

---

# Rethinking Image Quality Assessment through the Lens of Task Utility in Embodied Settings

---

Jirong Zha<sup>\*1</sup> Yemin Wang<sup>\*2</sup> Xiangmin Yi<sup>1</sup> Siqi Peng<sup>3</sup> Yingfeng Chen<sup>4</sup> Chen Gao<sup>5</sup> Xinlei Chen<sup>1</sup>

## Abstract

Image quality assessment (IQA) underpins embodied imaging pipelines by judging whether visual quality satisfies downstream tasks, yet most methods learn task-agnostic scores aligned with generic human ratings on static benchmarks. This objective mismatches the embodied and interactive settings, where image adequacy depends on task goals, context, and action requirements that shape an agent’s decisions. We argue that **IQA should shift from score regression to goal-conditioned judgment defined by the utility of embodied tasks**. Such utility-aware assessment demands models with strong reasoning, grounding, and tool-use capabilities, as enabled by multimodal large language models (MLLMs) agent. We advocate rethinking IQA from the perspective of embodied task utility and outline benchmarks, evaluation protocols, and research directions for developing MLLM-based embodied IQA agents.

## 1. Introduction

Image quality assessment (IQA) is a fundamental component of computer vision systems, addressing the critical question of whether visual signals are usable for various downstream applications (Zhang et al., 2023; Li et al., 2025c; Zha et al., 2025). IQA encompasses the quantitative and qualitative evaluation of visual content to determine its adequacy for specific purposes, ranging from technical fidelity metrics to perceptual quality judgments (Zhou et al., 2025). For example, IQA may determine whether a drone’s aerial image supports reliable target localization, whether an autonomous vehicle’s camera frame enables traffic sign

<sup>1</sup>Shenzhen International Graduate School, Tsinghua University, Shenzhen, China <sup>2</sup>Xiamen University, Xiamen, China <sup>3</sup>Central South University, Changsha, China <sup>4</sup>Weiyang College, Tsinghua University, Beijing, China. <sup>5</sup>BNRist, Tsinghua University, Beijing, China. Correspondence to: Xinlei Chen <chen.xinlei@sz.tsinghua.edu.cn>, Chen Gao <chgao96@gmail.com>.

Preprint. April 15, 2026.

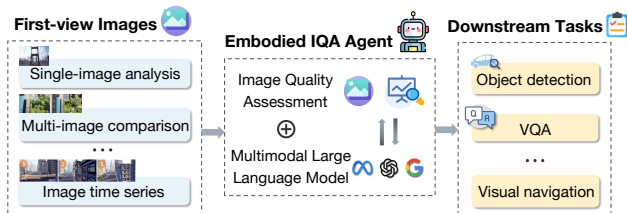


Figure 1. By coupling MLLM reasoning with image quality assessment, the MLLM-based embodied IQA agent evaluates quality from a first-person perspective, incorporating task context, environmental constraints, and goal-driven decision requirements to support autonomous decision-making.

recognition and pedestrian detection, or whether a robotic arm’s visual input is sufficient for precise grasping and manipulation. IQA methods inform image acquisition and enhancement pipelines, guide compression and streaming algorithms, support frame selection for video analysis, and benchmark restoration and generation models (Wang et al., 2004; Hore & Ziou, 2010; Saad et al., 2012; Mittal et al., 2012a; Ma et al., 2025). As embodied intelligence advances, IQA plays a pivotal role in embodied perception, where the visual signal in vision–language–action (VLA) systems directly shapes perception, planning, and control, making image adequacy fundamental to reliable robotic behavior.

Existing IQA research has evolved through several phases: early work focused on full-reference metrics (e.g., PSNR (Hore & Ziou, 2010) and SSIM (Wang et al., 2004)) that compare images against pristine references; reduced-reference methods that operate with partial information; and no-reference approaches (Saad et al., 2012; Mittal et al., 2012a) that assess quality without any reference. Deep learning models (Bosse et al., 2017) have achieved state-of-the-art performance by learning quality representations from large-scale datasets annotated with human opinion scores. Recent comprehensive surveys (Ma et al., 2025; Li et al., 2025a) provide detailed analysis of the field’s evolution and current state. These developments have established IQA as a mature field with well-established benchmarks and evaluation protocols.

Despite recent advances, traditional scalar IQA metrics remain inadequate for embodied settings, where image quality should be evaluated by its downstream utility and in-

interpretability for intelligent agents (Zhang et al., 2024a; Zhu et al., 2025). Existing IQA methods are largely task-agnostic, predicting scalar scores or discrete labels based on general perceptual principles without accounting for the intended use of the image. Their evaluation typically relies on correlation with human opinion scores on static benchmarks, which fails to reflect real-world utility in goal-driven scenarios. This limitation is particularly evident in embodied and interactive systems, where image quality is inherently task-dependent and must be assessed with respect to mission objectives, environmental constraints, and downstream decision-making.

**Dilemma of Traditional IQA.** In embodied settings, traditional scoring-based IQA methods break down when quality requirements are task-dependent, when task-relevant regions occupy only a subset of the image (Chen et al., 2024a; Song et al., 2025), and when semantic coverage outweighs pixel-level fidelity (Kerkouri et al., 2025). Consider UAV infrastructure inspection: motion blur in background regions (e.g., sky) is largely irrelevant, whereas blur on structural surfaces renders images unusable for crack detection. Conversely, a visually pristine image that fails to capture the target surface may score highly yet provide no utility. A similar issue arises in robotic manipulation, where visual inputs support pose estimation, tracking, and grasp planning; quality should be judged by the preservation of task-critical cues (e.g., edges, keypoints, depth), with distortions being tolerable or harmful depending on task requirements. These cases expose a fundamental misalignment: a single task-agnostic score, optimized for correlation with human ratings, cannot capture the diverse and task-specific quality demands of embodied downstream applications.

**Supporting position:** IQA should be reconceptualized as a judgmental process rather than a regression task. Quality assessment should consider *why* an image is useful for a specific goal, not only how it deviates from an ideal reference. This requires reasoning about task relevance, semantic content, and the relationship between visual properties and downstream performance. When a human operator evaluates a robot’s camera feed, they implicitly condition on the current objective (e.g., grasping, inspection, navigation), tolerate some degradations, and demand others to be resolved; an embodied IQA agent should do the same, as illustrated in Figure 2.

**The Turning Point.** IQA stands at an inflection point driven by two converging technological trends. First, the proliferation of *embodied AI agents*, autonomous systems that perceive and act in physical environments, has created new demands for context-aware quality assessment. These agents include unmanned aerial vehicles (UAVs) performing infras-

tructure inspection (Morgenthal & Hallermann, 2014), mobile robots engaged in navigation and manipulation (Zhang et al., 2024a), and augmented reality (AR) systems providing real-time visual assistance (Lv et al., 2025). Second, the rapid advancement of *multimodal large language models* (MLLMs) (Liu et al., 2023) has demonstrated unprecedented capabilities in combining visual perception with language-conditioned reasoning, enabling systems that can understand context, generate explanations, and adapt to diverse tasks.

These two technological trends converge to create a transformative opportunity: MLLMs can underpin embodied IQA agents that assess image quality from a first-person, goal- and context-conditioned perspective, as shown in Figure 1. MLLMs embodied agents create an opportunity to rethink IQA as *goal-conditioned judgment* rather than task-agnostic regression, because their multimodal reasoning, contextual understanding, and tool-use capabilities enable them to assess image quality with respect to explicit goals, task requirements, and environmental context rather than relying on generic perceptual scores. Recent work shows that MLLMs can express quality judgments in natural language (Wu et al., 2023; Huang et al., 2024b), support comparative evaluation (Zhu et al., 2024; Wu et al., 2024b), and provide localized rationales (You et al., 2024; Chen et al., 2024b). The missing step is to integrate IQA into the embodied loop, conditioning judgments and explanations on specific perception task and context so they can drive actionable decisions under changing goals and constraints.

**Primary position:** *IQA should embrace MLLM-based embodied agents, reframing image quality assessment as goal-conditioned, explainable, and adaptive judgment that is evaluated by task utility rather than by correlation with task-agnostic scores.*

**Why This Position Paper?** As AI systems transition from passive predictors to agents that perceive and act, “quality” should be defined relative to goals, constraints, and actions. Existing IQA objectives and benchmarks do not capture this dependency, and they provide little support for explanation, interaction, and policy-level decisions (e.g., retake, reposition, or proceed). The recent maturation of MLLMs, combined with the growing deployment of embodied systems, makes it timely to articulate a principled research agenda for embodied IQA agent that bridges the gap between general quality assessment and task-specific utility. This position paper reveals a critical gap: while MLLM-based IQA has shown promise, current approaches treat IQA as an isolated task rather than an integral component of embodied perception and decision-making. We advocate for a fundamental reframing that integrates IQA into the embodied loop, enabling agents to assess quality from a first-person perspective, explain their judgments, and

adapt to changing mission requirements.

**Contributions.** This position paper makes three contributions: (1) we analyze failure modes of scoring-based IQA in embodied settings and motivate a task-utility view of quality; (2) we review MLLM-based IQA across single-image assessment, comparative evaluation, and grounded analysis, highlighting gaps for embodied use; and (3) we propose a framework and provide research directions for datasets, objectives, evaluation, and deployment toward developing embodied IQA agents.

## 2. Background: MLLM-based IQA and Related Work

Recent work has begun to recast IQA using multimodal foundation models, offering richer outputs than scalar scores. We summarize this line of work along three themes: (1) single-image assessment, (2) comparative and relational evaluation, and (3) grounded quality analysis. While these directions improve interpretability and flexibility, they still rarely model the *task-conditioned* nature of quality required by embodied agents. Table 1<sup>1</sup> highlights the gap between current MLLM-based IQA and embodied requirements.

### 2.1. Single-Image IQA with MLLMs

Early work leveraged pre-trained vision-language models (VLMs) for IQA, establishing foundations for MLLM-based quality assessment. CLIP-IQA (Wang et al., 2023) demonstrates that CLIP embeddings can assess quality without task-specific training, achieving reasonable performance on general quality assessment. This zero-shot approach reveals that VLMs capture quality-relevant features.

Q-Instruct (Wu et al., 2024a) moves beyond zero-shot prompting by constructing instruction-tuning datasets that teach MLLMs to perform low-level visual quality judgments through natural language reasoning. By aligning quality assessment with instruction-following capabilities, it demonstrates that instruction tuning substantially improves MLLM performance on IQA-related tasks.

Q-ALIGN (Wu et al., 2023) further advances this line of work by replacing continuous quality scores with discrete, text-defined quality levels that better reflect how human raters perform subjective evaluation (e.g., “excellent,” “good,” “fair”). Framing IQA as a classification problem rather than regression, Q-ALIGN achieves state-of-the-art results on KonIQ-10K (Lin et al., 2018).

Other recent work includes PromptIQA (Chen et al., 2024c), which uses prompts to boost performance and generalization, and multi-modal prompt learning approaches (Pan

et al., 2024) that enhance blind image quality assessment. Teaching-based methods (Zhang et al., 2025a; You et al., 2025a) explore how to train LMMs for image quality scoring and interpreting. Next Token approaches (Li et al., 2025b) demonstrate realistic image quality and aesthetic scoring with MLLMs. Align-IQA (Yang et al., 2024) addresses diverse human preferences through customizable guidance, enabling more flexible quality assessment aligned with different user needs. A comprehensive study (Wu et al., 2024c) systematically investigates psychophysical experimental designs (single-stimulus, double-stimulus, and multiple-stimulus methods) combined with prompting strategies (standard, in-context, and chain-of-thought) for MLLM-based IQA, revealing that while GPT-4V provides reasonable quality assessment, it struggles with fine-grained quality discrimination and multi-image comparison tasks.

Despite these advances, these methods treat IQA as an isolated task without considering task context. CLIP-IQA provides no task-specific reasoning, Q-Instruct focuses on general quality improvements, and Q-ALIGN, while more interpretable, still produces task-agnostic assessments. They do not integrate with embodied decision-making pipelines where quality judgments should inform actions and adapt to changing mission requirements. The gap between general quality assessment and task-aware judgment represents a fundamental challenge that these methods do not address.

### 2.2. Comparative and Relational IQA

Beyond single-image IQA, comparative and relational methods align quality assessment with human judgment by enabling relative, context-aware reasoning across multiple images. Compare2Score (Zhu et al., 2024) uses adaptive soft comparison to estimate pairwise preferences. Q-Insight (Li et al., 2025c) employs group relative policy optimization for ranked image sets. Co-Instruct (Wu et al., 2024b) enables open-ended visual quality comparison and introduces MICBench, a benchmark for multi-image comparison tasks.

The advances in comparative assessment enable richer judgments than scalar scores but remain detached from task context. Asking “which image is better?” is more informative than assigning a single score, yet asking “which image better supports obstacle avoidance?” is far more meaningful for embodied applications. Bridging this gap between general and task-aware comparison is a key opportunity for embodied IQA.

### 2.3. Grounded Quality Assessment

Recent quality assessment work incorporates spatial grounding, recognizing that quality issues are often localized. Q-Ground (Chen et al., 2024b) detects and describes local distortions via multi-scale feature learning and introduces the QGround-100K dataset, while DepictQA (You et al.,

<sup>1</sup>Please find the detailed concept explanations in Appendix H.

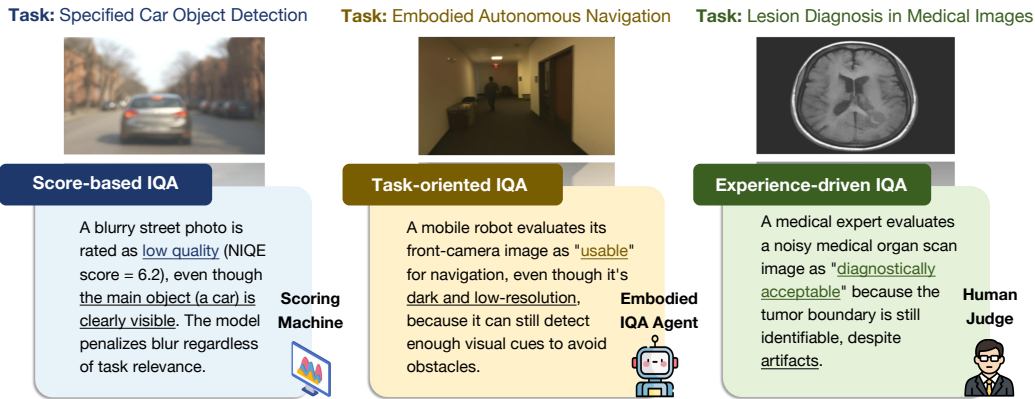


Figure 2. Illustration of low-quality yet high-utility image cases for different downstream tasks, showing how score-based IQA, embodied IQA agents, and human experts prioritize different criteria. While traditional IQA reduces quality to a single score, embodied agents and human experts focus on task-relevant regions and utility. This contrast highlights the need to rethink IQA as goal-conditioned, task-aware judgment rather than task-agnostic scoring.

Table 1. Capability comparison of existing IQA approaches with embodied IQA agents.

Capability	Traditional IQA Scoring	MLLM-based IQA Reasoning	Embodied IQA Agent
Numerical assessment	✓	✓	✓
Task awareness	×	~ (if prompted)	✓
Explanation	×	✓	✓
Task-grounded explanation	×	×	✓
Context adaptation	×	~ (if fine-tuned)	✓
Embodied integration	×	×	✓

Note: ✓ indicates full support; × indicates lack of support; ~ indicates partial support.

2024) generates descriptive quality comparisons across image pairs.

However, these explanations are not task-grounded: they identify what is wrong, but not why it matters for a specific task. Q-Ground can localize blur but cannot assess its task relevance (e.g., structural surfaces vs. sky for crack detection). DepictQA provides descriptive comparisons but cannot relate them to task requirements. As a result, they offer limited actionable guidance for embodied systems where explanations need to inform decisions.

**Extended position:** Our position extends the advances of MLLM-based IQA by calling for *task-aware grounding*: explanations should not just describe quality degradations but relate them to task utility. The transition from “motion blur in the central region” (descriptive) to “motion blur obscures the target structure needed for navigation” (task-grounded) represents a fundamental shift.

### 3. The Necessity of Embodied IQA

#### 3.1. Empirical Motivation: Typical Embodied Scenarios

We highlight concrete settings where “quality” is inherently conditional and a task-agnostic score can be misleading. Each illustrates a different reason that embodied systems

need goal-conditioned judgment rather than a single global score, supported by real-world applications, empirical evidence, and references to actual deployment scenarios <sup>2</sup>.

**Robotic perception and manipulation.** In robot manipulation and navigation, camera inputs are consumed by downstream modules (e.g., pose estimation, tracking, or grasp planning). Quality is therefore tied to whether task-relevant cues (edges, keypoints, depth cues) are preserved under the current objective (Zhang et al., 2024a). Empirical studies on embodied IQA (Zhang et al., 2024a) suggest that noise or compression may be tolerable for coarse navigation but harmful for fine manipulation tasks requiring precise geometric information. Large-scale robot perception datasets further emphasize the diversity of real-world conditions (lighting, motion, sensor artifacts) that induce such task- and context-dependent requirements (Nguyen et al., 2024). Conversely, “clean” imagery can still be inadequate if lighting or viewpoint obscures critical geometry needed for object recognition or pose estimation. For instance, in robotic grasping, an image with high overall quality but poor lighting on the target object’s edges may fail to support reliable

<sup>2</sup>The presented examples are grounded in existing literature (Morgenthal & Hallermann, 2014; Zhang et al., 2024a; Lv et al., 2025; Nguyen et al., 2024; Abedin et al., 2020) and illustrate why task-agnostic scoring fails in embodied settings.

grasp planning, while a slightly noisier image with adequate edge visibility may enable successful manipulation. This demonstrates how task-specific requirements (edge detection for grasping) override general quality metrics.

**UAV infrastructure inspection.** In UAV-based structural inspection, operators care about the visibility of task-relevant surfaces and defects, and generic no-reference IQA scores can correlate weakly with inspection outcomes because they do not encode coverage or task-relevant defect visibility (Morgenthal & Hallermann, 2014). UAV deployments also impose constraints on energy and data freshness (Abedin et al., 2020), making quality judgments inherently multi-objective: mild blur can be acceptable when the inspected region remains resolvable, whereas a sharp image that misses the target surface is unusable. A useful IQA judgment should incorporate *coverage* (is the relevant structure present?) and *localized adequacy* (are critical regions sufficiently clear?). For example, during bridge inspection, an image with slight motion blur in non-critical regions (sky, background) may score lower on traditional metrics yet remain fully adequate for crack detection if the structural surface is clearly visible and in focus. This indicates that task-relevant regions (structural surfaces) require different quality criteria than background regions.

**AR-guided assembly.** In augmented reality (AR) assistance, visual quality should balance detail, contrast, and latency under real-time constraints, and the relevant criterion depends on whether the system needs to recognize parts, align overlays, or support user verification (Lv et al., 2025). A task-agnostic score does not capture these multi-objective trade-offs or the need for actionable feedback (e.g., increase exposure, adjust viewpoint). For instance, in manufacturing assembly tasks, AR systems should overlay virtual instructions onto physical components; quality assessment is expected to prioritize alignment accuracy and contrast in regions where overlays appear, rather than applying uniform quality criteria across the entire image. Research in AR-assisted assembly (Lv et al., 2025) highlights the importance of task-aware visual assessment for effective overlay alignment, though the specific relationship between quality metrics and alignment precision requires further empirical investigation.

**Summary:** Across these scenarios, the bottleneck is not merely scoring accuracy but *conditional adequacy*: quality depends on task, context, and action, and effective systems should explain what limits utility and what can be done next. The examples illustrate why traditional scoring-based IQA fails when quality requirements vary across tasks, when semantic coverage matters more than pixel fidelity, and when multi-objective trade-offs should be balanced under real-time constraints.

### 3.2. Theoretical Foundation: A Task-utility Conditional View of Quality

We formalize the distinction between traditional and embodied IQA through a conditional framework. Let  $I$  denote an image. Traditional IQA typically learns a task-agnostic predictor  $Q(I)$  from human opinion scores  $H(I)$  by minimizing a loss  $\mathbb{E}_{I \sim \mathcal{D}}[\ell(Q(I), H(I))]$ , and reports rank/linear correlation (e.g., the Spearman rank correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC)) on static test sets (Wang et al., 2004; Hore & Ziou, 2010; Saad et al., 2012; Mittal et al., 2012a). Embodied IQA instead requires a *conditional* judgment  $Q(I | T, C)$ , where  $T$  denotes a task specification and  $C$  denotes context (environment, constraints, goals). The objective is to maximize expected *task utility*:

$$\mathbb{E}_{(I,T,C) \sim \mathcal{D}}[U(Q(I | T, C), \text{task\_success}(I, T))], \quad (1)$$

where  $U(\cdot)$  measures how well the assessment supports downstream performance and decision-making.

**Proposition 3.1.** *When task-relevant features differ across tasks, access to  $(T, C)$  can yield higher expected task utility than an unconditional assessment  $Q(I)$  trained only to match task-agnostic ratings.*

*Proof sketch.* This result follows from the observation that when different tasks require different quality criteria, a single unconditional predictor compromises between conflicting requirements (Zhang et al., 2024a). Assume there exist two tasks  $T_1$  and  $T_2$  for which the utility-relevant cue depends on different image regions or features (e.g., crack visibility on a surface versus global aesthetics). Any unconditional  $Q(I)$  induces a single ranking/decision rule over images, which compromises between the conflicting criteria implied by  $T_1$  and  $T_2$ . A conditional assessor  $Q(I | T, C)$  can implement different rules for different  $(T, C)$ , aligning its judgments with the task-specific utility signal; this flexibility can improve expected utility whenever the optimal criteria differ across tasks and both tasks occur with non-zero probability. This aligns with information-theoretic principles (Cover, 1999) where conditional assessment provides more task-relevant information than unconditional assessment when task requirements diverge.  $\square$

### 3.3. Generalization Capability: Multi-task and Transfer Learning

When tasks share underlying structure, embodied IQA can benefit from transfer learning: features learned for one task can inform quality judgments for related tasks. This perspective motivates training and evaluation regimes that explicitly model task families rather than treating IQA as a single, universal score. A conditional framework with task and context information naturally supports transfer by enabling shared

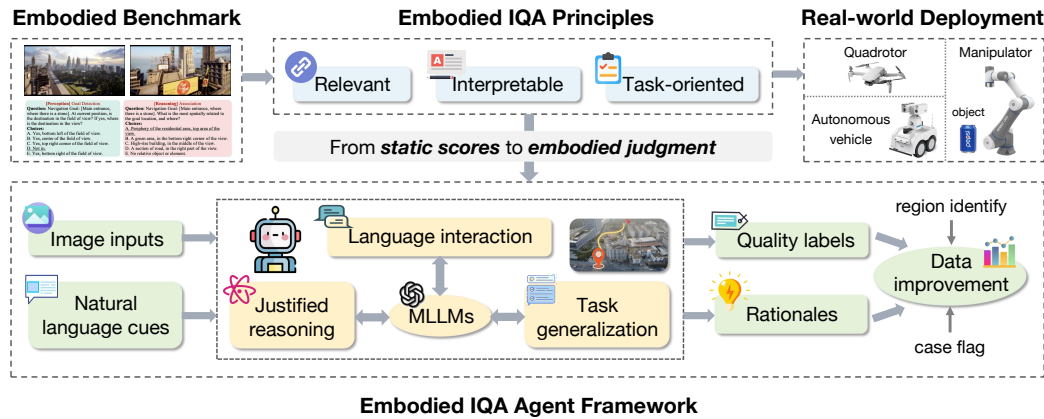


Figure 3. Multi-step decision pipeline of MLLM-based embodied IQA agents: from task understanding to knowledge grounding, visual rationale, tool refinement, and context-aware decision. The framework reconceptualizes IQA as a task-driven, interpretable, and context-relevant reasoning process rather than an absolute perceptual scoring problem.

representations across related tasks while retaining task-specific decision rules, allowing agents to generalize quality assessment to novel but structurally similar task scenarios.

#### 4. Framework for Embodied IQA Agent

This section formalizes the paradigm in Figure 3 as a structured decision pipeline for an MLLM-based embodied IQA agent, where image quality assessment is no longer treated as a static scoring problem but as a goal-driven, knowledge-augmented, and tool-enabled reasoning process coordinated by the MLLM agent.

**Step 1: Task understanding, decomposition, and knowledge grounding.** Given image inputs and natural language cues of the downstream goal, the MLLM-based embodied IQA agent first interprets the task and decomposes it into sequential subtasks, forming intermediate decision points for quality assessment. To establish appropriate evaluation criteria, the agent retrieves domain-specific expert knowledge through retrieval-augmented generation (RAG), grounding its reasoning in task-relevant standards rather than generic perceptual heuristics.

**Step 2: Rationale-guided quality judgment with visual grounding.** Conditioned on the task requirements and retrieved knowledge, the MLLM agent produces quality judgments such as rankings or usability decisions together with explicit textual rationales. These rationales are visually grounded by identifying task-relevant regions in the image, enabling region-level analysis that explains how specific quality factors influence task performance.

**Step 3: Tool-assisted refinement and utility-aware decision making.** When quality deficiencies are detected, the MLLM agent can invoke external tools to enhance the image for the current task, such as denoising, contrast adjustment, or cropping task-relevant areas. Throughout this pro-

cess, the agent estimates its confidence and flags ambiguous or borderline cases. By integrating perception, language-conditioned reasoning, knowledge retrieval, and tool usage, the MLLM-based IQA agent ultimately produces utility-aware quality decisions along with actionable feedback for improving image acquisition or processing parameters.

#### 5. Implementation of Embodied IQA Agent

Building embodied IQA agents coordinate design across four key components: embodied datasets, instruction-based training, grounded evaluation, and edge device deployment. Each component should support task-aware, interpretable, and context-grounded assessment. Specifically, embodied data defines the task contexts and utility signals that ground what “usable quality” means; training enables models to internalize these signals through goal-conditioned objectives and shared representations; evaluation measures success by downstream task support rather than score correlation; and deployment situates judgments within real pipelines. Together, these components form a closed loop that links task demands to practical, context-aware quality assessment in embodied systems.

##### 5.1. Embodied Datasets

Current IQA datasets (Lin et al., 2018) focus on general quality or specific distortions. Embodied IQA requires datasets that pair images with task specifications, environmental context, and goal-driven annotations. Recent embodied datasets (Wang et al., 2024; Yao et al., 2024; Gao et al., 2024) provide visual data from embodied platforms but lack quality annotations. Extending these with task-aware quality labels would support embodied IQA development. The annotation schema should include task specifications, quality judgments, explanations, and task performance outcomes.

## 5.2. Instruction-based Training

MLLMs should be trained to follow task-specific prompts like “assess image quality for obstacle detection” and produce reasoned outputs. Instruction tuning (Wu et al., 2024a) enables this capability, but current approaches focus on general quality rather than task-specific judgment. Training should incorporate pairwise comparisons, soft-label regression, and human feedback to align model judgments with task utility.

Direct Preference Optimization (DPO) (Rafailov et al., 2023) and related techniques from human feedback reinforcement learning (Christiano et al., 2017) can align IQA judgments with task performance rather than abstract quality scores. This addresses the evaluation misalignment problem by optimizing for true utility. The training objective combines utility loss (correlation with task performance), explanation loss (similarity to expert annotations), and consistency loss (stability across similar tasks). See 1, Equation 2, and Equation 3 in the appendix for detailed formulations.

## 5.3. Grounded Evaluation and Deployment

Standard IQA metrics, such as the Pearson linear correlation coefficient (PLCC) and Spearman rank correlation coefficient (SRCC), measure correlation with human ratings but do not assess task utility. Embodied IQA evaluation should measure three dimensions: usefulness (alignment with task performance), explanation fidelity (alignment with visual evidence), and cross-context consistency (appropriate adaptation to changing goals). Deployment requires lightweight, real-time systems that integrate with decision-making pipelines. Knowledge distillation (Hinton et al., 2015; Zhang et al., 2024a) and quantization (Rastegari et al., 2016) can optimize for edge devices. See subsection B.4 in the appendix for detailed evaluation protocols and deployment strategies.

## 6. Alternative Views

This section addresses alternative perspectives on our position, focusing on two practical questions: whether MLLMs are mature enough and how closely embodied IQA should mirror human judgment.

### 6.1. Are MLLMs Mature Enough for Embodied IQA?

A valid concern is whether MLLMs are sufficiently mature for embodied IQA. They can hallucinate, struggle with fine-grained visual details, and require significant computational resources. These limitations are real and must be addressed, particularly in safety-critical applications where errors can have serious consequences.

Recent advances suggest these limitations are surmount-

able. VisualCritic (Huang et al., 2024b) explores mechanisms to reduce ungrounded judgments while providing natural language rationales. Q-ALIGN (Wu et al., 2023) improves alignment by moving to discrete, text-defined quality levels that are more consistent with how human raters operate in subjective studies. Efficiency techniques (Zhang et al., 2024a) indicate that knowledge distillation and quantization can make deployment more practical on resource-constrained platforms.

The question is not whether MLLMs are perfect but whether they are better suited for embodied IQA than traditional methods. Given the fundamental limitations of regression-based approaches in task-aware, explainable, and adaptive assessment, MLLMs represent the most promising path forward despite current imperfections. Traditional methods are mature but fundamentally inadequate; MLLMs are less mature but architecturally appropriate. Maturity improves through use: embracing MLLMs now, while acknowledging limitations and working to address them, accelerates progress toward embodied IQA.

### 6.2. Human-Like vs. Machine-Optimized Perception

A philosophical alternative questions whether IQA should mimic human judgment or develop machine-optimized assessment principles. Human perception has limitations: it is context-dependent, sometimes inconsistent, and may miss task-critical signals that machines can detect.

A complementary approach is needed: embodied IQA agents should selectively adopt human-like traits (interpretability, task awareness, contextual reasoning) while leveraging machine capabilities (consistency, scalability, fine-grained analysis). The goal is not to replicate human perception exactly but to build systems that combine human and machine strengths.

This balance is already evident in related domains. Medical imaging systems use both human expertise and machine analysis, with each contributing unique strengths (Litjens et al., 2017). Embodied IQA should follow this pattern, creating systems that are interpretable and human-aligned while leveraging machine capabilities for consistency and scalability.

The human-like versus machine-optimized debate is often framed as a choice, but it is better understood as a design space. Different applications require different balances. Safety-critical applications may prioritize human-like interpretability, while high-throughput applications may prioritize machine consistency. The key is designing systems that can adapt this balance based on context and requirements, as depicted in Figure 4.

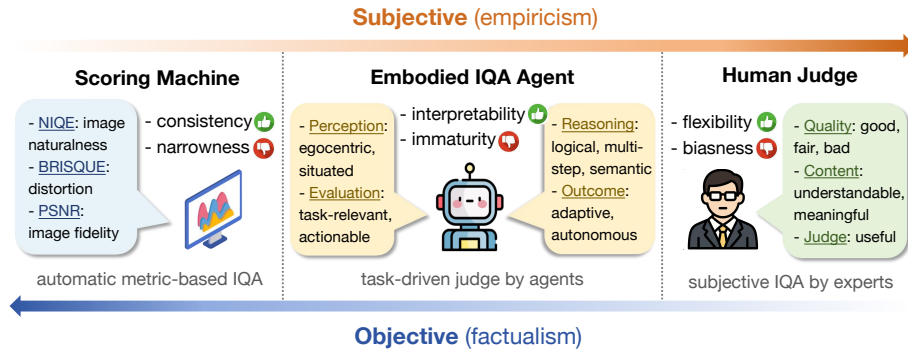


Figure 4. Embodied IQA Agent integrates strengths of human judge and scoring machine, enabling task-driven autonomous assessment while preserving interpretability and consistency.

## 7. Discussion

Although early signs of this paradigm appear in embodied AI, multimodal agents, and task-aware perception, these efforts remain fragmented and have not been unified into a coherent framework for embodied IQA.

**(1) Evaluation beyond correlation.** Since embodied IQA aims to support action, evaluation should reflect decision utility rather than score–rating correlation. While some embodied vision works implicitly judge perception by task success, IQA evaluation still relies almost exclusively on SRCC/PLCC against human opinion scores on static datasets. A key missing component is benchmarks that explicitly connect quality judgments to decision outcomes. This calls for task-conditioned datasets with  $(I, T, C)$ , downstream success labels, and expert rationales, as well as closed-loop metrics that measure whether an agent correctly decides to retake, reposition, or abstain. Real scenarios such as UAV inspection or robotic manipulation provide measurable task outcomes and natural testbeds.

**(2) Generality vs. task specificity.** Embodied IQA should generalize across tasks while remaining sensitive to task-critical cues. Advances such as parameter-efficient adaptation, retrieval augmentation, and modular routing in foundation models make this technically feasible, yet no existing IQA system separates generic distortion perception from task-utility reasoning. A promising direction is to combine shared quality encoders with task-conditioned adapters, retrieval of task guidelines at inference time, and routing mechanisms that decouple distortion perception from utility reasoning.

**(3) Safety assurance.** In high-stakes settings, systems should express uncertainty and avoid overconfident judgments. Although uncertainty estimation and abstention are studied in safety-critical ML, IQA models are still designed to always output a score, creating risk when deployed in embodied agents. Integrating confidence-aware abstention, grounded rationales linked to visual evidence, and stress

tests under lighting, motion, and sensor shifts, together with task-dependent safety thresholds, is therefore necessary but largely unexplored.

**(4) Human-in-the-loop collaborative IQA.** Embodied IQA provides a concrete setting to study evaluation misalignment and human–AI collaboration. Preference learning, constrained optimization, and actionable explanations are active research topics, but rarely examined in the context of perception quality for physical tasks. Examples include learning task-dependent quality preferences (e.g., navigation versus manipulation), balancing latency and adequacy in real-time systems, and producing explanations that suggest actionable adjustments rather than descriptive labels.

## 8. Conclusion

This position paper argues that traditional image quality assessment is misaligned with embodied downstream tasks, where the key question is not how good an image looks, but whether it is usable for a given goal and context. We advocate a shift from task-agnostic scoring to goal-conditioned image usability judgment, framing IQA as a decision problem tied to downstream utility rather than perceptual fidelity alone. This perspective connects IQA with embodied AI, multimodal reasoning, and safety-aware decision making, and calls for new benchmarks, models, and evaluation protocols aligned with action requirements. We hope this work motivates the community to treat embodied IQA as a distinct research direction rather than an incremental extension of traditional quality prediction.

## Impact Statement

This position paper advocates reframing IQA toward embodied, reasoning-based agents. Positive impacts include improved reliability in autonomous systems by prioritizing task utility, and improved transparency via grounded explanations that support debugging and human oversight. Potential negative impacts include increased compute demands,

new failure modes from ungrounded rationales, and risks of over-specialization. Responsible progress requires careful evaluation, uncertainty-aware deployment, and validation in safety-critical settings.

## References

- Abedin, S. F., Munir, M. S., Tran, N. H., Han, Z., and Hong, C. S. Data freshness and energy-efficient uav navigation optimization: A deep reinforcement learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 22(9):5994–6006, 2020.
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.
- Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T., and Samek, W. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017.
- Chen, C., Mo, J., Hou, J., Wu, H., Liao, L., Sun, W., Yan, Q., and Lin, W. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024a.
- Chen, C., Yang, S., Wu, H., Liao, L., Zhang, Z., Wang, A., Sun, W., Yan, Q., and Lin, W. Q-ground: Image quality grounding with large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 486–495, 2024b.
- Chen, Z., Qin, H., Wang, J., Yuan, C., Li, B., Hu, W., and Wang, L. Promptiqa: Boosting the performance and generalization for no-reference image quality assessment via prompts. In *European Conference on Computer Vision*, pp. 247–264. Springer, 2024c.
- Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- Cover, T. M. *Elements of information theory*. John Wiley & Sons, 1999.
- Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.
- Gao, C., Zhao, B., Zhang, W., Mao, J., Zhang, J., Zheng, Z., Man, F., Fang, J., Zhou, Z., Cui, J., et al. Embodiedcity: A benchmark platform for embodied agent in real-world city environment. *arXiv preprint arXiv:2410.09604*, 2024.
- Gao, X., Lu, W., Tao, D., and Li, X. Image quality assessment and human visual system. In *Visual Communications and Image Processing 2010*, volume 7744, pp. 316–325. SPIE, 2010.
- Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- Hore, A. and Ziou, D. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pp. 2366–2369. IEEE, 2010.
- Huang, Y., Yuan, Q., Sheng, X., Yang, Z., Wu, H., Chen, P., Yang, Y., Li, L., and Lin, W. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*, 2024a.
- Huang, Z., Zhang, Z., Lu, Y., Zha, Z.-J., Chen, Z., and Guo, B. Visualcritic: Making llms perceive visual quality like humans. *arXiv preprint arXiv:2403.12806*, 2024b.
- Kerkouri, M. A., Tliba, M., Chetouani, A., Aburaed, N., and Bruno, A. Modeling beyond mos: Quality assessment models must integrate context, reasoning, and multimodality. *arXiv preprint arXiv:2505.19696*, 2025.
- Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., and Levy, O. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.
- Li, C., Tian, Y., Ling, X., Zhang, Z., Duan, H., Wu, H., Jia, Z., Liu, X., Min, X., Lu, G., et al. Image quality assessment: From human to machine preference. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7570–7581, 2025a.
- Li, M., Wang, R., Sun, L., Bai, Y., and Chu, X. Next token is enough: Realistic image quality and aesthetic scoring with multimodal large language model. *arXiv preprint arXiv:2503.06141*, 2025b.
- Li, W., Zhang, X., Zhao, S., Zhang, Y., Li, J., Zhang, L., and Zhang, J. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679*, 2025c.

- Li, W., Zhang, X., Zhao, S., Zhang, Y., Li, J., Zhang, L., and Zhang, J. Q-insight: Understanding image quality via visual reinforcement learning, 2025d. URL <https://arxiv.org/abs/2503.22679>.
- Li, Y., Wu, S., Duan, H., Zhu, Y., Jia, Q., and Zhai, G. Exploring instruction data quality for explainable image quality assessment. *arXiv preprint arXiv:2510.03880*, 2025e.
- Lin, H., Hosu, V., and Saupe, D. Koniq-10k: Towards an ecologically valid and large-scale iqa database. *arXiv preprint arXiv:1803.08489*, 2018.
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88, 2017.
- Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- Lv, C., Liu, B., Wu, D., Lv, J., Li, J., and Bao, J. Ar-assisted assembly method based on instance segmentation. *International Journal of Computer Integrated Manufacturing*, 38(2):271–287, 2025.
- Ma, C., Shi, Z., Lu, Z., Xie, S., Chao, F., and Sui, Y. A survey on image quality assessment: Insights, analysis, and future outlook. *arXiv preprint arXiv:2502.08540*, 2025.
- Mittal, A., Moorthy, A. K., and Bovik, A. C. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012a.
- Mittal, A., Soundararajan, R., and Bovik, A. C. Making a ”completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012b.
- Morgenthal, G. and Hallermann, N. Quality assessment of unmanned aerial vehicle (uav) based visual inspection of structures. *Advances in Structural Engineering*, 17(3): 289–302, 2014.
- Nguyen, T.-M., Yuan, S., Nguyen, T. H., Yin, P., Cao, H., Xie, L., Wozniak, M., Jensfelt, P., Thiel, M., Ziegenbein, J., et al. Mcd: Diverse large-scale multi-campus dataset for robot perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22304–22313, 2024.
- Pan, W., Gao, T., Zhang, Y., Hu, R., Zheng, X., Zhang, E., Gao, Y., Liu, Y., Shen, Y., Li, K., et al. Multi-modal prompt learning on blind image quality assessment. *arXiv preprint arXiv:2404.14949*, 2024.
- Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741, 2023.
- Rastegari, M., Ordonez, V., Redmon, J., and Farhadi, A. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, pp. 525–542. Springer, 2016.
- Saad, M. A., Bovik, A. C., and Charrier, C. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE transactions on Image Processing*, 21(8):3339–3352, 2012.
- Song, C., Hui, C., Zhu, H., Jiang, F., Mi, Y., Zhang, W., and Liu, S. Segmenting and understanding: Region-aware semantic attention for fine-grained image quality assessment with large language models. *arXiv preprint arXiv:2508.07818*, 2025.
- Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., and Zhang, Y. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3667–3676, 2020.
- Wang, J., Chan, K. C., and Loy, C. C. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023.
- Wang, T., Mao, X., Zhu, C., Xu, R., Lyu, R., Li, P., Chen, X., Zhang, W., Chen, K., Xue, T., et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19757–19767, 2024.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al. Q-align: Teaching llms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.
- Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Xu, K., Li, C., Hou, J., Zhai, G., et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25490–25500, 2024a.

- Wu, H., Zhu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Li, C., Wang, A., Sun, W., Yan, Q., et al. Towards open-ended visual quality comparison. In *European Conference on Computer Vision*, pp. 360–377. Springer, 2024b.
- Wu, T., Ma, K., Liang, J., Yang, Y., and Zhang, L. A comprehensive study of multimodal large language models for image quality assessment. In *European Conference on Computer Vision*, pp. 143–160. Springer, 2024c.
- Xiao, J., Yan, B., Zhang, J., Wang, J., Li, C., Cheng, Z., and Zhai, G. Data assessment for embodied intelligence. *arXiv preprint arXiv:2511.09119*, 2025.
- Yang, J., Fu, J., Zhang, Z., Liu, L., Li, Q., Zhang, W., and Cao, W. Align-iqa: aligning image quality assessment models with diverse human preferences via customizable guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10008–10017, 2024.
- Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., and Yang, Y. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1191–1200, 2022.
- Yao, F., Yue, Y., Liu, Y., Sun, X., and Fu, K. Aeroverse: Uav-agent benchmark suite for simulating, pre-training, finetuning, and evaluating aerospace embodied world models. *arXiv preprint arXiv:2408.15511*, 2024.
- You, Z., Li, Z., Gu, J., Yin, Z., Xue, T., and Dong, C. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *European Conference on Computer Vision*, pp. 259–276. Springer, 2024.
- You, Z., Cai, X., Gu, J., Xue, T., and Dong, C. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14483–14494, 2025a.
- You, Z., Cai, X., Gu, J., Xue, T., and Dong, C. Teaching large language models to regress accurate image quality scores using score distribution, 2025b. URL <https://arxiv.org/abs/2501.11561>.
- Zha, J., Fan, Y., Zhang, T., Chen, G., Chen, Y., Gao, C., and Chen, X. Aircopbench: A benchmark for multi-drone collaborative embodied perception and reasoning. *arXiv preprint arXiv:2511.11025*, 2025.
- Zhang, J., Li, C., Hao, J., Jia, J., Duan, H., Zheng, G., Yuan, L., and Zhai, G. Embodied image quality assessment for robotic intelligence. *arXiv preprint arXiv:2412.18774*, 2024a.
- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.
- Zhang, W., Zhai, G., Wei, Y., Yang, X., and Ma, K. 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*, pp. 14071–14081, 2023.
- Zhang, Z., Wu, H., Zhang, E., Zhai, G., and Lin, W. Q-bench: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.
- Zhang, Z., Wu, H., Jia, Z., Lin, W., and Zhai, G. Teaching llms for image quality scoring and interpreting. *arXiv preprint arXiv:2503.09197*, 2025a.
- Zhang, Z., Wu, H., Jia, Z., Lin, W., and Zhai, G. Teaching llms for image quality scoring and interpreting, 2025b. URL <https://arxiv.org/abs/2503.09197>.
- Zhao, Z., Yue, X., Sun, J., Xie, Y., Shao, T., Yao, L., Xia, F., and Deng, Y. idetex: Empowering mllms for intelligent detailed explainable iqa. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3944–3953, 2025.
- Zhou, M., Shen, W., Wei, X., Luo, J., Jia, F., Zhuang, X., and Jia, W. Blind image quality assessment: Exploring content fidelity perceptibility via quality adversarial learning. *International Journal of Computer Vision*, pp. 1–17, 2025.
- Zhu, H., Wu, H., Li, Y., Zhang, Z., Chen, B., Zhu, L., Fang, Y., Zhai, G., Lin, W., and Wang, S. Adaptive image quality assessment via teaching large multimodal model to compare. *Advances in Neural Information Processing Systems*, 37:32611–32629, 2024.
- Zhu, H., Tian, Y., Ding, K., Chen, B., Chen, B., Wang, S., and Lin, W. Agenticiqa: An agentic framework for adaptive and interpretable image quality assessment. *arXiv preprint arXiv:2509.26006*, 2025.

## A. Literature Review

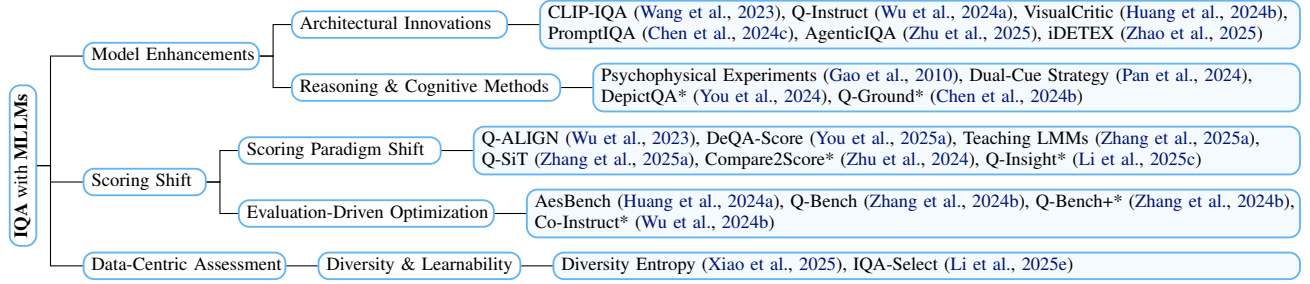


Figure 5. A Taxonomy of models for IQA with MLLMs: Model Enhancements, Scoring Shift, and Data-Centric Assessment. Methods marked with \* are for comparative multi-image assessment, while others are for single images.

## B. Training Formulations and Algorithms

### B.1. Training Objective

The training objective for embodied IQA can be formalized as follows. Given a dataset  $\mathcal{D} = \{(I_i, T_i, C_i, P_i, E_i)\}$  where  $I_i$  is an image,  $T_i$  is a task,  $C_i$  is context,  $P_i$  is task performance outcome, and  $E_i$  is an expert explanation, we optimize:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{utility}} + \lambda_2 \mathcal{L}_{\text{explanation}} + \lambda_3 \mathcal{L}_{\text{consistency}}, \quad (2)$$

where  $\mathcal{L}_{\text{utility}} = -\mathbb{E}[\text{corr}(Q(I|T, C), P)]$  measures correlation between quality assessment and task performance,  $\mathcal{L}_{\text{explanation}} = \text{sim}(E_{\text{model}}, E)$  measures explanation similarity to expert annotations, and  $\mathcal{L}_{\text{consistency}}$  enforces that similar tasks produce similar quality criteria while different tasks allow adaptation.

### B.2. Direct Preference Optimization

For preference-based training, we use pairs  $(I_1, I_2)$  where preference is determined by task performance: if task success is higher with  $I_1$  than  $I_2$ , the model should prefer  $I_1$  even if  $I_2$  scores higher on traditional metrics. Let  $x_i = (I_i, T, C)$  denote the conditioning input. The DPO objective becomes:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma \left( \beta \left[ \log \frac{\pi_{\theta}(y_1 | x_1)}{\pi_{\theta}(y_2 | x_2)} - \log \frac{\pi_{\text{ref}}(y_1 | x_1)}{\pi_{\text{ref}}(y_2 | x_2)} \right] \right), \quad (3)$$

where  $y_1, y_2$  are quality judgments,  $\pi_{\theta}$  is the policy being optimized,  $\pi_{\text{ref}}$  is a reference policy, and  $\beta$  controls the strength of the preference signal.

### B.3. Training Algorithm

The training procedure begins by initializing model parameters from a pre-trained MLLM, leveraging transfer learning from large-scale vision-language pretraining. During training, batches of images paired with task specifications and context are processed, extracting task-relevant features through multimodal encoding of  $(I_i, T_i, C_i)$ . This conditioning on task and context enables task-aware representations that differ fundamentally from traditional IQA methods that process images in isolation.

Feature extraction conditions on  $(I_i, T_i, C_i)$  rather than just  $I_i$ , enabling the model to learn task-dependent quality representations. For crack detection, the model emphasizes edge clarity and surface visibility; for object manipulation, it prioritizes pose-relevant information. The utility loss uses correlation with actual task performance  $P_i$  rather than human quality ratings, directly addressing evaluation misalignment by optimizing for task success. Explanation generation is integrated through similarity to expert annotations, ensuring quality judgments come with interpretable rationales. DPO operates on preference pairs determined by task performance rather than quality scores, aligning model preferences with actual utility. The model learns to prefer images that lead to task success, even if they score lower on traditional quality metrics.

**Algorithm 1** Embodied IQA Training with Task-Utility Alignment**Require:** Dataset  $\mathcal{D} = \{(I_i, T_i, C_i, P_i, E_i)\}$  where  $P_i$  is task performance**Require:** Pre-trained MLLM  $f_\theta$ , learning rate  $\alpha$ , DPO coefficient  $\beta$ **Ensure:** Fine-tuned model  $f_{\theta^*}$ 

```

1: Initialize  $\theta$  from pre-trained MLLM weights
2: for epoch = 1 to  $E$  do
3:   for batch  $(I_b, T_b, C_b, P_b, E_b) \sim \mathcal{D}$  do
4:     Extract task-relevant features:  $h_i = f_\theta(I_i, T_i, C_i)$ 
5:     Generate quality judgment:  $y_i = \text{Head}(h_i)$ 
6:     Compute utility loss:  $\mathcal{L}_{\text{utility}} = -\text{corr}(y_i, P_i)$ 
7:     Compute explanation loss:  $\mathcal{L}_{\text{expl}} = \text{sim}(\text{Gen}(h_i), E_i)$ 
8:     Sample preference pairs  $(I_1, I_2)$  where  $P_1 > P_2$ 
9:     Compute DPO loss:  $\mathcal{L}_{\text{DPO}}$  using Equation 3
10:    Total loss:  $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{utility}} + \lambda_2 \mathcal{L}_{\text{expl}} + \lambda_3 \mathcal{L}_{\text{DPO}}$ 
11:    Update:  $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}$ 
12:   end for
13: end for

```

**B.4. Evaluation Protocols**

We refer to subsection 5.3 for the main evaluation principles. In the appendix, we note two practical additions: (i) include *decision-centric* tests (e.g., whether the system correctly triggers retakes, repositioning, or abstention), and (ii) evaluate explanation fidelity via region-level grounding and expert verification (Lin et al., 2018; Kirstain et al., 2023). Concretely, benchmark samples can include  $(I, T, C)$  paired with task outcomes  $P$  and expert rationales  $E$ , enabling evaluation of both utility prediction and rationale faithfulness.

**B.5. Deployment Strategies**

Deployment requires lightweight, real-time systems that integrate with decision-making pipelines. Knowledge distillation can compress large MLLMs into efficient models (Zhang et al., 2024a). Specifically, a large teacher MLLM generates quality judgments and explanations, while a smaller student model learns to replicate these outputs with reduced computational cost. Quantization and hardware-aware neural architecture search can further optimize for edge devices, enabling real-time inference on resource-constrained platforms.

In UAV navigation (Abedin et al., 2020), AR inspection (Lv et al., 2025), and robot perception (Nguyen et al., 2024), IQA agents must provide low-latency feedback, support user interaction, and operate within hardware constraints. UAV systems require energy-efficient navigation optimization (Abedin et al., 2020), while AR-assisted assembly methods (Lv et al., 2025) need real-time quality assessment for manufacturing tasks. Robot perception systems (Nguyen et al., 2024) benefit from diverse multi-campus datasets that capture real-world variability. Integration with control systems enables closed-loop optimization, transforming IQA from post-hoc analysis to real-time perceptual support. For example, a drone’s IQA agent can trigger image retaking when quality is insufficient for the current mission, or adjust flight parameters to improve image quality for specific tasks.

**C. Preliminary Evaluation****C.1. Baselines**

In our experiments, we compare the proposed embodied IQA agent with the following state-of-the-art MLLM-based IQA baselines:

- **Q-Insight** (Li et al., 2025d) introduces a Reinforcement Learning (RL) based framework for image quality understanding. It utilizes Group Relative Policy Optimization (GRPO) to jointly optimize score regression and degradation perception tasks using carefully designed reward functions, demonstrating strong visual reasoning capabilities with limited data.

- **Q-SIT** (Zhang et al., 2025b) proposes a unified framework capable of performing both image quality scoring and interpreting simultaneously. It incorporates a "Scoring & Interpreting Balance Strategy" that efficiently determines the optimal data mixture ratio via a lightweight proxy model to mitigate task interference and enhance cross-task knowledge transfer.
- **DeQA-Score** (You et al., 2025b) focuses on accurate score regression by addressing the mismatch between continuous quality scores and discrete MLLM outputs. It employs a distribution-based soft label to preserve score distribution characteristics and introduces a fidelity loss based on Thurstone’s model to capture intra-dataset relationships during multi-dataset co-training.

## C.2. Metrics

To quantitatively evaluate image quality, we employ a diverse set of metrics ranging from traditional signal-fidelity measures to state-of-the-art deep learning-based perceptual models.

### C.2.1. FULL-REFERENCE (FR) IQA METRICS

These metrics compare the distorted image  $x$  against the pristine ground-truth image  $y$ .

- **MSE (Mean Squared Error)**: MSE measures the average squared difference between the estimated values and the actual value. It represents the signal fidelity in terms of pixel-wise error.

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (x_i - y_i)^2 \quad (4)$$

where  $N$  is the total number of pixels. A lower MSE indicates less error and higher similarity to the reference.

- **PSNR (Peak Signal-to-Noise Ratio)**: PSNR is derived from MSE and expresses the ratio between the maximum possible power of a signal and the power of corrupting noise. It is widely used to quantify reconstruction quality.

$$\text{PSNR} = 10 \cdot \log_{10} \left( \frac{L^2}{\text{MSE}} \right) \quad (5)$$

where  $L$  is the maximum possible pixel value (e.g., 255 for 8-bit images). Higher PSNR values imply better quality.

- **SSIM (Structural Similarity Index Measure)** (Wang et al., 2004): SSIM mimics the human visual system (HVS) by evaluating structural information degradation rather than pixel-wise error. It assesses three components: luminance ( $l$ ), contrast ( $c$ ), and structure ( $s$ ).

$$\text{SSIM}(x, y) = [l(x, y)]^\alpha \cdot [c(x, y)]^\beta \cdot [s(x, y)]^\gamma \quad (6)$$

Values range from 0 to 1, where 1 indicates perfect structural identity.

- **LPIPS (Learned Perceptual Image Patch Similarity)** (Zhang et al., 2018): LPIPS evaluates the distance between image patches in the deep feature space of a pre-trained convolutional neural network (e.g., VGG or AlexNet). It correlates better with human perceptual judgment than traditional metrics like MSE/PSNR.

$$d(x, y) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}_{hw}^l - \hat{y}_{0hw}^l)\|_2^2 \quad (7)$$

Lower LPIPS scores indicate that the distorted image is perceptually closer to the reference.

- **DISTS (Deep Image Structure and Texture Similarity)** (Ding et al., 2020): DISTS is a deep learning-based metric that specifically focuses on texture and structure invariance. It combines structural similarity measurements with deep feature maps to be robust against texture substitution, aligning closely with human texture perception. Lower values indicate higher similarity.

C.2.2. NO-REFERENCE (NR) IQA METRICS

These metrics assess image quality without access to a ground-truth reference, relying on learned statistical models of “naturalness.”

- **NIQE (Natural Image Quality Evaluator) (Mittal et al., 2012b):** NIQE is a completely blind IQA model based on constructing a “quality-aware” collection of statistical features derived from a corpus of natural, undistorted images. It measures the distance between the multivariate Gaussian (MVG) model of the test image and the MVG model of natural images. Lower scores indicate better perceptual quality (more “natural”).
- **BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator) (Mittal et al., 2012a):** BRISQUE operates in the spatial domain using scene statistics of locally normalized luminance coefficients. It quantifies the loss of “naturalness” using a support vector regression (SVR) model trained on human-rated datasets. Lower scores indicate better quality.
- **MANIQA (Multi-dimension Attention Network for IQA) (Yang et al., 2022):** MANIQA is a state-of-the-art deep learning method that utilizes a Multi-dimension Attention Network. It leverages Vision Transformer (ViT) features and applies channel-wise and spatial attention mechanisms to capture interaction across different regions, predicting a quality score that aligns with human mean opinion scores (MOS). Higher scores indicate better quality.
- **HyperIQA (Hyper-network for IQA) (Su et al., 2020):** HyperIQA addresses the content-dependency problem in blind IQA. It uses a self-adaptive hyper-network to separate content understanding from quality prediction, generating specific weight parameters for different image contents to predict the final quality score. Higher scores indicate better quality.

C.3. Task Scenarios

**Task 1 (Robotic Grasping):** *In a precise manipulation scenario, the objective is to resolve the 6D pose of the target object and target coordinates, ensuring visual cues are robust against quantization artifacts for successful grasp execution.*

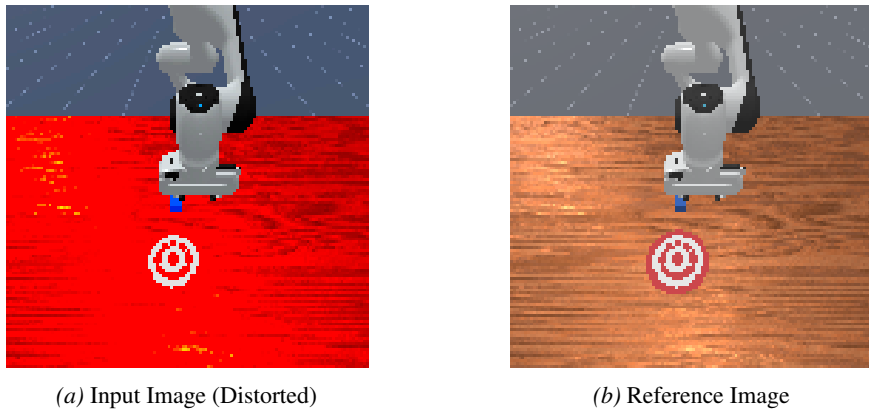


Figure 6. Visual visualization for Task 1 (Robotic Grasping). Left: The distorted input; Right: The reference ground truth.

Detailed Results for Task 1:

Traditional & MLLM-based Methods
<p><b>[Traditional IQA Scores]</b></p> <p>% Full-Reference (FR-IQA)</p> <p>MSE: 5925.8981 (↓ Low)</p> <p>PSNR: 10.4033 (↑ High)</p> <p>SSIM [0, 1]: 0.1505 (↑ High)</p> <p>LPIPS [0, 1]: 0.7858 (↓ Low)</p> <p>DISTS [0, 1]: 0.4567 (↓ Low)</p> <p>% No-Reference (NR-IQA)</p>

NIQE: 8.2755 (↓ Low)  
 BRISQUE: 22.9999 (↓ Low)  
 MANIQA [0, 1]: 0.1501 (↑ High)  
 HyperIQA [0, 1]: 0.2694 (↑ High)

**[MLLM-based IQA methods]**

Q-Insight:

Model Response:

<think>

This image appears to be a digitally generated or pixelated image, possibly a sprite from a video game or an illustration, given the style. The image contains recognizable elements such as a humanoid figure and a target, suggesting it's a constructed or intended image. There's a level of detail present, but the pixelation and resolution give the image a somewhat retro or stylized aesthetic. The colors are distinct, and there's a clear contrast between elements, which aids in visibility and comprehensibility. Overall, it's clear and legible, but lacks fine detail. Judging by the pixelation, lack of fine detail, and resolution, the image quality isn't quite at a high level, but the content is still understandable and appealing, especially for what it seems to be.

</think>

<answer>

"rating": 3.00

</answer>

Q-SIT: 4.542420948398246

DeQA-Score: 2.0176

**Embodied IQA Agent**

**1. Context Adaptation Report**

The aerial perspective provides a wide field of view with neutral-toned background elements (gray asphalt, green foliage). The lighting is diffuse (overcast), which minimizes harsh specular reflections that might otherwise distort color perception. While there is a slight global softness typical of high-altitude surveillance footage, the color saturation of red remains a dominant visual feature in the scene.

**2. Task-Grounded Explanation**

Localization: The target red vehicle is located in the center-right quadrant, specifically on the paved shoulder/sidewalk area adjacent to the row of trees. Impact Analysis: The visual conditions are optimal for the specific detection task. The red color signature exhibits high chromatic contrast against both the gray road surface and the green grass, allowing for immediate color-based thresholding. Furthermore, despite the high altitude, the vehicle's geometric boundaries are sharp enough to allow a detection algorithm to generate a precise bounding box without confusion from neighboring objects or background textures.

**3. Embodied Integration (Control Decision Support)**

Perception Action: The vision system should prioritize a color-space segmentation (e.g., HSV filtering) to isolate the target. The current signal-to-noise ratio is high enough to support automated tracking with minimal risk of identity switching or loss of lock.

Confidence Threshold: High. No additional image enhancement or altitude adjustment is required for successful task execution.

**4. Embodied Utility Score: 5.0**

Listing 1. Quantitative and reasoning results for Task 1.

**Task 2 (Aerial Detection):** *In a high-altitude surveillance scenario, the objective is to locate and bound a specific target red vehicle based on its chromatic signature amidst neutral background elements.*



Figure 7. Input Image for Task 2.

**Detailed Results for Task 2:**

**Traditional & MLLM-based Methods**

**[Traditional IQA Scores]**

NIQE: 3.0114 (↓ Low)  
 BRISQUE: 1.8237 (↓ Low)  
 MANIQA [0,1]: 0.2950 (↑ High)  
 HyperIQA [0,1]: 0.5463 (↑ High)

**[MLLM-based IQA methods]**

Q-Insight:

Model Response:

<think>

This is a picture that has a good level of detail and clear lines, as well as good contrast between the roads, vehicles, and surrounding greenery. The colors are balanced, and the subject matter is in focus. The composition provides a clear view of the subjects, though there is a slight graininess or noise in the image, which slightly decreases the sharpness but does not significantly impact the overall image quality. There is a lack of motion blur, which is evident in the car and vehicles, giving it a clear and legible appearance.

</think>

<answer>

"rating": 3.80

</answer>

Q-SIT: 4.542420948398246

DeQA-Score: 2.0176

**Embodied IQA Agent**

**1. Context Adaptation Report**

The aerial perspective provides a wide field of view with neutral-toned background elements (gray asphalt, green foliage). The lighting is diffuse (overcast), which minimizes harsh specular reflections that might otherwise distort color perception. While there is a slight global softness typical of high-altitude surveillance footage, the color saturation of red remains a dominant visual feature in the scene.

**2. Task-Grounded Explanation**

Localization: The target red vehicle is located in the center-right quadrant, specifically on the paved shoulder/sidewalk area adjacent to the row of trees.

**Impact Analysis:** The visual conditions are optimal for the specific detection task. The red color signature exhibits high chromatic contrast against both the gray road surface and the green grass, allowing for immediate color-based thresholding. Furthermore, despite the high altitude, the vehicle's geometric boundaries are sharp enough to allow a detection algorithm to generate a precise bounding box without confusion from neighboring objects or background textures.

### 3. Embodied Integration (Control Decision Support)

**Perception Action:** The vision system should prioritize a color-space segmentation (e.g., HSV filtering) to isolate the target. The current signal-to-noise ratio is high enough to support automated tracking with minimal risk of identity switching or loss of lock.

**Confidence Threshold:** High. No additional image enhancement or altitude adjustment is required for successful task execution.

### 4. Embodied Utility Score: 5.0

Listing 2. Quantitative and reasoning results for Task 2.

**Task 3 (Auto. Driving):** In a dynamic roadway environment, the objective is to evaluate the visual utility for safety-critical perception (e.g., lane topology, signal states) and identify fatal navigational violations such as wrong-way driving.



Figure 8. Input Image for Task 3.

### Detailed Results for Task 3:

#### Traditional & MLLM-based Methods

##### [Traditional IQA Scores]

NIQE: 3.1552 (↓ Low)  
BRISQUE: 16.7502 (↓ Low)  
MANIQA [0,1]: 0.3850 (↑ High)  
HyperIQA [0,1]: 0.5344 (↑ High)

##### [MLLM-based IQA methods]

Q-Insight:  
Model Response:

```
<think>
This picture is clear and detailed, providing a sense of depth and perspective,
as well as good lighting and composition. The image is well-balanced, with a
clear focus on the intersection, and it is free from blurriness or distracting
elements. The color balance appears natural and vibrant. The photograph
effectively captures the environment, suggesting good image quality that is
suitable for its subject.
</think>
<answer>
"rating": 4.17
</answer>
```

Q-SIT: 4.8053

DeQA-Score: 4.4023

## Embodied IQA Agent

### 1. Context Adaptation Report

The environment presents high-contrast lighting with a direct solar source in the upper right. Sharp shadows from roadside trees are cast across the asphalt. Despite these high-frequency visual interruptions, sensor reliability is excellent; the dynamic range is sufficient to maintain high edge-contrast for lane markings within the shadows. No lens flare or atmospheric occlusion (fog/rain) impairs the distant gantry or ground instructions.

### 2. Task-Grounded Explanation

Visual Robustness: Image quality is extremely high. The resolution and contrast support high-confidence detection of fine-grained semantics, including ground arrow contours and the mechanical structure of distant signal housings.

Compliance Check:

Directional Arrows: A clear white arrow on the road surface at the intersection entrance is oriented facing the vehicle.

Signal Orientation: The primary overhead traffic signals are seen from the rear (housing side), indicating they are intended for traffic traveling in the opposite direction.

Lane Logic: The vehicle is positioned to the left of a yellow dashed line; in standard right-hand traffic jurisdictions, this confirms the vehicle has crossed into the opposing flow of traffic.

Finding: CRITICAL STATE DETECTED - Wrong-Way Driving.

Safety Conclusion: While the visual data is "perfect" in terms of clarity, it provides undeniable evidence of a fatal navigational violation. The situation is inherently unsafe for continued autonomous operation.

### 3. Embodied Integration (Control Decision Support)

Driving Policy: Immediate Handover (Critical Failure/Violation).

Reasoning: The perception system has identified a high-confidence "Wrong-Way" scenario. Based on the Safety Veto logic, the viability of continuous autonomy is zero. The system must immediately trigger an emergency alert and hand over control to a human operator or execute a safe-stop maneuver to prevent a head-on collision.

### 4. Embodied Utility Score

1.0

(Justification: Even though the image is crystal clear (Sensor Quality = 5.0), the detection of a Critical Violation (Wrong-Way) mandates a score of 1.0. The data confirms that the current trajectory is illegal and life-threatening.)

Listing 3. Quantitative and reasoning results for Task 3.

**Task 4 (Authenticity Check):** *In a mixed-reality power grid inspection scenario, the objective is to filter out invalid synthetic data by detecting physical inconsistencies (e.g., floating artifacts without gravity support) that violate environmental realism.*



Figure 9. Input Image for Task 4.

### Detailed Results for Task 3:

#### Traditional & MLLM-based Methods

##### [Traditional IQA Scores]

NIQE: 3.1446 (↓ Low)  
BRISQUE: 24.9523 (↓ Low)  
MANIQA [0,1]: 0.3698 (↑ High)  
HyperIQA [0,1]: 0.4684 (↑ High)

##### [MLLM-based IQA methods]

Q-Insight:

Model Response:

<think>

This image appears to be an outdoor scene captured with a good level of detail and focus. The resolution seems adequate, with distinguishable details on the subject and clear outlines of the landscape. The image is not particularly vibrant in color, which could indicate a more natural or neutral environment. The photo shows some slight grain or noise, which is common in some types of image capture, but it does not significantly impact the overall composition or readability.

</think>

<answer>

"rating": 3.56

</answer>

Q-SIT: 4.0642

DeQA-Score: 3.6074

#### Embodied IQA Agent

##### 1. Context Adaptation Report

The image is a clear instance of a synthetic simulation or a "phantom" composite. While the background (fields, roads, and distant towers) appears to be a real-world aerial capture, the primary subject—the insulator string—exhibits massive physics violations. There is a total absence of mechanical support (it is floating in mid-air), no interaction with environmental lighting (it casts no shadow on the ground or nearby towers), and it possesses "jagged" edge artifacts that indicate a poor digital insertion.

Verdict: Synthetic / Fake Data.

## 2. Task-Grounded Explanation

**Authenticity Verification:** The insulator string is floating without any physical connection to the tower arms or transmission cables. It violates the law of gravity and the fundamental structural logic of power grid infrastructure.

**Defect Analysis:** Because the object is a digital "phantom" and not a real physical component in the scene, performing a health monitoring assessment is scientifically invalid. Any apparent "defects" would be artifacts of the rendering/compositing process rather than actual grid hazards.

**Inspection Conclusion:** Invalid Data. This frame is a "hallucination" or a test case designed to verify the perception system's ability to filter out non-existent objects.

## 3. Embodied Integration (Control Decision Support)

**Recommended Action:** Discard/Ignore.

**Reasoning:** The data is non-authentic. Recording this as a "healthy" or "defective" component would corrupt the grid's asset database with false information. The system must flag this frame as a sensor/processing anomaly and proceed to the next valid frame.

## 4. Embodied Utility Score

1.0

(Justification: The image is a synthetic "phantom" that provides zero utility for actual grid health monitoring. It is correctly identified as invalid data.)

Listing 4. Quantitative and reasoning results for Task 4.

**Task 5 (Defect Detection):** *In a verified real-world transmission line inspection, the objective is to assess the structural health of insulators and identify specific physical hazards, such as foreign objects (e.g., bird nests) or component breakage.*



Figure 10. Input Image for Task 5.

## Detailed Results for Task 5:

### Traditional & MLLM-based Methods

#### [Traditional IQA Scores]

NIQE: 2.6471 (↓ Low)  
 BRISQUE: 14.0626 (↓ Low)  
 MANIQA [0,1]: 0.5057 (↑ High)  
 HyperIQA [0,1]: 0.6792 (↑ High)

#### [MLLM-based IQA methods]

Q-Insight:

Model Response:

<think>

This picture appears to be clear with good focus, and the details of the power tower and the surrounding environment are well-preserved. There's a slight lack of sharpness, but the lighting is even, and the colors are natural. No obvious issues in color balance, exposure, or distortion. Overall, the picture seems balanced and the subject is well-defined within the frame.

</think>

<answer>

"rating": 3.87

</answer>

Q-SIT: 4.4320

DeQA-Score: 4.0547

### Embodied IQA Agent

#### 1. Context Adaptation Report

The scene exhibits high environmental realism. The lighting is diffuse and consistent with an overcast sky, casting soft, natural shadows across the lattice structure. Material textures—specifically the oxidation/weathering on the galvanized steel and the organic debris of the bird nest—show high-fidelity details that match real-world physical aging. The background (water ponds and levees) displays realistic atmospheric perspective and water-surface ripples. No "floating" artifacts or compositing halos are present.

Verdict: Real-world capture.

#### 2. Task-Grounded Explanation

Authenticity Verification: The image is physically valid. The insulator strings are correctly under tension and mechanically anchored to the tower's cross-arms using standard hardware (shackles and ball-eye links). The spatial relationship between the drone's perspective, the tower, and the ground is geometrically consistent.

Defect Analysis: A significant Foreign Object Hazard (Bird Nest) is identified on the left-hand cross-arm, located directly at the critical junction where the insulator string attaches to the tower. While the insulator discs themselves appear structurally intact (no visible broken sheds), the proximity of the nest to high-voltage components poses a high risk of flashover or fire during damp conditions.

Inspection Conclusion: Defect Hazard. This is a valid, high-quality inspection frame that has successfully captured a maintenance-critical anomaly.

#### 3. Embodied Integration (Control Decision Support)

Recommended Action: Trigger Defect Alert.

Reasoning: The detection of a bird nest in a high-voltage clearance zone is a primary maintenance trigger. The visual data is clear and unambiguous, providing sufficient evidence for a maintenance crew to be dispatched for removal.

#### 4. Embodied Utility Score

3.0

(Justification: This score represents a real, clear image where a specific defect/foreign object has been detected, necessitating system intervention.)

Listing 5. Quantitative and reasoning results for Task 5.

**Task 6 (Indoor Navigation):** *In a resource-constrained indoor environment (e.g., dimly lit corridors), the objective is to determine navigational actionability by verifying the visibility of essential geometric boundaries and obstacles despite severe sensor noise.*



Figure 11. Input Image for Task 6.

**Detailed Results for Task 6:**

**Traditional & MLLM-based Methods**

**[Traditional IQA Scores]**

NIQE: 11.8701 (↓ Low)  
 BRISQUE: 79.0786 (↓ Low)  
 MANIQA [0,1]: 0.2994 (↑ High)  
 HyperIQA [0,1]: 0.3356 (↑ High)

**[MLLM-based IQA methods]**

Q-Insight:  
 Model Response:  
 <think>  
 This picture appears to be of a hallway or corridor in an office or a similar building. The resolution is not very high, and there's a slight lack of sharpness, which may indicate motion blur or poor image focus. Lighting seems adequate but does not brighten the image enough, potentially making the subject slightly darker or obscured. It's slightly blurry, potentially because it was taken with a handheld device, or there was motion when the picture was taken, giving it a grainy texture that could mask further detail. The person in the picture is not fully defined, and there isn't a high level of fine detail or fine-textured elements in the scene that would be expected in a high-quality image.  
 </think>  
 <answer>  
 "rating": 2.30  
 </answer>

Q-SIT: 0.5875

DeQA-Score: 1.8262

## Embodied IQA Agent

### 1. Context Adaptation Report

The image is characterized by low light, significant sensor noise (grain), and noticeable motion blur. The dynamic range is poor, with the overhead light source causing localized blooming. However, despite being "low-fidelity" by aesthetic standards, the environmental geometry is not destroyed. The contrast between the light-colored walls and the dark-toned carpet remains sufficient for a standard Canny or Sobel edge-detection filter to identify the primary structural planes.

### 2. Task-Grounded Explanation

Geometry Check: The floor-wall boundaries are continuous and visible. The perspective lines of the corridor converge toward a clear vanishing point, which allows the navigation stack to maintain a reliable heading. Although the edges are "soft" due to blur, the transition from vertical to horizontal planes is semantically distinct.

Obstacle Detection:

Dynamic: The silhouette of the person walking ahead is high-contrast against the lighter background at the end of the hall, ensuring they are mapped as a dynamic obstacle in the occupancy grid.

Static: The trash bin on the left and the doorframe on the right have clear enough silhouettes for the robot to calculate a clearance buffer.

Navigability: There is a clear "traversable path" to the right of the bin and behind the pedestrian. The robot has enough visual data to calculate a local trajectory vector.

### 3. Embodied Integration (Control Decision Support)

Recommended Control State: Reduce Speed (Low Confidence).

Reasoning: The image is "Usable," but the low resolution and motion blur increase the margin of error for depth estimation. Reducing speed will allow the robot to perform temporal filtering (averaging multiple frames) to compensate for sensor noise and ensure that obstacle positions are stable before committing to a high-speed path.

### 4. Embodied Utility Score

3.5

(Justification: The data is noisy and dim, but the safety-critical information—the location of the floor, the walls, and the obstacles—is fully actionable. The robot can navigate safely, albeit with higher caution than in a well-lit environment.)

Listing 6. Quantitative and reasoning results for Task 6.

### C.4. Evaluation Results

Table 2. **Quantitative Comparison on Task-Oriented IQA.** We compare Embodied Agent against Traditional methods (FR/NR) and MLLM-based IQAs, including Q-Insight (Li et al., 2025d), Q-SIT (Zhang et al., 2025b), and DeQA-Score (You et al., 2025b). **FR:** Full-Reference, **NR:** No-Reference. Note that despite high aesthetic scores, existing MLLM-based IQAs fail to identify safety hazards in some tasks.

Task Scenario	Full-Reference (FR)					No-Reference (NR)				MLLM-based IQA			Embodied IQA Agent
	MSE↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	BRIS↓	MANI↑	Hyp.↑	Q-Ins.	Q-SIT	DeQA.	
Robotic Grasping	5925.9	10.40	0.15	0.79	0.46	8.28	23.00	0.15	0.27	3.00	1.43	1.97	3.50
Aerial Detection	–	–	–	–	–	3.01	1.82	0.30	0.55	3.80	4.54	2.02	5.00
Auto. Driving	–	–	–	–	–	3.16	16.75	0.39	0.53	4.17	4.81	4.40	1.00*
Authenticity Check	–	–	–	–	–	3.14	24.95	0.37	0.47	3.56	4.06	3.61	1.00*
Defect Detection	–	–	–	–	–	2.65	14.06	0.51	0.68	3.87	4.43	4.05	3.00
Indoor Navigation	–	–	–	–	–	11.87	11.87	0.30	0.34	2.30	0.59	1.83	3.50

\*Low scores indicate correct detection of critical safety issues (Safety Veto).

#### Key Insights:

- **Semantic Safety Veto (Auto. Driving):** While existing MLLMs (Q-SIT: 4.81, DeQA-Score: 4.40) assign high scores based on visual clarity, the Embodied IQA Agent correctly triggers a safety veto (1.00). It prioritizes the semantic hazard (“Wrong-Way Driving”) over aesthetic fidelity.
- **Physical Consistency Check (Authenticity Check):** Baselines fail to identify synthetic artifacts (3.61–4.06). In contrast, the Embodied IQA Agent detects physics violations (e.g., floating objects lacking gravity support) and rejects the data as operationally invalid.
- **Task-Grounded Actionability (Indoor Navigation):** Despite low-fidelity sensor noise that severely penalizes traditional metrics (Q-SIT: 0.59), the Embodied IQA Agent assesses the image as actionable (3.50), verifying that the essential traversable geometry remains intact for path planning.

## D. Detailed Case Study: UAV Infrastructure Inspection

This appendix provides a detailed analysis of how traditional IQA fails in UAV infrastructure inspection and how embodied IQA addresses these limitations.

### D.1. Scenario Description

A commercial drone inspection system captures images of bridge structures to detect structural damage, particularly cracks and corrosion. The system operates in varying conditions: different lighting conditions, wind speeds, and flight altitudes. Operators need to assess whether captured images are adequate for damage detection before proceeding with analysis.

### D.2. Traditional IQA Failure Modes

Traditional IQA methods fail in this scenario through several mechanisms. First, they penalize motion blur uniformly, but in bridge inspection, blur in non-critical regions (e.g., sky, background) does not affect crack detection while blur in the target structure does. Second, traditional methods cannot reason about whether critical structures are visible: an image might score highly on general quality metrics but fail to capture the bridge surface needed for inspection. Third, traditional methods provide no explanation, making it difficult for operators to understand why an image is inadequate or how to improve capture parameters.

Prior reports suggest that generic no-reference IQA scores can correlate weakly with inspection outcomes, because they do not encode coverage of the inspected surface or the task-relevant visibility of defects (Morgenthal & Hallermann, 2014). This motivates task-aware assessment criteria that explicitly account for region-level adequacy and semantic coverage.

### D.3. Embodied IQA Approach

An embodied IQA agent would assess images differently. Given a task prompt “Assess whether this image is adequate for detecting structural cracks in the bridge surface,” the agent would identify whether the bridge surface is visible and in focus, evaluate whether motion blur affects critical regions, consider lighting conditions relative to crack visibility requirements, and generate an explanation like “Image is adequate for crack detection. Bridge surface is clearly visible with minimal motion blur. Lighting is sufficient to reveal surface texture needed for crack identification.”

This approach addresses the limitations of traditional IQA by providing task-specific reasoning, actionable explanations, and

judgments that align with actual inspection needs rather than abstract quality metrics.

## E. Theoretical Extensions

### E.1. Information-Theoretic Analysis

From an information-theoretic perspective, task-aware IQA can be understood as maximizing mutual information between quality assessment and task performance (Cover, 1999). Let  $H(\text{task\_success}|I)$  denote the entropy of task success given image  $I$ , and  $H(\text{task\_success}|I, Q(I|T, C))$  denote the entropy after conditioning on quality assessment. Task-aware IQA maximizes  $I(\text{task\_success}; Q(I|T, C)|I)$ , the information that quality assessment provides about task success beyond what is available from the image alone. This information-theoretic framework formalizes why conditional assessment  $Q(I|T, C)$  provides more task-relevant information than unconditional assessment  $Q(I)$ , following the principle that conditional mutual information captures task-specific relevance (Cover, 1999).

Traditional IQA, by contrast, maximizes  $I(H(I); Q(I))$ , the information shared between human ratings and quality scores. When task-relevant features differ from general quality features, this optimization fails to capture information relevant to task performance, leading to evaluation misalignment (Zhang et al., 2024a).

### E.2. Generalization Bounds

We can analyze the generalization properties of embodied IQA through the lens of domain adaptation (Ben-David et al., 2010). Traditional IQA methods trained on general quality datasets may not generalize to task-specific scenarios due to domain shift. Embodied IQA, by explicitly conditioning on task and context, reduces this domain shift and improves generalization.

Formally, if we treat different tasks as different domains, embodied IQA performs domain adaptation by conditioning on task specifications. This reduces the effective domain gap compared to methods that assume a single domain (general quality assessment), leading to better generalization across tasks (Ganin et al., 2016).

## F. Implementation Roadmap

This appendix outlines a concrete roadmap for implementing embodied IQA systems, organized by component and timeline.

### F.1. Phase 1: Dataset Construction

The first phase focuses on constructing embodied IQA datasets. This involves collecting images from embodied platforms (drones, robots, AR systems) with associated task specifications, annotating images with task-relative quality judgments and explanations, and measuring actual task performance outcomes. Key challenges include ensuring task diversity, annotation quality, and scale. The required dataset scale depends on task diversity, annotation granularity, and desired generalization; meaningful embodied IQA systems would benefit from diverse task-conditioned examples across multiple domains.

### F.2. Phase 2: Model Development

The second phase develops MLLM-based IQA models. This involves fine-tuning existing MLLMs (e.g., LLaVA (Liu et al., 2023), GPT-4V (Achiam et al., 2023)) on embodied IQA datasets, implementing DPO-based training for task-utility alignment, and developing explanation generation capabilities. Key technical challenges include handling multimodal inputs, generating structured outputs, and ensuring real-time inference. We estimate this phase requires access to large-scale compute resources and expertise in MLLM fine-tuning.

### F.3. Phase 3: Evaluation and Benchmarking

The third phase establishes evaluation protocols and benchmarks. This involves defining metrics for usefulness (alignment with task performance), explanation fidelity (alignment with visual evidence), and cross-context consistency (appropriate adaptation to changing goals), creating benchmark datasets with ground-truth task performance and expert explanations, and evaluating models on these benchmarks. Key challenges include defining appropriate metrics and ensuring benchmark quality. We estimate this phase requires collaboration with domain experts and extensive validation. See subsection 5.3 in the main text for detailed evaluation principles.

### F.4. Phase 4: Deployment and Integration

The final phase focuses on real-world deployment. This involves optimizing models for efficiency through distillation and quantization, integrating with embodied systems (UAVs, robots, AR), and validating performance in operational settings. Key challenges include meeting latency requirements, ensuring robustness, and building trust with operators. We estimate this phase requires partnerships with industry and extensive field testing.

This roadmap demonstrates that embodied IQA is not merely a theoretical proposal but a concrete research program with

clear milestones and deliverables. The timeline is ambitious but achievable given current MLLM capabilities and the growing availability of embodied datasets.

## G. Practical Risks

Key risks of the implementation of MLLM-based embodied IQA agents include ungrounded justifications, over-specialization to narrow task prompts, and deployment constraints on latency and compute. Addressing them likely requires joint progress in grounding, efficiency (e.g., distillation (Zhang et al., 2024a)), and benchmark design that rewards safe behavior under uncertainty.

## H. Glossary of Key Concepts

To clarify terminology used throughout this paper, we provide definitions of key concepts:

**Embodied IQA:** Image quality assessment performed from a first-person perspective by an agent that considers task utility, environmental context, and goal-driven reasoning, rather than abstract distortion metrics.

**Task-aware assessment:** Quality evaluation that considers why an image is useful for a specific goal, reasoning about task relevance, semantic content, and the relationship between visual properties and downstream performance.

**Conditional quality judgment:** Quality assessment  $Q(I | T, C)$  that depends on task specification  $T$  and contextual information  $C$ , as opposed to unconditional assessment  $Q(I)$  that treats quality as an absolute property.

**Evaluation misalignment:** A machine learning problem where optimization objectives (e.g., correlation with human ratings) do not match desired outcomes (e.g., task performance), leading to suboptimal system behavior.

**Task-grounded explanation:** Explanations that relate visual properties to task utility (e.g., “motion blur obscures the target structure needed for navigation”) rather than merely describing distortions (e.g., “motion blur in central region”).