 | **0.709** | **0.709** | 0.504 | 0.532 | **0.799** | 0.834 | **0.768** | **0.826** | **0.787** | **0.809** |
| 0.30 | 0.946 | 0.956 | 0.899 | 0.899 | 0.885 | 0.908 | 0.696 | 0.696 | 0.507 | 0.538 | 0.798 | **0.834** | 0.766 | 0.824 | 0.785 | 0.808 |
| 0.45 | 0.948 | 0.957 | 0.900 | 0.900 | 0.888 | 0.911 | 0.696 | 0.697 | 0.513 | 0.543 | 0.779 | 0.814 | 0.767 | 0.824 | 0.784 | 0.806 |
| 0.60 | 0.946 | 0.957 | 0.902 | 0.902 | 0.887 | 0.908 | 0.701 | 0.701 | 0.507 | 0.536 | 0.787 | 0.820 | 0.764 | 0.824 | 0.785 | 0.807 |
| 0.75 | 0.946 | 0.957 | 0.900 | 0.900 | 0.889 | 0.911 | 0.702 | 0.701 | **0.514** | **0.543** | 0.773 | 0.807 | 0.766 | 0.822 | 0.784 | 0.806 |
| 0.90 | 0.948 | 0.958 | 0.899 | 0.899 | 0.887 | 0.909 | 0.699 | 0.700 | 0.507 | 0.538 | 0.777 | 0.812 | 0.766 | 0.821 | 0.783 | 0.805 |
| 1 | 0.948 | 0.958 | **0.902** | **0.904** | 0.890 | 0.911 | 0.700 | 0.702 | 0.507 | 0.539 | 0.773 | 0.809 | 0.764 | 0.819 | 0.783 | 0.806 |

Table 8: Comparison of distillation loss functions.

| Distillation loss | KonIQ | | SPAQ | | LIVE-W | | KADID | | PIPAL | | CSIQ | | AGIQA | | AVG. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC |
| Reverse KL | 0.946 | **0.956** | 0.900 | 0.900 | **0.885** | **0.908** | **0.709** | **0.709** | 0.504 | 0.532 | **0.799** | **0.834** | **0.768** | **0.826** | **0.787** | **0.809** |
| Forward KL | 0.946 | 0.955 | **0.902** | **0.902** | 0.884 | 0.908 | 0.696 | 0.697 | **0.509** | **0.539** | 0.785 | 0.820 | 0.767 | 0.824 | 0.784 | 0.806 |
| Jeffreys | **0.947** | 0.956 | 0.900 | 0.901 | 0.884 | 0.907 | 0.699 | 0.700 | 0.509 | 0.539 | 0.792 | 0.826 | 0.765 | 0.823 | 0.785 | 0.807 |
| Hard-label CE | 0.942 | 0.953 | 0.897 | 0.899 | 0.879 | 0.902 | 0.702 | 0.703 | 0.490 | 0.523 | 0.797 | 0.830 | 0.764 | 0.825 | 0.781 | 0.805 |

## C Additional Implementation Details

**SAM-2 Region Selection.** SAM-2 produces candidate masks with a predicted-IoU threshold of $0.88$. We discard empty masks and those covering less than $4\%$ or more than $16\%$ of the image, rank the rest by predicted IoU times the square root of the area ratio, and apply bounding-box NMS at an IoU of $0.3$, keeping at most ten crops per image. During training, we randomly sample at most three of these crops, together with their bounding boxes, for the self-teacher.

**PLCC Calculation.** Following the common practice[2], we map model predictions to MOSs using a monotonic four-parameter logistic function to compensate for prediction nonlinearity before computing PLCC.

## D Additional Ablation Analyses

**Effect of the Token-Retention Ratio.** Table 7 studies the proportion $\rho$ of tokens retained by quality-sensitive selection, including the two endpoints: $\rho = 0$ completely disables token-level distillation, whereas $\rho = 1$ distills all tokens. The default $\rho = 0.15$ yields the best average SRCC/PLCC of 0.787/0.809, with especially strong cross-dataset performance, particularly on KADID, CSIQ, and AGIQA.

**Effect of the Distillation Loss.** Table 8 compares different choices of distillation loss. Reverse KL performs best, achieving an average SRCC/PLCC of 0.787/0.809, while Forward KL and Jeffreys divergence remain competitive, and hard-label cross-entropy shows a larger drop. This suggests that preserving the teacher's soft distribution provides richer supervision than one-hot targets. Reverse

[2]VQEG, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment," 2000.

KL is particularly well suited to our on-policy setting, as it emphasizes student-preferred tokens that receive low probability from the privileged teacher, thereby discouraging potentially unreliable quality-related predictions along student-generated trajectories.

## E Subjective Human Preference Study

Compared with MLLM-as-a-judge evaluation, a well-controlled subjective study provides a more direct and faithful measure of human preference for model-generated quality descriptions. However, exhaustively evaluating all images is highly labor-intensive and time-consuming, and is likely unnecessary for obtaining reliable model-level conclusions. Following Feng et al. (2025), we sample description pairs produced by two models that differ most substantially from each other. To quantify textual discrepancy, we use Qwen3-VL-Embedding-2B (Li et al., 2026) to compute the semantic distance between descriptions generated by the two models, and select the 12 samples with the largest distances from each of KonIQ, CSIQ, and AGIQA. In each trial, participants are shown an image, its MOS, and two anonymous candidate descriptions in randomized order, and are asked to select the description that better matches the image and its perceptual quality. The study involves 12 participants with normal or corrected-to-normal vision and general familiarity with image processing and computer vision, none of whom are informed of the study background or model identities. Figure 1 applies this protocol to compare ROSA-IQA with and without privileged information, while Figure 6 compares ROSA-IQA with Q-Insight. In both cases, the results consistently favor ROSA-IQA, indicating that it produces quality descriptions that are more aligned with human preference across all three datasets.

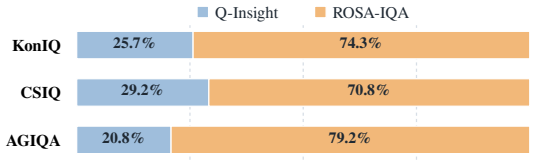



Figure 6: Human preference between ROSA-IQA and Q-Insight.

## F Limitations and Future Directions

**Cross-Dataset Quality Calibration.** Our framework aligns quality descriptions with MOS supervision under single-dataset training, where samples share a common annotation protocol and score distribution. This assumption may break under joint training across datasets, since identical normalized MOS values can still correspond to different perceptual quality levels (Zhang et al., 2021). Directly using MOS as privileged information may therefore introduce inconsistent quality anchors across datasets. In future work, we plan to explore constructing pairwise samples within each dataset as privileged information, providing relative quality supervision without requiring globally calibrated MOS across datasets.

**Agentic IQA.** Recent agent-based NR-IQA methods (Liang et al., 2026; Wu et al., 2026) have shown promise for IQA by actively acquiring additional visual evidence. In contrast, ROSA-IQA performs quality assessment directly from the given visual input without invoking external tools. A natural extension is to incorporate tool use into the privileged-teacher framework. The teacher could either supervise the student in learning when, where, and which tools to invoke, enabling adaptive evidence acquisition, or privately leverage tools to obtain richer quality evidence and distill it into a tool-free student. The latter is particularly appealing, as it may transfer the benefits of tool-assisted perception without requiring additional tool calls at inference.

## G Additional Qualitative Examples

Figures 7 and 8 show examples of how the reasoning changes as additional design components are incorporated for low- and high-quality images, respectively. The examples show a clear progression from global calibration to local grounding and quality-focused refinement. MOS improves the calibration of overall quality and degradation severity, while SAM-2 regions ground these judgments in specific semantic regions. Quality-sensitive distillation further emphasizes perceptually diagnostic evidence and its consequences, yielding more calibrated, localized, and quality-focused descriptions.

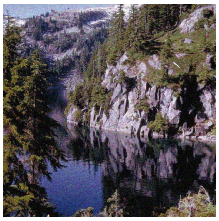
| | | |
|---|---|---|
| **Privileged Information** |  |  |
| **Pretrained** | To evaluate the quality of this picture, I'll consider several factors: 1. Resolution and Clarity: The image is **clear enough to make out the details** of the building and the people. Some **minor blurring or graininess** is visible, but it **doesn't significantly impact** the overall picture. 2. Lighting: The scene is **well-lit**, with warm tones that contribute to the visibility of the subject. The lighting does, however, cause some shadows and reflections that could obscure fine details, although these **aren't too pronounced**. 3. Color Contrast: The colors are warm, and while the contrast is sufficient for a readable image, there's some noise which might slightly detract from the image's detail. 4. Focus: The focus is mostly on the barriers in the foreground, while the background building and people are slightly blurred. 5. Noise and Graininess: There is a **minimal amount of noise** visible in the image, but it's **not detrimental** to the overall experience. In the end, it seems to be a **clear photo with reasonable lighting and colors** that give it a **good level of quality**, but the minor grain and lack of fine detail are noticeable. | This image appears to have a **medium to high level of detail**. The scene is **not overly blurry** and the objects, such as the trees and rocks, are distinguishable. The overall tone of the image seems to be **natural and balanced**, offering a good representation of the subject matter. There is a bit of noise visible, which could **slightly affect the clarity**, yet **it's not prominent**. The colors are **vivid and well-defined**, especially in the green foliage, which enhances the appeal. The slight noise, along with the overall quality, allows for a **rating in the mid to upper range**. Upon evaluating the image, the noise level and lack of fine detail appear to hold it back from the highest scores. Thus, while the image **looks good overall** and captures the essence well, it falls short in a few areas. |
| **MOS** | The image captures a vibrant nighttime scene with **motion blur**, suggesting movement among **pedestrians** near a **historic landmark**. Warm, golden lighting illuminates the architecture and flags, creating a rich, atmospheric glow. The **foreground barriers** introduce a sense of depth and context, while the **blurred figures** convey energy and liveliness. However, the **motion blur** slightly **reduces sharpness**, and the overall image appears slightly **overexposed in highlights**, with some **areas losing detail**. The composition balances architectural grandeur with human activity, evoking a dynamic urban ambiance. Colors are saturated yet harmonious, with the yellow tones dominating the palette. Despite minor technical imperfections, the image successfully conveys mood and movement, offering a compelling visual narrative of a bustling evening in a historic setting. | The image presents a rugged, natural landscape with **steep, rocky cliffs** and **dense evergreen trees** framing a **calm, reflective lake**. The water mirrors the surrounding terrain, enhancing depth and tranquility. However, the visual quality is compromised by **noticeable digital noise and grain**, particularly in the **darker areas**, which **reduce clarity and texture detail**. Colors appear somewhat muted and slightly oversaturated in highlights, creating an **inconsistent tonal balance**. The overall impression is one of a high-contrast, moody scene, yet the image's technical execution detracts from its natural beauty. While the composition is compelling, the **visual artifacts hinder immersion**, making the scene feel less refined and more digitally processed than its serene subject matter suggests. |
| **MOS + SAM-2 Regions** | The image captures a vibrant nighttime scene with warm, golden lighting illuminating the **historic architecture** and **bustling crowd**. **Motion blur** suggests activity and movement, particularly among **pedestrians**, adding a dynamic feel. The **foreground barriers with red-and-white stripes** create a **strong visual anchor**, contrasting with the softly lit background. The composition balances architectural grandeur with human presence, evoking a lively urban atmosphere. Lighting is rich but slightly uneven, with some areas appearing **overexposed under the glow**. The overall impression is immersive and energetic, though slight focus inconsistencies and **motion blur reduce sharpness**. The image successfully conveys atmosphere and motion, making it engaging despite minor technical imperfections. | The image presents a serene natural landscape with a calm lake reflecting rugged, tree-clad cliffs. The composition draws the eye along the **water's edge**, where **evergreen trees frame the scene**. However, the visual quality is compromised by noticeable **digital artifacts—pixelation and graininess**—especially in the **foliage and rock textures**. Colors appear muted and slightly distorted, with an **unnatural blue tint** dominating the **water and sky**. The overall impression is one of a **digitally degraded photograph**, **lacking the crispness and richness** expected in high-quality imagery. Despite its tranquil subject matter, the technical flaws detract from the visual experience, making the scene feel less immersive and more processed than natural. |
| **Full** | The image captures a vibrant nighttime scene with **motion blur**, suggesting activity and movement among **pedestrians**. The warm, golden lighting illuminates the historic architecture, particularly the **prominent castle-like structure**, while the foreground features a **red-and-white striped barrier**. The composition balances the grandeur of the background with the dynamic human presence. However, the **motion blur softens details, reducing sharpness and clarity**. The lighting creates strong contrasts, with **deep shadows** and **bright highlights**, contributing to a moody, atmospheric feel. The overall impression is energetic yet slightly indistinct, with the blur lending a sense of motion but **sacrificing fine detail**. The image conveys a lively urban ambiance, though its **visual clarity is compromised** by the motion effect. | The image presents a rugged, natural landscape with **steep, rocky cliffs and dense evergreen trees** bordering a **calm, reflective lake**. The water mirrors the **surrounding cliffs and foliage**, creating a serene, symmetrical effect. However, the image suffers from **significant digital noise and pixelation**, particularly in the **foreground and midground**, which **obscures fine details and reduces clarity**. Colors appear muted and slightly distorted, with an **unnatural blue tint dominating the water and sky**. The **overall texture is grainy**, diminishing visual smoothness. Despite these flaws, the composition retains a **strong sense of depth and natural grandeur**, with the **interplay of rock, tree, and water** offering visual interest. The scene's inherent beauty is partially masked by technical imperfections. |

Figure 7: Qualitative comparison of quality descriptions generated by model variants. The full model is trained with all privileged information and the quality-sensitive distillation. Colored text marks the quality-relevant descriptions emphasized by each variant. The complete model better identifies the dominant degradation and leverages localized semantic regions to produce more spatially specific quality descriptions.

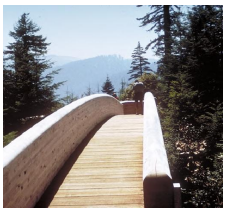
| | | |
|---|---|---|
| **Privileged Information** |  |  |
| **Pretrained** | The image seems to convey a scenic atmosphere but **lacks fine detail** and **fine resolution** that is usually sought in photographs. The foreground has a darker tone, which may hide some detail, but the sunset and the horizon are relatively well-defined. Lighting and clarity are somewhat subdued, and it appears that the **colors are saturated, possibly due to a filter or camera error**, which affects color accuracy. The composition is intriguing, giving the viewer a sense of the scene but the **lack of fine detail** and the slightly flat nature of the image reduce its impact. It appears clear enough in composition but **not vibrant in detail**. | In assessing the quality of the picture, there are a few factors to consider: 1. Clarity and Detail: The image is relatively clear and the bridge, trees, and background are distinguishable. 2. Color and Contrast: The colors are somewhat muted, which could indicate age or **low image quality**. The contrast between the wooden bridge and the green trees is not very pronounced. 3. Sharpness and Resolution: The image appears **slightly blurry** and **lacks fine detail**, suggesting **lower image quality**. 4. Composition: The picture is well-composed, which partially compensates for issues in the image's quality. Overall, the photo conveys its subject but **could stand to be sharper** and have more vibrant colors. The issue of resolution could contribute to the overall feeling of slightly **lower image quality**. |
| **MOS** | The image captures a **serene sunset scene** with **silhouetted figures** standing by the water's edge, framed by dark foliage in the foreground. The **gradient sky transitions** from warm orange near the sun to deep purple, creating a tranquil, atmospheric mood. The **sun's reflection glows** along the water's surface, adding a striking vertical line of light. The silhouettes, including a **child held between two adults**, convey intimacy and stillness. While the image is emotionally evocative, the overall clarity is moderate, with slight softness and muted contrast. The composition is balanced, but the foreground leaves slightly obscure the view, reducing depth perception. Colors are rich yet slightly desaturated, lending a dreamy, nostalgic quality. The image successfully evokes calm and connection, though its technical sharpness and detail could be enhanced. | The image presents a serene forest scene with a **wooden bridge curving gently** through tall pines, leading toward **distant, misty mountains**. The lighting is bright and natural, casting soft shadows and highlighting the **texture of the wood and bark**. Colors are vivid yet slightly faded, suggesting mild age or film grain. The composition draws the eye along the bridge's curve, creating depth and inviting exploration. A **lone figure** stands near the bridge's end, adding scale and human presence. While the image is clear and well-composed, subtle blurring in the background and **slight overexposure in the sky** reduce overall sharpness. The atmosphere feels tranquil and slightly nostalgic, capturing a moment of quiet wilderness beauty with a gentle, timeless quality. |
| **MOS + SAM-2 Regions** | The image captures a serene sunset over water, with **silhouetted figures and foliage** framing the scene. The gradient sky transitions from **warm orange to deep purple**, creating a tranquil, atmospheric mood. The **sun's reflection glows along the horizon**, adding depth and warmth. Silhouettes of **two adults and a child** stand together, evoking a sense of connection and peace. The **foreground foliage** adds texture and natural framing, enhancing composition. While the image is visually harmonious, the overall clarity and color saturation are moderate, with slight softness in the distant horizon. The lighting is evocative but not exceptionally sharp, resulting in a pleasing, contemplative aesthetic that prioritizes mood over technical precision. | The image presents a serene forest scene with a **wooden bridge curving through tall pines**, leading toward distant mountains. The lighting is bright and natural, casting soft shadows and highlighting the **textures of wood and bark**. Colors are vivid yet balanced, with **deep greens and earthy browns** contrasting against the **pale sky**. The composition draws the eye along the bridge's curve, creating a sense of depth and tranquility. While the image is clear, slight overexposure in the sky and minor graininess suggest it may be a **scanned or older photograph**. Overall, the visual experience is peaceful and immersive, capturing the quiet majesty of a woodland trail. The natural elements are rendered with pleasing detail, evoking a calm, contemplative mood. |
| **Full** | The image captures a serene sunset scene with **silhouetted figures against a vibrant, gradient sky** transitioning from golden yellow to deep purple. The **sun's reflection glows warmly on the water**, creating a **striking path of light**. **Silhouetted foliage in the foreground** adds depth and framing, enhancing the tranquil mood. The composition balances natural elements with human presence, evoking a sense of peace and contemplation. While the image is visually harmonious, the overall clarity and color saturation are moderate, with slight softness in the horizon and **subtle noise visible in the darker areas**. The emotional resonance is strong, though technical execution could be refined for greater sharpness and vibrancy. The scene remains evocative and aesthetically pleasing despite minor imperfections. | The image presents a scenic, **sunlit wooden bridge arching** through a forested landscape. The composition draws the eye along the bridge's curve toward **distant, misty mountains**. Natural lighting creates a warm, inviting atmosphere, with dappled shadows from trees enhancing depth. The **textures of the wood and stone** are visible, adding tactile realism. **A lone figure on the bridge** provides scale and a sense of peaceful solitude. The color palette is rich with earthy tones, accented by the vibrant green of pines and the soft blue of the sky. While the image is clear and well-composed, **slight softness in the background** suggests a slightly older photographic style, lending a nostalgic quality to the scene. Overall, it evokes tranquility and natural beauty. |

Figure 8: Qualitative comparison of quality descriptions generated by model variants. The full model is trained with all privileged information and the quality-sensitive distillation. Colored text marks the quality-relevant descriptions emphasized by each variant. The complete model better distinguishes silhouettes, atmospheric perspective, and depth-dependent softness from genuine global quality degradation.