Q-Probe: Scaling Image Quality Assessment to High Resolution via Context-Aware Agentic Probing

Xiang Li[1,2*] **Xueheng Li**[1,2*†] **Yu Wang**[1,2*] **Xuanhua He**[3‡]
Zhangchi Hu[1] **Chengjun Xie**[2‡]
[1]University of Science and Technology of China
[2]Hefei Institutes of Physical Science, Chinese Academy of Sciences
[3]The Hong Kong University of Science and Technology



Figure 1: Challenges in detecting subtle distortions via global perception versus local zooming. (a) Existing MLLMs fail to capture subtle local artifacts. (b) Even when visible via cropping, Semantic Robustness Bias causes models to ignore defects in key semantic areas (e.g., face). (c-d) Naive zooming leads to *Logic Collapse*, where the model misinterprets natural bokeh as blur (c) or falsely learns that zooming implies low quality (d). (e) Q-Probe mimics human active viewing (Global Perception → Local Scrutiny → Critical Thinking) to correctly distinguish artifacts from natural effects. (f-g) Data distribution of the high-resolution Vista-Bench and performance comparison showing Q-Probe's superiority.

Abstract

Reinforcement Learning (RL) has empowered Multimodal Large Language Models (MLLMs) to achieve superior human preference alignment in Image Quality Assessment (IQA). However, existing RL-based IQA models typically rely on coarse-grained global views, failing to capture subtle local degradations in high-resolution scenarios. While emerging "Thinking with Images" paradigms enable multi-scale visual perception via zoom-in mechanisms, their direct adaptation to IQA induces spurious "cropping-implies-degradation" biases and misinterprets natural depth-of-field as artifacts. To address these challenges, we propose Q-Probe, the first agentic IQA framework designed to scale IQA to high resolution via context-aware probing. First, we construct Vista-Bench, a pioneering benchmark tailored for fine-grained local degradation analysis in high-resolution IQA settings. Furthermore, we propose a three-stage training paradigm that progressively aligns

*Equal Contribution. †Project Leader. ‡Corresponding Authors.

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

the model with human preferences, while simultaneously eliminating causal bias through a novel context-aware cropping strategy. Extensive experiments demonstrate that Q-Probe achieves state-of-the-art performance in high-resolution settings while maintaining superior efficacy across resolution scales.

1 Introduction

Image Quality Assessment (IQA) serves as a pivotal fundamental technique in computer vision, designed to emulate human perception of visual fidelity [7, 25, 14]. Within this domain, No-Reference (NR) methods [33, 20] are highly valued in practical scenarios due to their independence from pristine reference images. However, the transition from traditional handcrafted features [1] to deep learning paradigms [27, 44] has long been plagued by challenges regarding overfitting and out-of-distribution generalization [50]. The emergence of Multimodal Large Language Models (MLLMs) [34, 2, 35, 31] has provided a novel solution to this dilemma. Through the Chain-of-Thought (CoT) mechanism, MLLMs integrate low-level distortions (e.g., noise, blur) and high-level semantics (e.g., aesthetics, content), markedly improving model generalization [40, 43]. However, current Supervised Fine-Tuning (SFT) paradigms are still hindered by high annotation expenses and shallow reasoning, failing to fully unleash the inherent cognitive potential of MLLMs.

Recent research paradigms are shifting towards leveraging Reinforcement Learning (RL) to further align models with human perception [4, 48]. This direction advocates for the explicit design of reward functions to unify the complementary dimensions of "response consistency" and "preference alignment," thereby overcoming the limitations of traditional approaches that optimize solely for ranking or regression accuracy. Two dominant RL paradigms have emerged in IQA. Regression-based methods (e.g., Q-Insight [18]) treat IQA as absolute score prediction, employing verifiable tolerance rewards to align accuracy with human preference. Conversely, comparison-based methods (e.g., VisualQuality-R1 [41]) derive quality from relative differences. By optimizing ranking distributions against human preferences, they induce emergent CoT capabilities that balance ranking accuracy with logical interpretability.

Nevertheless, RL-based IQA methods suffer from a heavy reliance on global macroscopic views, making it difficult to precisely capture subtle local degradation features (Figure 1(a)). Even when such artifacts are detected, they are frequently overlooked due to insufficient attentional allocation. In real-world scenarios, degradation in high-value semantic regions (e.g., faces or license plates) should trigger a severe penalty in the quality metric, whereas existing models frequently overestimate quality in such cases due to semantic robustness bias (Figure 1(b)). Moreover, these methods are typically restricted to processing global quality. In high-resolution contexts [38, 26], this coarse-grained approach fails to resolve small objects and subtle local artifacts, leading to marked biases in quality prediction.

The emergence of the "Thinking with Images" [6, 24] and multi-turn agentic reasoning [8, 30] paradigm has spurred the use of multi-scale tools for high-resolution tiny object perception [49, 32]. While this offers a promising avenue for fine-grained perception in IQA, directly transposing this local zoom-in strategy presents severe challenges. On one hand, traditional visual zoom-in methods relying on region-level supervision benefit from the diversity of target objects, merely requiring SFT trajectories to encompass the target regions. In the IQA context, however, training exclusively on degraded regions is prone to inducing overfitting, causing the model to establish a spurious causal correlation that "cropping implies low quality" (Figure 1(d)). Consequently, the model loses its discriminative capability regarding content, exhibiting a systematic low-score bias across all local views. On the other hand, even when equipped with local zoom-in tools to facilitate fine-grained semantic analysis, existing MLLMs often overlook the critical dependency of quality assessment on the global context. Taking the Depth-of-Field (DoF) [23] effect in photography as an example, a sharp foreground contrasted with a blurred background typically signifies high-quality imaging. Yet, when isolated from the holistic view and focused solely on the background, models tend to misinterpret natural optical bokeh as degradation (e.g., blur), thereby erroneously assigning a low quality score (Figure 1(c)). Furthermore, the current IQA landscape lacks a benchmark dedicated to local fine-grained degradations in high-resolution scenarios, which significantly constrains the evolution of algorithms toward more refined perception.

To address the aforementioned challenges, we propose **Q-Probe** (Figure 1(e)), the first agentic IQA model grounded in the Thinking with Images paradigm and designed specifically for high-resolution scenarios. To enable such fine-grained perception given the lack of dedicated benchmarks, we first construct **Vista-Bench** (Figure 1(f)), covering a wide range of scenes and local distortion patterns. To ensure the authenticity and precision of the synthesized artifacts, we employ the wavelet transform [45] to decouple image structure from texture, selectively injecting degradations into texture-rich regions to simulate realistic impairments. To guarantee authoritative ground truth, we deploy a dedicated panel of six human specialists to perform a hierarchical evaluation, calculating weighted assessments based on local degradation severity and the semantic significance of the regions.

Building on this foundation, we introduce a three-stage training paradigm mimicking the human visual mechanism of "global perception, local scrutiny," enabling Q-Probe to dynamically optimize tool invocation for comprehensive assessment across resolution scales. Initially, we conduct RL pre-training using a subset of low-resolution images to establish a foundational perception of image quality. This phase is crucial for decoupling the overall learning complexity. By first anchoring the model in the fundamental concept of Mean Opinion Scores (MOS), we establish a stable baseline that allows subsequent stages to focus exclusively on mastering intricate agentic behaviors without the burden of basic quality alignment.

Following this, we curate **Probe-CoT-3K**, a dataset comprising mixed-resolution imagery and reasoning trajectories, to facilitate an SFT warm-up phase. Here, we devise a generation strategy that integrates a global overview with diverse local cropping mechanisms (e.g., degradation capture, clarity localization, and distant view assessment). Specifically, the trajectories simulate an initial global reasoning phase, utilizing a holistic view to determine the necessity of further fine-grained perception. During the local cropping phase, unlike traditional VQA methods that focus solely on target semantics, we mandate that cropped regions encompass all degraded patches while simultaneously preserving areas exhibiting pristine natural DoF or clear foregrounds. This design not only effectively eliminates the spurious causal correlation associating "local cropping" with "low quality" but also empowers the model to discern the optimal timing of tool invocation, grounded in a balanced integration of global and local perspectives. However, the strategy of incorporating background context to ensure logical robustness inevitably reduces the model's precision in localizing subtle degradations, requiring further optimization. To maximize the model's capacity to identify such fine-grained artifacts, we construct the **Probe-RL-4K** dataset and implement Reinforcement Fine-Tuning, synchronously optimizing reasoning trajectories and tool invocation precision for accurate decision-making in complex scenarios. Extensive experimental results demonstrate that Q-Probe significantly outperforms existing models in high-resolution settings, accompanied by varying degrees of performance gains in low-resolution scenarios.

To quantify the practical relevance of this setting, we provide a preliminary prevalence analysis in Appendix F.2: among 1,000 high-resolution AIGC samples, 38.5% are globally acceptable but contain minor local defects; a re-examination of UHD-IQA and DPED indicates regional degradations in roughly 28% and 22% of images, respectively. These dataset-dependent estimates motivate dedicated high-resolution local-degradation evaluation rather than replacing existing global-IQA benchmarks.

Our contributions are summarized as follows:

- We propose Q-Probe, the first agentic IQA model for high-resolution scenarios, leveraging adaptive cropping to achieve superior capability in assessing both global quality and subtle local degradations.
- We introduce Vista-Bench, the pioneering benchmark dedicated to high-resolution fine-grained degradation analysis, encompassing a broad spectrum of scenes and local distortion patterns.
- We construct the Probe-CoT-3K and Probe-RL-4K datasets, leveraging a context-aware cropping strategy to eliminate spurious correlations and generate high-quality CoT trajectories for mastering tool usage.
- We devise a progressive three-stage training curriculum that first aligns global perception with human preference, then stabilizes logical reasoning via SFT, and finally leverages decoupled reward guidance to recover localization precision.

2 Related Work

2.1 Image Quality Assessment

Image Quality Assessment (IQA) has evolved from early reliance on handcrafted features based on Natural Scene Statistics (NSS) [21, 22] to data-driven deep learning models. While subsequent CNN and Transformer-based approaches achieved performance gains, they often treated IQA as a black-box regression task. The recent advent of Multimodal Large Language Models (MLLMs) introduced a semantic turn, with methods like Q-Align [39] and DeQA-Score [43] leveraging foundation models for quality scoring via supervised fine-tuning (SFT). However, these SFT-based paradigms rely heavily on expensive annotations and often suffer from shallow reasoning patterns. To address these limitations, Reinforcement Learning (RL) has emerged as a pivotal mechanism for aligning MLLMs with human preferences. Notably, Q-Insight [18] pioneered the application of Group Relative Policy Optimization (GRPO) to visual quality understanding, while VisualQuality-R1 [41] reformulated IQA as a ranking task via RL-to-Rank.

2.2 Thinking with Images

The "Thinking with Images" paradigm, introduced by OpenAI-o3 [24], has spurred the development of agentic visual reasoning MLLMs [13, 29] that interleave image and text reasoning with iterative visual analysis [5, 12]. DeepEyes [49] provides an open-source implementation, demonstrating that end-to-end RL can incentivize models to adopt this behavior for fine-grained tasks. However, subsequent works reveal that pure RL is insufficient for complex, multi-turn interactions. Pixel Reasoner [32] identifies a critical "learning trap" where models bypass nascent visual tools, proposing a two-phase approach: a cold-start phase to establish tool use followed by curiosity-driven RL. Similarly, other works have adopted multi-stage training to activate deep trajectories for hard visual searches. Despite these advances, directly transplanting this agentic probing to IQA remains non-trivial. Unlike object detection, cropping in IQA introduces a causal ambiguity: models trained on cropped degradation patches tend to overfit, learning a spurious correlation that "zooming implies degradation," thereby penalizing high-quality crops. Q-Probe addresses this gap by introducing a context-aware probing mechanism.

3 Methodology

3.1 Overview

Q-Probe emulates the human "coarse-to-fine" visual mechanism via a progressive three-stage curriculum, balancing global aesthetic perception with local defect scrutiny. To facilitate this, we constructed Vista-Bench, a high-resolution benchmark derived from public datasets through wavelet-based artifact injection and manual scoring by an expert panel. The training pipeline of Q-Probe begins with perception alignment using Group Relative Policy Optimization (GRPO) [9] to align with the human concept of Mean Opinion Scores (MOS). It then progresses to a hybrid-resolution SFT phase driven by a data synthesis pipeline, employing context-aware cropping to dissociate tool usage from negative quality bias. Finally, a decoupled RL post-training stage utilizes a bifurcated reward mechanism to jointly optimize defect localization and scoring, resolving the conflict between exploration diversity and detection precision. Details are shown in Figure 3.

3.2 Data Construction: Vista-Bench

Existing datasets usually consist of low-resolution images (< 1K) where degradations are typically applied globally rather than to localized regions. To evaluate fine-grained perception in the high-resolution quality assessment era, we constructed the **Vista-Bench**.

The dataset is curated from high-quality samples within the HR-Bench 4K [36], Pixel-Reasoner [32] and UHD-IQA [11] datasets. As shown in Figure 2, the construction pipeline proceeds as follows: we first utilize wavelet transforms to extract global high-frequency components and apply specific degradations (e.g., blur, compression, mosaic). This approach ensures that the degradation is applied to the semantic foreground regions as much as possible. Subsequently, we employ a dedicated panel of six human experts to conduct a rigorous, two-stage manual annotation. The evaluators first assess

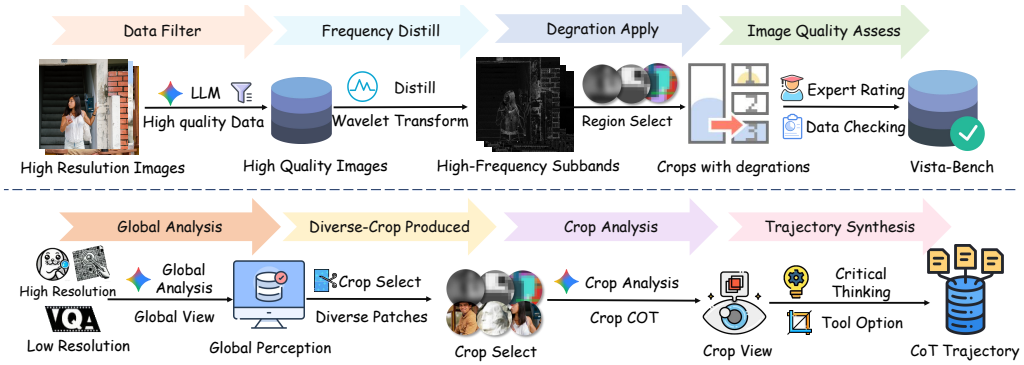


Figure 2: This diagram illustrates the construction pipeline of Vista-Bench and the data synthesis pipeline for SFT. Specifically, we utilize wavelet transforms to decouple structure from texture, selectively injecting artifacts into texture-rich semantic regions. To ensure annotation accuracy, we employ a panel of domain experts to manually generate importance-weighted scores for fine-grained perception probing. To support SFT, we generate traces that interleave global overviews, defect zooming, and context verification (scrutinizing clear regions), thereby preventing the model from associating tool usage solely with defects.

the global visual quality, and then scrutinize the explicitly cropped regions to generate fine-grained quality scores based on the severity of the local degradation and the semantic importance of the region. The resulting benchmark encompasses over 3,000 high-resolution (> 4K) images, characterized by authoritative, multi-domain, and locally annotated degradations. The six specialists are senior PhD researchers in our laboratory specializing in Image Quality Assessment. Vista-Bench is split into training and test sets at a 9:1 ratio. Probe-CoT-3K and Probe-RL-4K are drawn exclusively from the Vista-Bench training split. For KADID-10K, the 3,000 Stage-1 images are randomly sampled only after excluding the test set. We provide more details in Appendix Section A.

3.3 Stage 1: Perception Alignment via Pre-RL

Why Pre-alignment is Required. Before introducing complex tool usage, we strategically decouple the learning objectives to prevent the model from being overwhelmed by simultaneous tasks. Requiring the policy to concurrently learn whether to zoom in, where to zoom in, how to weigh the importance of local crops against the global image, and the fundamental concept of Mean Opinion Scores (MOS) makes the optimization highly unstable. Therefore, we employ a dataset of 3,000 low-resolution images randomly sampled from the KADID-10k [19] dataset to train the model via a pairwise variant of GRPO. This initial stage explicitly isolates and anchors the learning of the MOS concept and foundational aesthetic perception. Consequently, Stage 2 can focus exclusively on mastering the three intricate crop-related capabilities without the confounding burden of basic quality alignment.

Why not use SFT as Stage 1. A fundamental innovation of our Q-Probe framework lies in its pioneering three-stage training curriculum: GRPO $\rightarrow$ SFT $\rightarrow$ GRPO. To the best of our knowledge, this is the first framework to adopt such an interleaved reinforcement learning paradigm for multimodal visual quality assessment. We observed that employing SFT as Stage 1 significantly entrenches the model's fixation on global quality analysis, rendering it exceedingly difficult to learn the relevant zoom-in invocation capabilities during the second stage. Crucially, utilizing standard SFT to directly instill the concept of Mean Opinion Scores (MOS) inevitably forces the model to anchor its judgments on macroscopic, global-view heuristics. This deepens the "Semantic Robustness Bias," causing the model's cognitive process to be disproportionately dominated by holistic image semantics rather than objective local fidelity. Once trapped in this global bias, the model exhibits severe resistance during Stage 2; it struggles to learn the intricate mechanics of executing cropping actions, analyzing subtle localized defects, and reconciling these micro-imperfections with the overall global context. For more extended discussion, please refer to Appendix Section F.3.



Figure 3: Overview of the three-stage training framework. Initially, RL Pre-training leverages ranking rewards to align global perception with human preferences. Subsequently, hybrid-resolution SFT enables the model to acquire robust logical reasoning. Finally, the RL Post-training stage fine-tunes the model for precise degradation detection and adaptive tool invocation.

Probabilistic Ranking via Thurstone Model. Unlike standard regression approaches that treat quality scores as deterministic point estimates, we model visual quality as an intrinsically relative concept governed by the Thurstone model [28] Case III. For a training pair of images $\{x_i, x_j\}$ (effectively a batch size of $B = 2$), the policy model π_θ generates K distinct reasoning paths and quality scores for each image: $q(x_i) = [q_1(x_i), \ldots, q_K(x_i)]^\intercal$. This distribution inherently captures predictive uncertainty, which serves as a cornerstone for reliable ranking.

We calculate the asymmetric comparative probability that x_i is perceptually superior to x_j by standardizing the difference in their predicted means against their joint uncertainty. To improve numerical stability, we first define the standardized difference score $\mathcal{Z}_{ij}^k$:

$$\mathcal{Z}_{ij}^k = \frac{q_k(x_i) - \mu(q(x_j))}{\sqrt{\sigma^2(q(x_i)) + \sigma^2(q(x_j)) + \gamma}} \tag{1}$$

where $\mu(\cdot)$ and $\sigma^2(\cdot)$ denote the sample mean and variance of the predicted scores group, $q_k(x_i)$ represents the k-th score estimate for image x_i, and γ is a small constant ensuring numerical stability. The comparative probability is then derived via the standard Gaussian Cumulative Distribution Function (CDF):

$$P_k(x_i \succ x_j) = \Phi\left(\mathcal{Z}_{ij}^k\right) \tag{2}$$

where $\Phi(\cdot)$ is the Gaussian CDF, representing the probability P_k that image x_i is perceptually superior to x_j. Crucially, we explicitly leverage the sample variances derived from the GRPO group outputs to dynamically accommodate predictive uncertainty, preventing over-confident errors on ambiguous images.

Fidelity Reward and Optimization. To rapidly align the model's predictions with human perception, we define the ground truth preference y_{ij} based on Mean Opinion Scores (MOS): $y_{ij} = 1$ if $\text{MOS}(x_i) > \text{MOS}(x_j)$, 0.5 if equal, and 0 otherwise. The optimization is driven by a continuous fidelity reward. We define the ranking reward R_{rank} as:

$$R_{rank}(x_i) = y_{ij} \cdot P_k(x_i \succ x_j) + (1 - y_{ij}) \cdot (1 - P_k(x_i \succ x_j)) \tag{3}$$

where $y_{ij} \in \{0, 0.5, 1\}$ represents the ground truth preference label derived from human ratings. For tied pairs with $y_{ij} = 0.5$, the displayed reward reduces to $R_{rank} = 0.5$ regardless of $P_k(x_i \succ x_j)$. This reward is designed to capture fine-grained distinctions in quality ranking. For each prompt, let $\mathcal{R}_k$ denote the reward of the k-th rollout and $\mathcal{R}_{group} = \{\mathcal{R}_k\}_{k=1}^{K}$. The group-relative advantage used in the GRPO objective is

$$A_k = \frac{\mathcal{R}_k - \text{mean}(\mathcal{R}_{group})}{\text{std}(\mathcal{R}_{group}) + \epsilon}, \tag{4}$$

where ϵ is a small numerical-stability constant. To ensure training stability, we maximize the GRPO objective $\mathcal{J}(\theta)$, which integrates the clipped surrogate advantage and a KL-divergence penalty:

$$\mathcal{J}(\theta) = \mathbb{E}\left[\min\left(\frac{\pi_\theta}{\pi_{old}} A_k, \text{clip}\left(\frac{\pi_\theta}{\pi_{old}}, 1-\epsilon, 1+\epsilon \right) A_k \right) - \beta D_{KL}(\pi_\theta || \pi_{ref}) \right] \tag{5}$$

where π_θ and π_{old} represent the current and previous policies respectively, A_k is the advantage score computed from reward signals, ϵ is the clipping threshold to constrain policy updates, π_{ref} is the reference policy (typically the initial SFT model), and β is a coefficient controlling the strength of the KL-divergence penalty D_{KL}. This pre-alignment stage establishes a solid baseline for global image understanding, converging towards the mechanisms of human MOS.

3.4 Stage 2: Hybrid-Resolution SFT via Data Synthesis Pipeline

To bridge the gap between global perception and local scrutiny, we construct the **Probe-CoT-3K** dataset using a data synthesis pipeline (as illustrated in Figure 2). The dataset follows a 2:1 ratio of low-to-high resolution images.

The SFT Dilemma: Precision vs. Diversity. In traditional Visual Question Answering (VQA), crop content inherently exhibits high diversity. However, in Agentic IQA, if SFT data is invariably centered on degradations, the model overfits to a biased mode: *"Tool Usage implies Degradation and Low Quality,"* resulting in a fundamental breakdown of logical reasoning.

Context-Aware Trajectory Generation. To mitigate this, we synthesize CoT trajectories encompassing genuine degradations, natural depth-of-field, and pristine foregrounds. This strategy effectively severs the spurious causal correlation associating "cropping" with "low quality".

Specifically, the SFT data primarily includes two types. First, the model can directly observe the global details and does not need to zoom in. Second, for high-resolution images or when global details are unclear, the model decides to zoom in for further analysis. After zooming in, if a defect is discovered, the model applies a penalty based on the location's importance to the overall image; however, if no defects are found, or if the model reasons that the observed blurriness is simply a normal out-of-focus background (depth of field), it decides not to penalize.

We optimize the model using the standard cross-entropy loss over the mixed sequence of reasoning and action tokens y:

$$\mathcal{L}_{SFT} = -\sum_{t=1}^{L} \log P(y_t | y_{<t}, \mathbf{x}) \tag{6}$$

where $\mathcal{L}_{SFT}$ is the SFT loss, L denotes the sequence length, $\mathbf{x}$ represents the multimodal input (image and instruction), and y_t is the t-th token in the generated target sequence y given the preceding tokens $y_{<t}$. Following the SFT phase, the model establishes a robust understanding of the global image context and acquires the capability to accurately discern diverse local regional features. However, intentionally introducing a zero-penalty region negatively impacts the model's zoom-in accuracy. Experimental results regarding this can be found in Appendix Section F.5.

3.5 Stage 3: Precision Pursuit via Decoupled Post-RL

As the SFT phase strategically trades localization precision for logical robustness, we construct and leverage the **Probe-RL-4K** dataset in the final stage to further refine the model's fine-grained detection capabilities.

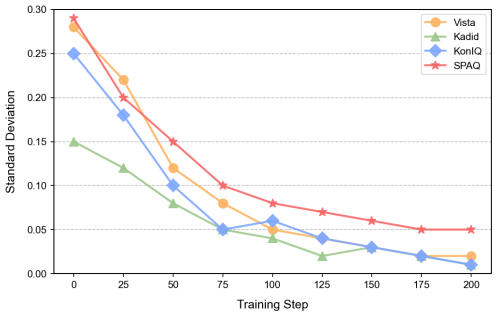


Figure 4: We calculated the average standard deviation of predicted scores across multiple inference runs at various checkpoints. The observed monotonic decrease in variance confirms that Stage-1 achieves great stability.

Table 3: Ablation of the Reward Mechanism in Stage 3 on Vista-Bench.

Reward Strategy	**SRCC**	**PLCC**
R_{acc} (Score Only)	0.770	0.821
$R_{acc} + R_{format}$	0.778	0.826
$R_{acc} + R_{format} + R_{loc}$ **(Ours)**	**0.801**	**0.850**

Table 4: Ablation of Crop Coverage Strategies regarding Degradation Regions on Vista-Bench.

Crop Strategy	**SRCC**	**PLCC**
Degradation Only	0.578	0.650
Partial Degradation & Partial Normal	0.767	0.819
All Degradation & Partial Normal	**0.801**	**0.850**

Decoupled Reward Mechanism. We implement a decoupled reward mechanism that disentangles the optimization of the "Looking Policy" from the "Scoring Policy." We independently define the accuracy reward R_{acc} and the localization reward R_{loc}:

$$R_{acc} = \exp\left(-\frac{|s_{pred} - s_{MOS}|}{\tau}\right) \tag{7}$$

$$R_{loc} = \mathbb{I}(has_defect) \cdot \text{IoU}(B_{pred}, B_{gt}) \tag{8}$$

where s_{pred} and s_{MOS} denote the predicted score and ground truth MOS respectively, and τ is a temperature hyperparameter controlling the sensitivity of the accuracy reward. For localization, $\mathbb{I}(\cdot)$ is an indicator function that equals 1 if a defect exists in the image, and $\text{IoU}(\cdot)$ computes the Intersection over Union between the predicted crop box B_{pred} and the ground truth defect region B_{gt}. This reward incentivizes the model to accurately capture degradations and precisely deploy tools for regional magnification.

For multi-defect and multi-call trajectories, each ground-truth box is matched independently by the maximum IoU over all tool calls:

$$R_{loc} = \frac{1}{|G|} \sum_{g \in G} \max_{b \in B_{pred}} \text{IoU}(b, g), \tag{9}$$

when defects exist, and $R_{loc} = 0$ otherwise. A missed ground-truth box contributes zero; extra failed crops do not cancel a successful match. We do not add an explicit false-call penalty, because a hard penalty destabilizes exploration in early training; the accuracy reward, format constraints, and termination behavior indirectly discourage unnecessary calls.

The total reward R_{total} is then composed as:

$$R_{total} = \underbrace{\alpha R_{acc}}_{\text{Scoring}} + \underbrace{\beta R_{loc}}_{\text{Looking}} + \gamma R_{format} \tag{10}$$

where α, β, and γ are hyperparameters weighting the contribution of the accuracy reward, localization reward, and format compliance reward R_{format}, respectively. For completeness, $R_{format} = 1$ if the response contains a valid final `<score>...</score>` tag and any tool call follows the prescribed XML schema, and $R_{format} = 0$ otherwise. This decoupled reward optimizes the model to refine its localization capabilities (βR_{loc}), building upon the robust logic reasoning capability acquired during SFT.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. To evaluate the comprehensive capabilities of Q-Probe, we conduct experiments across a diverse set of IQA benchmarks. For standard resolution assessments, we utilize

Table 1: Comprehensive performance comparison (SRCC / PLCC) on standard and high-resolution datasets. **Q-Probe** achieves SOTA results, especially on the high-resolution **Vista** benchmark. Methods are color-coded by category: Handcrafted, Deep Learning, and MLLMs-based.

Methods	Vista	SPAQ	KADID	PIPAL	TID13	KonIQ	AGIQA	Avg
Spearman Rank Correlation Coefficient (SRCC)								
BRISQUE [21]	0.211	0.614	0.429	0.242	0.548	0.385	0.497	0.418
NIQE [22]	0.247	0.676	0.487	0.357	0.532	0.421	0.533	0.465
MUSIQ [16]	0.357	0.720	0.647	0.317	0.670	0.473	0.494	0.525
UNIQUE [46]	0.373	0.751	0.513	0.393	0.703	0.649	0.608	0.570
MANIQA [42]	0.388	0.745	0.760	0.338	0.589	0.213	0.422	0.494
Qwen2.5-VL-7B [3]	0.450	0.848	0.787	0.390	0.787	0.754	0.735	0.679
LIQE [47]	0.406	0.815	0.809	0.371	0.718	0.684	0.653	0.637
DeQA-Score [43]	0.463	0.852	0.831	0.383	0.756	0.677	0.738	0.671
Q-Align [39]	0.424	0.767	0.832	0.406	0.769	0.573	0.682	0.636
UnifiedReward-T [37]	0.477	0.871	0.841	0.399	0.788	0.820	0.722	0.703
Q-Insight [18]	0.429	0.872	0.856	0.429	0.816	0.806	0.749	0.708
VisualQuality-R1 [41]	**0.517**	**0.875**	**0.871**	**0.469**	**0.848**	**0.855**	**0.805**	**0.749**
Q-Probe (Ours)	**0.801**	**0.892**	**0.901**	**0.474**	**0.829**	**0.871**	**0.837**	**0.801**
Pearson Linear Correlation Coefficient (PLCC)								
BRISQUE [21]	0.224	0.624	0.451	0.259	0.546	0.400	0.541	0.435
NIQE [22]	0.252	0.683	0.415	0.314	0.516	0.439	0.560	0.454
MUSIQ [16]	0.350	0.666	0.622	0.347	0.695	0.435	0.434	0.507
UNIQUE [46]	0.368	0.708	0.548	0.361	0.729	0.590	0.581	0.555
MANIQA [42]	0.381	0.753	0.780	0.364	0.617	0.257	0.448	0.514
Qwen2.5-VL-7B [3]	0.426	0.854	0.806	0.420	0.837	0.810	0.772	0.704
LIQE [47]	0.398	0.814	0.817	0.347	0.748	0.652	0.653	0.633
DeQA-Score [43]	0.439	0.858	0.873	0.381	0.793	0.703	0.743	0.684
Q-Align [39]	0.412	0.779	0.862	0.381	0.794	0.612	0.694	0.648
UnifiedReward-T [37]	0.460	0.846	0.877	0.421	**0.873**	0.804	0.745	0.718
Q-Insight [18]	0.410	0.872	**0.881**	**0.462**	0.851	0.829	0.794	0.728
VisualQuality-R1 [41]	**0.493**	**0.878**	0.821	0.458	0.871	**0.840**	**0.843**	**0.743**
Q-Probe (Ours)	**0.850**	**0.900**	**0.892**	**0.476**	**0.876**	**0.863**	**0.813**	**0.810**

Table 2: Ablation study of the progressive training strategy across multiple benchmarks.

Model Configuration	Vista		SPAQ		KADID-10k		KonIQ-10k		Average	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Stage 1 Only	0.352	0.404	0.760	0.780	0.700	0.725	0.650	0.680	0.615	0.647
Stage 2 + Stage 3	0.751	0.795	0.766	0.796	0.781	0.772	0.752	0.769	0.762	0.783
Stage 1 + Stage 2	0.762	0.808	0.850	0.860	0.880	0.875	0.820	0.815	0.828	0.839
Full (Stage 1+2+3)	**0.801**	**0.850**	**0.892**	**0.900**	**0.901**	**0.892**	**0.871**	**0.863**	**0.866**	**0.876**

KonIQ-10k [10], SPAQ [7], and KADID-10k [19] to represent in-the-wild and synthetic distortions. We also include PIPAL [15] to evaluate robustness against algorithm-processed distortions and AGIQA-3K [17] for AI-generated content. Crucially, to validate fine-grained perception in the high-resolution era, we employ our proposed Vista-Bench, which contains localized artifacts in high-resolution scenarios. We leverage the widely recognized Spearman Rank-order Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC) as our primary evaluation metrics. Following established protocols [39, 7], we compute the average performance across datasets

Table 5: Run-to-run stability for principal comparisons (mean ± standard deviation over independent runs).

Method	Vista SRCC	Vista PLCC	SPAQ SRCC
VisualQuality-R1	0.517 ± 0.021	0.493 ± 0.025	0.875 ± 0.009
Q-Probe (Ours)	$\mathbf{0.801 \pm 0.012}$	$\mathbf{0.850 \pm 0.015}$	$\mathbf{0.892 \pm 0.008}$

to assess overall generalization. Table 1 presents the SRCC and PLCC performance across seven benchmarks.

Implementation Details. We utilize Qwen-2.5-VL-7B [3] as our base model. During the perception alignment stage, we employ GRPO with a generation number $K = 6$ and a clip threshold $\epsilon = 0.2$. The model is trained using the AdamW optimizer with a learning rate of 1×10^{-6}. For the hybrid-resolution SFT, we utilize the Probe-CoT-3K dataset with a crop resolution of 768×768. All experiments are conducted on 8 NVIDIA A100 (80GB) GPUs.

Inference cost. With up to six interaction turns, Q-Probe averages 3.6 crop operations, 10.7 seconds of latency, and approximately 1,700 tokens per high-resolution image on our inference setup. This overhead is the cost of active probing; single-pass baselines use one forward pass and therefore do not provide a matched-compute comparison.

Table 5 reports run-to-run intervals for the principal comparisons.

4.2 Ablation Studies

Necessity of the Three-Stage Curriculum. Table 2 validates the efficacy of our progressive training curriculum. Although Stage 1 performs poorly on high-resolution details, it successfully pre-aligns the model with human aesthetic perception on standard benchmarks. Without Stage 1, model can't reach its best performance. Building on this foundation, Stage 2 bridges the granularity gap via hybrid-resolution training, while Stage 3 further refines the scoring precision through RL to achieve the best overall performance. Besides, three Stages consistently boost performance on standard low-resolution benchmarks, culminating in the best overall results.

Impact of Reward Components in Post-RL. In Stage 3, we analyze the contribution of different reward components. As shown in Table 3, using only the Accuracy Reward (R_{acc}) provides a solid baseline. Incorporating the Format Reward (R_{format}) yields a slight improvement by ensuring the structural validity of the CoT reasoning. However, the most significant gain comes from the Decoupled Localization Reward (R_{loc}), which explicitly guides the model's attention to defects, pushing the SRCC to 0.801. Finer-grained experiments can be seen in Appendix Section F.6.

Impact of Crop Coverage Strategy. Table 4 examines the spatial relationship between crops and degradation. Restricting crops to *exact* degradation regions (Strategy 1) leads to severe overfitting, where the model learns biased inference and associates all crops with degradation, dropping SRCC to 0.578. Strategy 2, covering *partial* degradation and normal areas, improves to 0.767 but introduces trade-off that the model may overlook omitted defects. Finally, Strategy 3 achieves optimal performance (SRCC 0.801) by capturing *all* degradation within a normal context, ensuring comprehensive defect assessment without hallucinations.

5 Conclusion

In this work, we introduced Q-Probe, the first agentic framework designed to scale IQA to high-resolution scenarios via context-aware probing. Recognizing the limitations of existing global-view RL methods in capturing fine-grained artifacts, we proposed a novel three-stage training curriculum that mimics the human "coarse-to-fine" perception mechanism, teaching the model to evaluate quality based on importance of degraded crops. Furthermore, by constructing the Vista-Bench and leveraging a data synthesis pipeline, we eliminated the spurious "cropping-implies-degradation" bias, enabling the model to intelligently distinguish between technical distortions and artistic effects. Our model achieves SOTA performance on both high-resolution and low-resolution datasets.

6 Acknowledgements and Disclosures

Funding. This work was supported by the National Natural Science Foundation of China (No. 32672516) and the Central-Oriented Foundation for Local Science and Technology Development of China (No. 202407a12020010).

Competing Interests. The authors declare no competing interests.

References

[1] Nisar Ahmed, Hafiz Muhammad Shahzad Asif, and Hassan Khalid. Image quality assessment using a combination of hand-crafted and deep features. In *International Conference on Intelligent Technologies and Applications*, pages 593–605. Springer, 2019.

[2] 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.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] 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.

[5] Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, et al. Sensenova-mars: Empowering multimodal agentic reasoning and search via reinforcement learning. *arXiv preprint arXiv:2512.24330*, 2025.

[6] Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, et al. Sensesearch: Empowering vision-language models with high-resolution agentic search-reasoning via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26508–26517, 2026.

[7] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3677–3686, 2020.

[8] Zhenkun Gao, Yicheng Bao, Jinlong Peng, Xueheng Li, Theo Huang, Bangwei Liu, Kunquan Li, Zhenye Gan, Tao Hu, Chengjun Xie, et al. Videosearcher: Empowering video deep research with multi-tool agentic reasoning via reinforcement learning. *arXiv preprint arXiv:2607.02927*, 2026.

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

[10] 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.

[11] Vlad Hosu, Lorenzo Agnolucci, Oliver Wiedemann, Daisuke Iso, and Dietmar Saupe. Uhd-iqa benchmark database: Pushing the boundaries of blind photo quality assessment. In *European Conference on Computer Vision*, pages 467–482. Springer, 2024.

[12] Tao Hu, Jiaxin Ai, Licheng Wen, Xueheng Li, Shu Zou, Siqi Li, Nianchen Deng, Xinyu Cai, Hongbin Zhou, Pinlong Cai, et al. Itercad: An iterative multimodal agent for visually-grounded cad generation and editing. *arXiv preprint arXiv:2606.13368*, 2026.

[13] Ziheng Huang, Yicheng Bao, Xueheng Li, Zhenkun Gao, Bangwei Liu, Kunquan Li, Yuxiang Shen, Bangyan Li, Xuejiao Wang, Changbo Wang, et al. Watch-think-interact: Bootstrapping long-horizon multi-turn streaming video reasoning with reinforcement learning. *arXiv preprint arXiv:2609.37035*, 2026.

[14] Ziheng Jia, Jiaying Qian, Zicheng Zhang, Zijian Chen, and Xiongkuo Min. Refine-iqa: Multi-stage reinforcement finetuning for perceptual image quality assessment. *arXiv preprint arXiv:2508.03763*, 2025.

[15] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European conference on computer vision*, pages 633–651. Springer, 2020.

[16] 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*, pages 5148–5157, 2021.

[17] 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, 2023.

[18] Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679*, 2025.

[19] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019.

[20] Qingyu Mao, Shuai Liu, Qilei Li, Gwanggil Jeon, Hyunbum Kim, and David Camacho. No-reference image quality assessment: Past, present, and future. *Expert Systems*, 42(3):e13842, 2025.

[21] 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.

[22] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[23] Hajime Nagahara, Sujit Kuthirummal, Changyin Zhou, and Shree K Nayar. Flexible depth of field photography. In *European Conference on Computer Vision*, pages 60–73. Springer, 2008.

[24] OpenAI. Thinking with images. `https://openai.com/index/thinking-with-images/`, 2025.

[25] Avinab Saha, Sandeep Mishra, and Alan C Bovik. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5846–5855, 2023.

[26] Yuxiang Shen, Hailong Huang, Zhenkun Gao, Xueheng Li, Man Zhou, Chengjun Xie, Haoxuan Che, Xuanhua He, and Jie Zhang. Svfeye: A semantic-visual fusion framework with multi-scale visual context for multimodal reasoning. *arXiv e-prints*, pages arXiv–2603, 2026.

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

[28] Louis L Thurstone. A law of comparative judgment. In *Scaling*, pages 81–92. Routledge, 2017.

[29] Shi-Yu Tian, Zhuo-Xia Wang, Xuan-Yi Zhu, Zhi Zhou, Xinwei Yang, Kun-Yang Yu, Ming Yang, Yang Chen, and Yu-Feng Li. Self-evolving neuro-symbolic skills for tool-augmented spatial reasoning. *arXiv preprint arXiv:2608.07955*, 2026.

[30] Shi-Yu Tian, Zhi Zhou, Kun-Yang Yu, Ming Yang, Yang Chen, Ziqiao Shang, Lan-Zhe Guo, and Yu-Feng Li. Last: Leveraging tools as hints to enhance spatial reasoning for multimodal large language models. *arXiv preprint arXiv:2604.09712*, 2026.

[31] Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, et al. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue. *arXiv preprint arXiv:2510.13747*, 2025.

[32] Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025.

[33] Lanjiang Wang. A survey on iqa. *arXiv preprint arXiv:2109.00347*, 2021.

[34] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[35] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[36] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7907–7915, 2025.

[37] Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025.

[38] Qianshan Wei, Yishan Yang, Siyi Wang, Jinglin Chen, Binyu Wang, Jiaming Wang, Shuang Chen, Zechen Li, Yang Shi, Yuqi Tang, et al. Agentic-mme: What agentic capability really brings to multimodal intelligence? *arXiv preprint arXiv:2604.03016*, 2026.

[39] 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.

[40] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25490–25500, 2024.

[41] Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. *arXiv preprint arXiv:2505.14460*, 2025.

[42] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1191–1200, 2022.

[43] 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*, pages 14483–14494, 2025.

[44] Zihao Yu, Fengbin Guan, Yiting Lu, Xin Li, and Zhibo Chen. Sf-iqa: Quality and similarity integration for ai generated image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6692–6701, 2024.

[45] Dengsheng Zhang. Wavelet transform. In *Fundamentals of image data mining: Analysis, Features, Classification and Retrieval*, pages 35–44. Springer, 2019.

[46] 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.

[47] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14071–14081, 2023.

[48] Shijie Zhao, Xuanyu Zhang, Weiqi Li, Junlin Li, Li Zhang, Tianfan Xue, and Jian Zhang. Reasoning as representation: Rethinking visual reinforcement learning in image quality assessment. *arXiv preprint arXiv:2510.11369*, 2025.

[49] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv:2505.14362*, 2025.

[50] Yan Zhong, Xingyu Wu, Li Zhang, Chenxi Yang, and Tingting Jiang. Causal-iqa: Towards the generalization of image quality assessment based on causal inference. In *Forty-first International Conference on Machine Learning*, 2024.

Appendix of "Q-Probe: Scaling Image Quality Assessment to High Resolution via Context-Aware Agentic Probing"

A Data Construction and Trajectory Synthesis Details

To ensure the utmost reliability and authority of **Vista-Bench** in scaling Image Quality Assessment (IQA) to high-resolution scenarios, we implemented a rigorous manual annotation protocol. This protocol relies on a dedicated panel of six specialists to generate fine-grained, hierarchical quality annotations. This section details the human evaluation workflow and the specific context-aware evaluation guidelines employed.

A.1 Wavelet-Based Degradation Injection

Existing IQA datasets often rely on global distortions (e.g., Gaussian blur applied uniformly), which allow models to bypass genuine visual scrutiny by learning simple global statistical deviations. To prevent such "shortcut learning" and simulate realistic high-resolution defects, we employ a targeted degradation injection strategy based on the Discrete Wavelet Transform (DWT).

Structure-Texture Decoupling. We utilize the Haar wavelet to decompose high-resolution source images into four distinct frequency subbands: the low-frequency approximation (LL) and three high-frequency detail components (LH, HL, HH). The LL band preserves the global semantic structure and flat regions, while the high-frequency bands capture fine-grained textures and edges where high-resolution artifacts typically manifest, as shown in Figure 5.

Targeted Artifact Synthesis. Instead of corrupting the entire image, we selectively inject degradations (e.g., compression ringing, sensor noise, and slight defocus) specifically into the high-frequency subbands. This strategy ensures that artifacts are embedded within complex texture regions—such as hair, foliage, or architectural details—making them visually subtle yet perceptually critical. By keeping the LL band intact, we maintain the overall structural integrity of the scene while introducing realistic local impairments.

Reconstruction. The final degraded samples are generated via the Inverse Discrete Wavelet Transform (IDWT). This process forces the IQA agent to actively "zoom in" and scrutinize local details to detect quality drops, effectively simulating the fine-grained perception required for 4K scenarios.

A.2 Complete Instructions for Human Expert Evaluators

As transparency and reproducibility are critical for human-annotated benchmarks, we formalize the exact guidelines provided to our panel of six specialists. To prevent subjective bias and ensure a standardized "coarse-to-fine" perception, all evaluators were required to strictly adhere to the following four-phase evaluation protocol during the construction of Vista-Bench:

Vista-Bench Human Annotation Guidelines

Objective: Provide a hierarchical quality assessment (Global Base Score followed by Local Penalties) for high-resolution images containing subtly injected artifacts.

Phase 1: Global Quality Assessment

- View the image at a standard global scale (fit-to-screen).
- Assign a **Target Base Score** (1.0 to 5.0) reflecting the overall structural integrity, lighting, and general aesthetic fidelity. Do not scrutinize microscopic details at this stage.

Phase 2: Local Fine-Grained Scrutiny

- Zoom in and actively inspect highly-textured semantic regions (e.g., hair, foliage, building edges) and distinct background areas.
- Identify the presence of specific man-made artifacts injected via wavelet transform: *Mosaic/Pixelation, JPEG Compression Ringing, Posterization, Scanlines, or Grid lines.*

Phase 3: Semantic Weighting and Final Scoring

- **The "Zero-Penalty" Rule:** If the blur or texture loss in a cropped region is strictly due to natural photographic Depth-of-Field (bokeh), apply a **0.0 penalty**.
- **Semantic Significance Weighting:** If artificial degradations are found, the penalty must be weighted by the semantic importance of the region:
  - *Primary Subject / Salient Foreground:* Apply a severe penalty (e.g., -1.0 to -2.0).
  - *Background / Non-Salient Regions:* Apply a minor to moderate penalty (e.g., -0.2 to -0.5), adjusting slightly based on spatial context (e.g., confined indoor backgrounds require higher penalties than distant outdoor backgrounds).
- Subtract the calculated penalties from the Target Base Score to yield the **Target Final Score**.

Phase 4: Consensus and Quality Assurance

- Flag the image sample for discarding if there is a high variance in scores among panel members, or if there is fundamental disagreement regarding the semantic classification (e.g., conflicting views on whether a degraded region belongs to the foreground subject or the background).

The panel consists of six senior PhD researchers specializing in IQA. All annotations were performed on calibrated professional displays following a standardized "fit-to-screen first, zoom-in second" protocol. For score aggregation, a sample was discarded if the variance of the six expert scores exceeded 0.8, or if the experts could not resolve whether the degraded region belonged to the foreground subject or the background. Approximately 8,000 samples were discarded by these filters. For retained samples, the final MOS is the arithmetic mean of all valid expert scores.

A.3 Asynchronous CoT Trajectory Generation Pipeline

Given the high-resolution nature of the source images and the need for dense cropping, sequentially generating reasoning trajectories for thousands of images would be computationally prohibitive. We implemented a Python-based asynchronous framework tailored for high-concurrency interaction with the Multimodal LLM (MLLM) to synthesize the Chain-of-Thought (CoT) data. The pipeline consists of three critical stages:

1. **Dynamic Image Preprocessing and Encoding:** To optimize the balance between visual fidelity and token consumption, we implemented an adaptive resizing strategy. Original high-resolution images are resized such that their maximum edge does not exceed 1,024 pixels ($H_{max} = 1024$) using Lanczos resampling. This ensures that global artifacts are preserved while significantly reducing bandwidth usage. Images are then serialized into Base64 format for API transmission.
2. **Semaphore-Based Concurrency Control:** To maximize throughput without exceeding API rate limits, the system utilizes an asynchronous semaphore mechanism (set to a concurrency limit of 5). This allows for parallel processing of image batches while maintaining system stability. The pipeline utilizes `aiohttp` for non-blocking network I/O, ensuring that image encoding and API waiting times do not bottle-neck the **trajectory synthesis process**.
3. **Robust Output Parsing:** Since MLLM outputs can vary in formatting, we implemented a resilient parsing logic that utilizes regular expressions to extract JSON structures from mixed-text responses. This ensures valid CoT reasoning extraction even if the model wraps outputs in Markdown fences or custom XML-like tags.

A.4 Context-Aware Prompt Design for Trajectory Synthesis

A critical challenge in generating SFT reasoning trajectories is ensuring the model outputs a logical path that leads to the human-annotated ground truth, *without* explicitly revealing to the trained model that the ground truth was provided. If the ground truth is exposed in the user prompt during SFT, the model will suffer from data leakage and fail to learn genuine visual evaluation.

To eliminate the risk of visual hallucination synthesis, our generation pipeline utilizes a strictly constrained **Hidden-State Anchored Prompt**. During this phase, the Multimodal LLM (acting as

the trajectory synthesizer) is secretly provided not only with the global target scores but also with the **exact degradation labels and types for each specific cropped region** (e.g., whether a crop contains mosaic, scanlines, or is clean). The MLLM is instructed to construct its reasoning strictly based on these ground-truth injected artifacts. Crucially, it must format the output as if it is evaluating the image autonomously, without ever revealing that the ground truth was provided.

The system prompt used for trajectory generation is presented below:

System Prompt for Autonomous-Style Trajectory Synthesis

SYSTEM INSTRUCTION:
You are an advanced Image Quality Assessment Expert. Your task is to generate a perfect, step-by-step reasoning trajectory for an SFT dataset.
HIDDEN DIRECTOR'S KNOWLEDGE (DO NOT REVEAL):
You are secretly provided with the **Target Base Score**, the **Target Final Score**, and the **Ground-Truth Degradation Labels for each crop** (e.g., Crop 1: Clean, Crop 2: Mosaic).
CRITICAL RULE: You must generate the response as if you are evaluating the image entirely from scratch. **NEVER** say "The expert score is..." or "Based on the provided labels...". You must act as if YOU discovered these defects autonomously.
CRITICAL KNOWLEDGE: TARGET DEGRADATIONS
You are specifically looking for 5 types of **MAN-MADE** degradations: Mosaic, JPEG Compression, Posterization, Scanlines/Grid, and Noise/Blur.
Step 1: Global Baseline Establishment
Evaluate the global tech-fidelity. State your initial global assessment clearly, ending with a base score that *exactly matches* the Target Base Score.
Step 2: Local Scrutiny & Mathematical Deduction
Examine the crops.

- **Clean Crops:** If the provided label indicates a crop is clean, declare it clean (Penalty = 0.0) due to natural bokeh/textures.
- **Degraded Crops:** If the provided label indicates a specific degradation (e.g., JPEG Compression), you **MUST strictly use this true degradation type** in your reasoning. Construct a logical justification for a penalty that mathematically bridges the Base Score to the Final Score. (e.g., "However, upon closer inspection, the background reveals severe mosaic blocks, requiring a 2.0 penalty.") **DO NOT invent or hallucinate degradations that are not in the provided labels.**

Output Format
Output ONLY the conversational trajectory (your thoughts, your tool calls, and the final score). Do not output JSON metadata here, just the natural text for the SFT assistant response.

By utilizing this hidden-state anchoring, the resulting trajectories perfectly align with human consensus while maintaining the linguistic appearance of independent, deductive reasoning, making them ideal for Supervised Fine-Tuning.

B Inference Pipeline Implementation Details

To evaluate Q-Probe's capability to detect fine-grained degradations in high-resolution images, we developed a specialized inference pipeline based on the `vLLM` framework. This pipeline supports multi-turn visual reasoning and dynamic tool invocation, allowing the model to actively "zoom in" on suspicious regions. The implementation details are described below.

B.1 Engine Configuration and Resolution Management

We deployed the model using the `vLLM` engine to maximize inference throughput and memory efficiency. To accommodate the ultra-high resolution of Vista-Bench images (often exceeding 4K resolution), we customized the multi-modal processor configuration. Specifically, we set the `min_pixels` to $256 \times 28 \times 28$ and `max_pixels` to $2560 \times 28 \times 28$. This configuration ensures that the model preserves sufficient high-frequency details during the initial global encoding phase, preventing the loss of small artifacts (e.g., subtle mosaic blocks or grid lines). The inference was



Figure 5: We utilize the Wavelet Transform to extract high-frequency components, serving as spatial guidance to concentrate the degradation injection within semantically rich foreground regions.

conducted using `bfloat16` precision with a tensor parallel size of 2, distributed across NVIDIA RTX 4090 GPUs. We also enabled `enable_prefix_caching` to accelerate the processing of shared system prompts across multiple turns.

B.2 The Agentic Reasoning Loop

Unlike traditional static inference, our pipeline implements a dynamic, multi-turn "Observe-Reason-Act" loop. The process is defined as follows:

1. **Initialization**: The session begins with the system prompt defining the five target degradation types (Mosaic, JPEG, Posterize, Scanlines, Grid) and the "Zero-Penalty" rule. The global view of the high-resolution image is fed into the model.
2. **Global Analysis and Hypothesis Generation**: The model first evaluates the global technical fidelity. If it detects potential local degradations that are ambiguous in the global view, it generates a special XML tag `<tool_call>` containing the normalized bounding box coordinates for a closer inspection.
3. **Dynamic Tool Execution**: A regular expression parser monitors the model's output stream. Upon detecting a `crop_image_normalized` function call, the pipeline pauses text generation and executes the cropping logic:

$$I_{crop} = \text{Crop}(I_{global}, [x_1, y_1, x_2, y_2]) \tag{11}$$

   The cropped region is extracted from the original high-resolution source (not the resized tokens) to ensure maximum clarity.
4. **Contextual Update**: The cropped image is encoded into Base64 format and appended to the conversation history as a new user message. The model then resumes generation, analyzing the new visual evidence to confirm or dismiss the degradation hypothesis.
5. **Termination**: This loop continues for a maximum of $T = 6$ turns. The process terminates early if the model ceases to call tools or provides a final score.

B.3 State Management and Reproducibility

To ensure reproducibility and facilitate error analysis, the pipeline implements a robust state management system. The full conversation history, including text reasoning, tool arguments, and the sequence of viewed crops, is serialized into a JSON format. Image data within the history is dynamically managed: during active inference, images are passed as Base64 strings; for archival, they are saved as local paths (e.g., `{image_id}_turn{i}_crop_{j}.png`) to maintain a lightweight log file. This structured logging allows for the quantitative analysis of the model's "looking" behavior (e.g., IoU of crops) as discussed in the main paper.

B.4 Prompt Engineering

To guide the agent in performing rigorous technical quality assessment, we designed a structured System Prompt. This prompt serves four key functions: Role Assignment, Defect Taxonomy, Workflow Enforcement, and Output Standardization.

The exact system prompt used in our inference pipeline is shown below.

System Prompt for Agentic IQA Inference

You are an expert Image Quality Assessment (IQA) assistant. Your task is to analyze the image for technical degradations, specifically focusing on digital artifacts like:

1. **Mosaic/Pixelation**: Blocky structures.
2. **JPEG Compression**: Ringing artifacts and blocking.
3. **Posterization**: Color banding or loss of gradient.
4. **Scanlines/Grid**: Artificial line overlays.
5. **Noise/Blur**: General degradation (common in KADID10K).

Inference Steps:

1. **Global Analysis**: Briefly describe the overall content and perceived quality. Is it sharp? Is it noisy?
2. **Defect Hunting**: Actively look for the artifacts listed above.
3. **Zoom/Crop Strategy**: If you suspect an artifact in a specific area (e.g., background textures, edges), use the `crop_image_normalized` tool to inspect it closely.
4. **Final Scoring**: Based on the severity of the defects found:
   - **5.0**: Excellent, no visible defects.
   - **4.0**: Good, minor defects visible only upon close inspection.
   - **3.0**: Fair, defects are annoying but content is visible.
   - **2.0**: Poor, defects significantly affect quality.
   - **1.0**: Bad, extremely degraded.

CRITICAL OUTPUT RULE:
At the very end of your final response, you MUST output the final score enclosed in XML tags, strictly like this: `<score>3.45</score>`.

C Supervised Fine-Tuning Data Construction

To equip the Q-Probe agent with both robust visual perception and reasoning capabilities, we constructed the **Probe-CoT-3K** dataset. The construction involves a rigorous synthesis strategy that pairs images with multi-turn conversation trajectories. Crucially, while the trajectories are synthesized using hidden ground-truth anchors (as described in Sec. A), the final SFT dataset strictly mimics a zero-shot real-world inference scenario to prevent data leakage.

C.1 Data Format and Templating

Each data entry in **Probe-CoT-3K** follows a standardized multi-turn conversation format.

- **System Prompt**: A fixed instruction set defining the agent's persona and the degradation taxonomy.
- **Clean User Input**: The user provides *only* the image and a generic request (e.g., "Analyze the quality of this image"). **No ground truth scores are exposed to the model during SFT.**
- **Autonomous Assistant Trajectory**: The assistant responds with a global assessment (establishing a base score), optionally invokes the `crop_image_normalized` tool, analyzes the returned crops, and mathematically deduces the final score. Because of our hidden-prompting generation strategy, this reasoning autonomously leads to the correct human-annotated score.

Table 6: The final SFT data template for the High-Res / Agentic Path. The Assistant's response contains the autonomous reasoning trajectory that arrives at the correct score.

(a) High-Res / Agentic Path (With Crop Tool Usage)

```
{
  "image_id": "HighRes_Sample_001",
  "message_list": [
    {
      "role": "system", "content": "...[Standard IQA System Prompt]..."
    },
    {
      "role": "user",
      "content": [
        {"type": "image", "image": "<path_to_full_image>"},
        {"type": "text", "text": "Analyze the quality of this image."}
        // Notice: NO scores provided here!
      ]
    },
    {
      "role": "assistant",
      "content": "The overall image structure is well-composed and clear, which warrants
      an initial base score of 4.5. However, I notice some complex textures in the
     background that might contain compression artifacts. I need to inspect them closer.
     <tool_call>{\"name\":\"crop_image_normalized\",...}</tool_call>"
    },
    {
      "role": "user",
      "content": [
        {"type": "image", "image": "<path_to_crop_1>"},
        {"type": "text", "text": "Here are the cropped images..."}
      ]
    },
    {
      "role": "assistant",
      "content": "Upon closer inspection of Crop 1, the background is not just natural
     bokeh; it contains distinct 8x8 'Mosaic' blocking artifacts. This artificial
     degradation is severe enough to justify a 2.0 penalty from the base score. Final
     Score: <score>2.5</score>"
    }
  ]
}
```

C.2 Trajectory Diversity and Quality Control

A key challenge in SFT for agentic tasks is avoiding "shortcut learning," where the model blindly predicts low scores whenever a crop tool is used. To mitigate this, our dataset includes a balanced mix of three trajectory types:

1. **Positive Trajectories**: The agent zooms in on complex textures (e.g., foliage, fabric) and correctly identifies them as *clean*, applying zero penalty.

2. **Negative Trajectories**: The agent detects actual degradations (e.g., JPEG blocks) in the crops and applies penalties that logically match the expert score drop.

3. **Global-Only Trajectories**: For obviously low-resolution or low-quality images, the agent decides *not* to zoom in and justifies the score directly based on the global view.

Tables 6 and 7 illustrate the final SFT data structures. Notice how the User prompt is completely devoid of target scores, forcing the Q-Probe model to learn the actual mapping from visual features to the complex reasoning trajectory.

Table 7: The final SFT data template for the Low-Res / Direct Path (No Crop Tool Usage).

(b) Low-Res / Direct Path (No Crop Tool Usage)

```
{
  "image_id": "LowRes_Sample_002",
  "message_list": [
    {
      "role": "system", "content": "...[Standard IQA System Prompt]..."
    },
    {
      "role": "user",
      "content": [
        {"type": "image", "image": "<path_to_full_image>"},
        {"type": "text", "text": "Analyze the quality of this image."}
      ]
    },
    {
      "role": "assistant",
      "content": "From the global view, it is immediately apparent that the image
     suffers from heavy, uniform sensor noise and severe blurring across all regions. The
      degradation is global, so no local zooming is required. The technical fidelity is
     extremely poor. \n\n<score>1.2</score>"
    }
  ]
}
```

D More Details about Vista-Bench

The significant performance superiority of Q-Probe over existing global-view IQA methods on Vista-Bench can be attributed to the fundamental paradigm shift from passive observation to active **Context-Aware Agentic Probing**. While traditional methods rely on coarse-grained global views that often miss high-frequency details, our framework introduces a nuanced evaluation capability centered on two key advantages:

1. Fine-Grained Perception vs. Coarse Global Scoring: Existing MLLMs typically process images from global perspective, causing subtle degradations in high-resolution scenarios (e.g., $4K$ inputs) to vanish during encoding. These models effectively "guess" the quality based on a low-fidelity global impression. In contrast, Q-Probe utilizes an agentic zooming mechanism to actively localize and inspect microscopic degradations. Crucially, our method does not merely detect defects; it evaluates the *perceptual impact* of the degradation relative to the **semantic importance** of the region. By integrating the severity of local defects with the saliency of the content (e.g., applying severe penalties for artifacts on a foreground subject while adhering to a "Zero-Penalty" rule for natural background bokeh), Q-Probe aligns its scoring logic with human visual attention, a capability largely absent in global-view baselines.

2. Generalization to Unseen Degradations: To rigorously validate that the performance gains stem from genuine visual understanding rather than distribution overfitting, we introduced five degradation types that were never used during training: fast fading, color shift, motion blur, salt-and-pepper noise, and ringing artifacts. We used half of the test set, retrieved the corresponding pristine images, and re-injected these degradations with random spatial injection rather than the wavelet-guided injection used in training; each type accounts for an equal proportion. Compared with in-distribution wavelet-injected test samples, Q-Probe's SRCC and PLCC dropped by only 0.01 and 0.008, respectively.

E Limitations and Bad Cases Study

To comprehensively evaluate the boundaries of Q-Probe, we analyze both successful applications and failure modes. As demonstrated in Figure 6, under standard high-resolution scenarios, our agent exhibits superior performance. The context-aware cropping strategy successfully distinguishes

between artificial degradations and natural depth-of-field effects (bokeh). By accurately localizing low-quality regions while ignoring non-subject blurry backgrounds, the model adheres strictly to the "Zero-Penalty" rule, ensuring fair and accurate quality assessment.

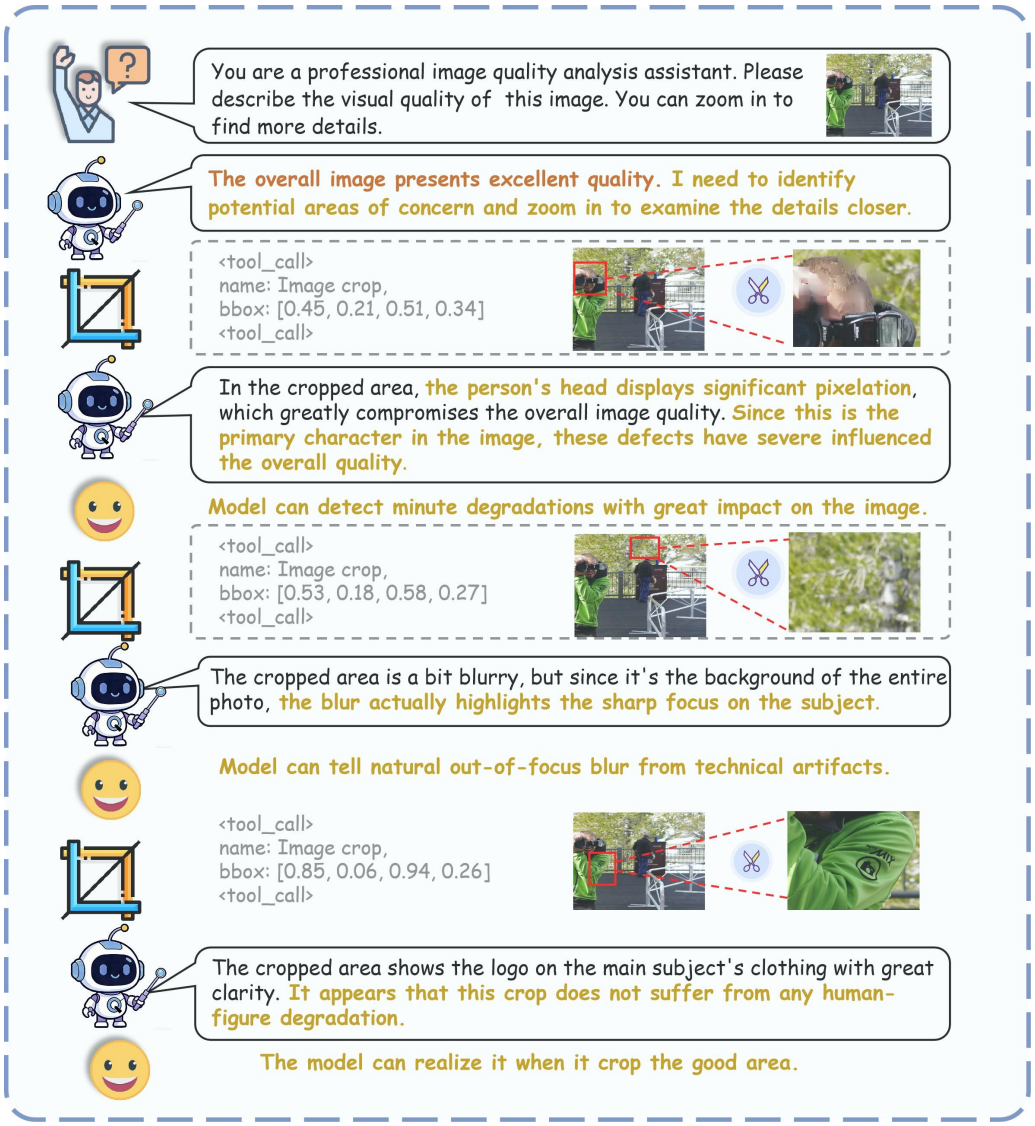


Figure 6: **Success Case:** The agent effectively identifies the main subject and applies the "Zero-Penalty" rule to the natural bokeh in the background, demonstrating the effectiveness of our context-aware cropping strategy in standard scenarios.

However, robustness challenges arise in scenes exhibiting complex semantic dependencies. The first limitation involves **semantic entanglement between foreground and background**. As illustrated in Figure 7, while the model ignores standalone blurry backgrounds, it struggles when distant backgrounds contain textures strongly correlated with the subject's narrative (e.g., a complex street scene behind a vehicle).

In these cases, the agent's attention mechanism becomes confused by the high correlation. Instead of focusing solely on the subject, it erroneously interprets complex background textures as "artificial noise" or "processing artifacts." This leads to **erroneous cropping** on irrelevant areas and subsequently applies **unjustified penalties**, indicating a need for a more sophisticated semantic understanding of scene composition.

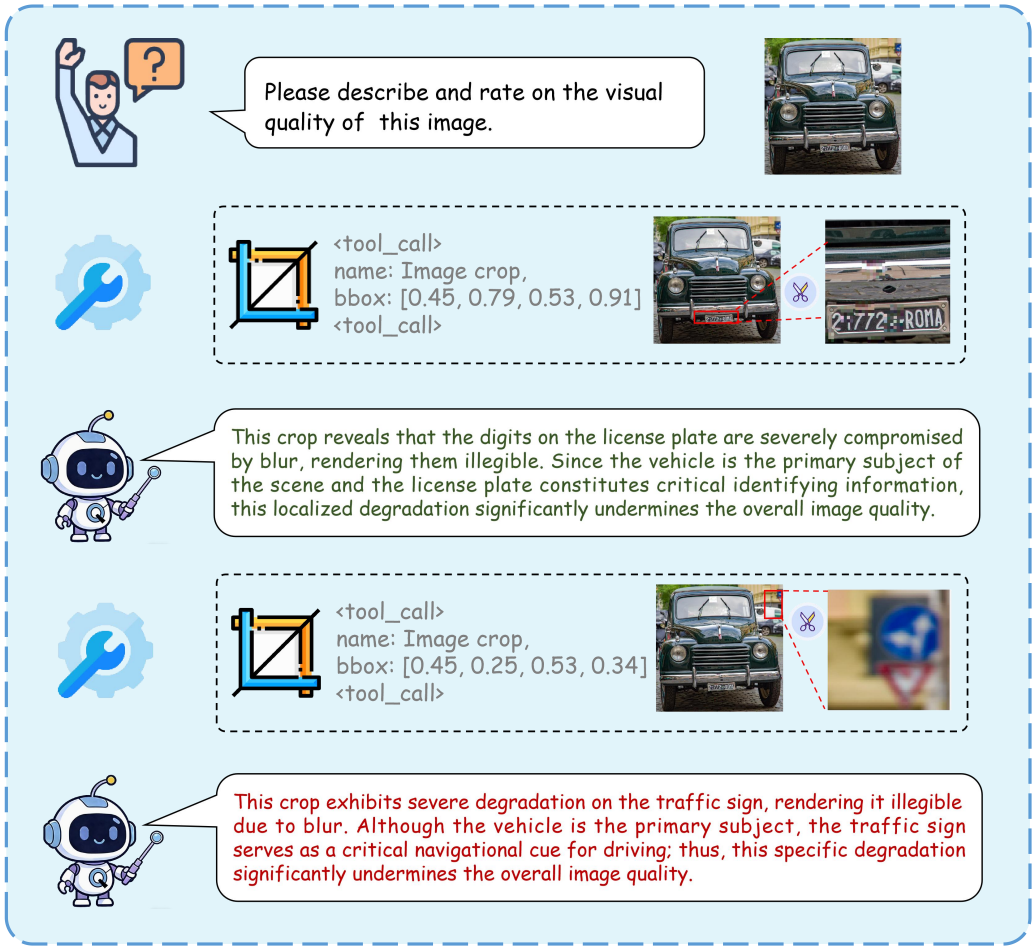


Figure 7: **Failure Mode I (Semantic Entanglement):** While the agent can distinguish bokeh, its discriminative capability weakens when the background is strongly correlated with the subject (e.g., a car on a street). The model may misinterpret complex background textures as artifacts, leading to erroneous cropping.

The second limitation pertains to the **insensitivity to spatial context differences**, particularly between indoor and outdoor environments. As shown in Figure 8, our agent is prone to uniformly assigning minimal penalties to degradations in all background regions. This generalization is flawed because indoor backgrounds cannot be treated equivalently to vast outdoor settings.

Due to the **confined spatial nature** of indoor scenes, background artifacts remain perceptually significant compared to distant outdoor backgrounds. Consequently, while the penalty for indoor background degradation should not be as severe as foreground defects, it must not be negligible either. Our current model fails to apply this necessary **moderate penalty**, often underestimating the impact of background noise in spatially constrained environments.

Finally, a third limitation arises from **ambiguity in subject definition**, leading to a **salience bias** during crop selection. As depicted in Figure 9, the agent incorrectly identifies the distant house as the primary subject because it is the sole distinct architectural element, while relegating the river to the background.

This classification ignores the fact that the river occupies nearly half the frame and dominates the visual experience. In the context of Image Quality Assessment (IQA), degradations on such a prominent surface significantly impair perceptual quality. By classifying the river as a non-salient background, the model generates a flawed reasoning trajectory and under-penalizes visible defects, resulting in an inflated quality score.

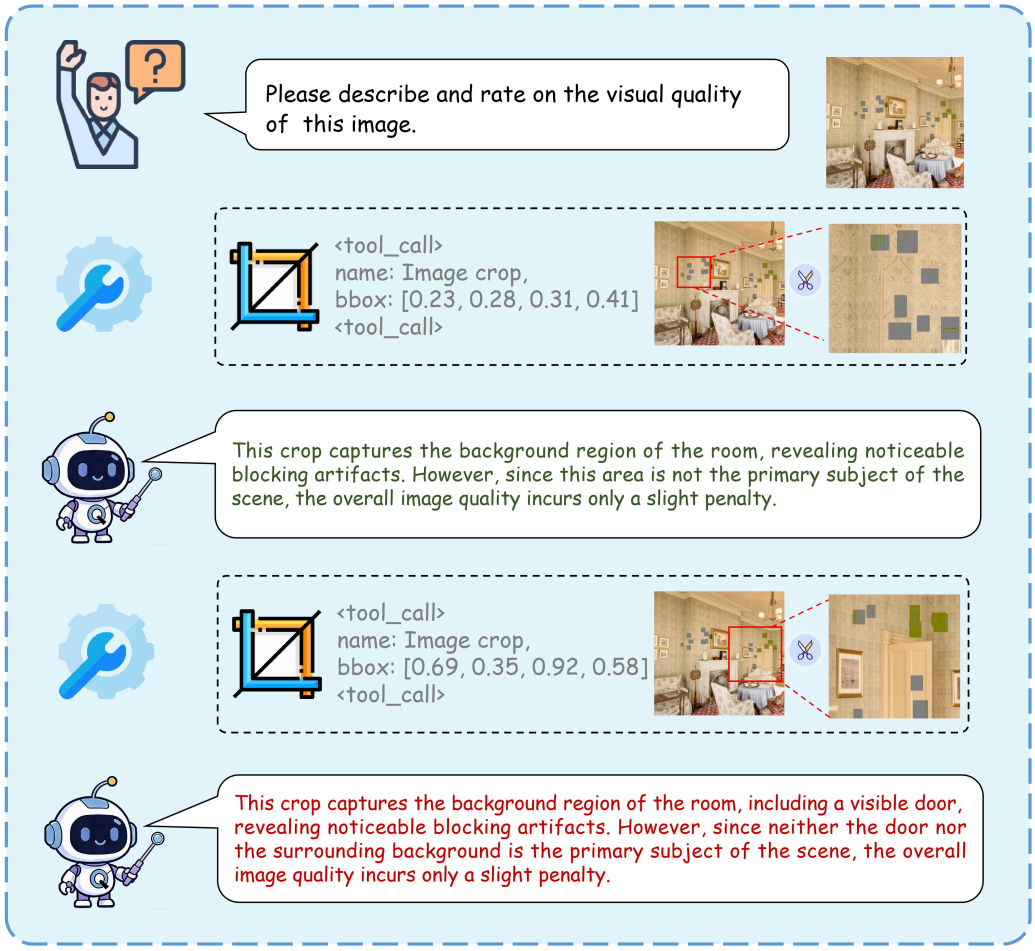


Figure 8: **Failure Mode II (Spatial Context Insensitivity):** The agent incorrectly applies minimal penalties to indoor background degradations. Unlike vast outdoor settings, indoor scenes are spatially confined, meaning background artifacts remain perceptually significant and require a moderate penalty rather than a negligible one.

F Additional Experimental Results and Analyses

This section provides extended experimental results, comprehensive ablations, and further details regarding the human-centric annotation process of our benchmark.

F.1 Exploratory Study on Naive Zoom-in Fine-Tuning

Before arriving at the proposed context-aware cropping strategy, we conducted an exploratory study in which SFT trajectories were constructed exclusively from cropped regions containing injected degradations. This naive zoom-in fine-tuning quickly induced the failure mode illustrated in Fig. 1(c-d): the model associated tool usage with low quality and misinterpreted natural bokeh as degradation. The exploratory result motivated the context-aware cropping strategy and the three-stage curriculum; we report it here as an empirical motivation rather than as a claim about prior work.

F.2 Prevalence of Local Degradations in Real High-Resolution Scenes

To assess whether localized degradation is practically relevant beyond the synthetic Vista-Bench construction, we conducted a preliminary prevalence study. We generated 1,000 high-resolution images with Nano Banana using the prompt "Please generate realistic and trustworthy images."



Figure 9: **Failure Mode III (Salience Bias):** Ambiguity in scene composition introduces a bias: the agent misidentifies the small distant house as the main subject due to object uniqueness, while dismissing the dominant river as background. Consequently, the model under-penalizes degradations on the water surface, which are actually critical to perceptual quality.

The same six-expert panel independently labeled each image into three categories. The results are summarized in Table 8.

Table 8: Preliminary prevalence of local defects in 1,000 high-resolution AIGC images.

Category	**Count**	**Percentage**
Overall excellent (no noticeable defects)	586	58.6%
Minor local defects (globally acceptable, locally imperfect)	385	38.5%
Overall poor (global quality compromised)	29	2.9%

A re-examination of existing high-resolution benchmarks gives dataset-dependent estimates: in UHD-IQA, approximately 28% of images contain regional over/under-exposure or local motion blur; in DPED, approximately 22% of smartphone photos exhibit localized artifacts such as lens flare, corner softness, or local noise. These statistics are preliminary and dataset-dependent; they are intended to show that local defects are neither rare nor negligible as resolution increases, not to claim a universal prevalence rate.

F.3 Mitigating Semantic Robustness Bias via Pairwise Optimization

We define "Semantic Robustness Bias" as the phenomenon where IQA models over-prioritize global semantic quality while neglecting critical local defects. Existing ranking-based RL paradigms often employ group-level image comparisons (e.g., VisualQuality-R1 uses a group size of 6), which inadvertently encourages the model to solidify its global assessments at the expense of local sensitivity.

To mitigate this, Q-Probe implements a pairwise contrastive optimization strategy (comparison pair size = 2). This allows the model to undergo more granular dynamic learning, effectively alleviating the global bias induced by standard training paradigms. As shown in Table 9, directly employing Supervised Fine-Tuning (SFT) as Stage 1 achieves competitive or even superior performance on low-resolution datasets dominated by global quality (e.g., SPAQ, KonIQ). However, it forces the model to heavily anchor its judgments on macroscopic semantics, causing it to completely overlook localized degradations and resulting in the lowest performance on the high-resolution Vista-Bench. Similarly, while group-level contrastive RL (e.g., VisualQuality-R1 with a generation group size of 6) slightly improves local sensitivity over SFT, it still prioritizes global consistency over subtle local artifacts. By isolating fine-grained relative differences through pairwise RL, our approach successfully preserves the model's exploratory plasticity for high-resolution details, establishing an optimal, unbiased foundation for the subsequent agentic cropping stage.

Table 9: Impact of Pre-alignment Strategy and Group Size in Stage 1

Model Config	Vista		SPAQ		KADID		KonIQ	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Stage 1 Only (Ours, Pairwise)	**0.352**	**0.404**	0.760	0.780	0.700	0.725	0.650	0.680
Stage 1 Only (VisualQuality-R1)	0.327	0.375	0.792	0.815	**0.735**	**0.758**	0.684	0.712
Stage 1 Only (SFT)	0.312	0.355	**0.805**	**0.822**	0.720	0.741	**0.695**	**0.721**

To further verify this, we sampled 500 images from Vista-Bench containing localized blurring in key regions. VisualQuality-R1 significantly overestimated the quality with an average score of 3.94, whereas Q-Probe yielded an average score of 2.06, which is much closer to the human Ground Truth (GT) average score of 1.82.

F.4 Training Stability: Uncertainty-Aware Alignment vs. Bradley-Terry

Due to the inherent subjectivity of IQA, relying solely on single-sample ranking losses (e.g., the Bradley-Terry model) makes Stage-1 RL susceptible to noise overfitting on ambiguous samples. Q-Probe mitigates this via a K-rollout uncertainty-aware alignment method, utilizing multi-sample variance to downweight gradients for high-uncertainty samples.

Table 10 demonstrates that our approach outperforms the Bradley-Terry model across all datasets. Furthermore, Table 11 tracks the average standard deviation of predicted scores across training steps. While the Bradley-Terry model hits an optimization bottleneck and its variance fluctuates (around 0.15–0.20) as it overfits to subjective noise, Q-Probe demonstrates a continuous and stable reduction in variance, confirming superior training stability.

Table 10: Performance Comparison of RL Modeling Strategies

Model Config	Vista (SRCC)	Vista (PLCC)	SPAQ (SRCC)	SPAQ (PLCC)	KADID (SRCC)	KonIQ (SRCC)
Stage 1 Only (Ours)	**0.352**	**0.404**	**0.760**	**0.780**	**0.700**	**0.650**
Stage 1 Only (Bradley-Terry)	0.337	0.388	0.745	0.762	0.682	0.635

F.5 Localization Precision and SFT Trade-offs

During the Stage-2 SFT phase, we intentionally relax rigid constraints on localization precision to allow the model to learn a mix of degradation and background regions. This prevents the model from developing a spurious causal bias where "cropping implies degradation." As shown in Table 12, this strategy slightly trades off localization accuracy (mIoU) for substantial gains in logical robustness (SRCC/PLCC). Stage-3 RL optimization subsequently recovers and maximizes the mIoU via a dedicated location reward, achieving optimal fine-grained detection.

Table 11: Average Standard Deviation Trajectory During Stage-1 Training. Blue values represent our method (Ours), and gray values represent the Bradley-Terry (BT) baseline.

Training Step	Vista		KADID		KonIQ	
0	0.34	0.35	0.15	0.15	0.25	0.25
50	0.18	0.26	0.08	0.12	0.10	0.22
100	0.11	0.22	0.04	0.10	0.06	0.18
200	**0.08**	0.21	**0.01**	0.11	**0.01**	0.16

Table 12: Trade-off Analysis for Localization Precision. Config A/B do not contain Stage-3.

Config	Training Strategy	mIoU	SRCC	PLCC
Config A	SFT (Strict degradation-only regions)	0.689	0.558	0.627
Config B	SFT (Mix of degradation & background)	0.643	0.762	0.808
Config C	**Full Pipeline (Stage-3 RL Optimization)**	**0.702**	**0.801**	**0.850**

F.6 Reward Parameter Analysis in Decoupled Post-RL

We conducted ablations on the reward parameters in Stage 3: α (accuracy), β (localization), and γ (format). The temperature parameter τ in the accuracy reward was empirically set to 1.0. As demonstrated in Table 13, lowering the weight of either the accuracy or localization reward strictly diminishes performance. The format reward has a marginal influence, while accuracy and localization rewards are the primary drivers of Q-Probe's precision.

Table 13: Parameter Ablation for Decoupled Rewards

α (Accuracy)	β (Localization)	γ (Format)	SRCC	PLCC
1.0	**1.0**	**1.0**	**0.801**	**0.850**
0.5	1.0	1.0	0.768	0.794
1.0	0.5	1.0	0.782	0.796
1.0	1.0	0.5	0.799	0.851

F.7 Rigorous Human Annotation and Ground Truth Validation

To guarantee the absolute reliability of the Vista-Bench evaluation, we emphasize that the final Ground Truth (GT) labels are derived entirely from rigorous manual annotation from scratch.

Given the inherently subjective nature of Image Quality Assessment, we eschewed any reliance on automated systems or large language models for scoring. Instead, we established a dedicated panel of domain experts to conduct systematic, cross-verified evaluations of every single image in the dataset. A sample was strictly discarded if it triggered either:

1. **High Scoring Ambiguity:** The variance among the human evaluators' independently assigned scores was too large, indicating a lack of human consensus.
2. **Semantic Disagreement:** The experts failed to reach a consensus on the semantic classification of the cropped regions (e.g., disagreeing on whether a region constituted Primary Foreground or Background).

Through this purely manual, expert-driven annotation process, all data splits are fully vetted and approved by human evaluators, successfully capturing the plurality of human opinions while ensuring the highest standard of data fidelity.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Our abstract and introduction clearly outline our contributions, including the Vista-Bench benchmark, the Q-Probe framework, and the three-stage training paradigm, all of which are extensively supported by our methodology and experiments sections.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We include a dedicated "Limitations and Bad Cases Study" section (Appendix E) that discusses and visualizes semantic entanglement, spatial context insensitivity, and salience bias.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Justification: Our work is primarily empirical and methodological, focusing on an agentic IQA framework and a new dataset benchmark, and we do not present formal theoretical proofs.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: We provide comprehensive details on our experimental setup, including base models (Qwen-2.5-VL-7B), hyperparameter settings (e.g., GRPO generation number, learning rate), evaluation metrics, and system prompts used for trajectory generation and inference.

Guidelines:

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

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

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

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

   Answer: [Yes]

   Justification: We have included the code and dataset in the supplementary materials. Furthermore, we provide detailed descriptions in both the main text and the appendix on how the code is implemented and how the dataset is constructed (including the ground-truth scores and Chain-of-Thought (CoT) reasoning).

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: In Section 4.1, we specify key training parameters, including the AdamW optimizer, learning rate, GRPO generation numbers, clip threshold, and batch/crop sizing.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We track the average standard deviation of predicted scores across multiple inference runs over various training steps (Figure 4, Table 9) to empirically demonstrate our training stability and variance reduction.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We explicitly state that our experiments were conducted on 8 NVIDIA A100 (80GB) GPUs and that we deployed the inference pipeline across NVIDIA RTX 4090 GPUs.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Our research focuses on advancing image quality assessment methodology safely and does not pose any obvious ethical violations, malicious use cases, or involve harmful data processing.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [N/A]

Justification: We focus intensely on the technical methodology of high-resolution IQA and this won't have potential positive societal impacts or negative societal impacts.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: Our proposed Q-Probe framework and Vista-Bench evaluate technical image fidelity and distortions, which present a very low risk for misuse compared to generative models or sensitive demographic datasets.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: We appropriately credit and cite existing datasets (HR-Bench 4K, UHD-IQA, KADID-10k) and foundation models (Qwen-2.5-VL-7B), adhering to academic standards, though we do not print specific license texts inline.

Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: We extensively detail our Vista-Bench construction, including our wavelet-based degradation methodology, annotation procedures, and SFT format guidelines (Probe-CoT-3K), in both the main text and appendices.

Guidelines:

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

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

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

Answer: [Yes]

Justification: We detail the human annotation process and instructions in Appendix A.2. Furthermore, all six specialists (domain experts and PhD researchers) were fairly compensated for their annotation workload through our laboratory's research funding at standard institutional rates, adhering to ethical labor practices.

Guidelines:

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

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

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

    Answer: [N/A]

    Justification: Our "human subjects" involvement was restricted to expert evaluators grading the technical quality of images, which is standard data annotation lacking the risk or intervention that typically requires IRB oversight.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: We explicitly state our use of an MLLM (Qwen-2.5-VL-7B) as the foundational architecture for our Q-Probe framework and for autonomous trajectory synthesis within our methodology section.

    Guidelines:

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