IQA-T1: Tool-based Visual Evidence Reasoning for Image Quality Assessment

Jinjian Wu[1]*, Jiaqi Tang[2]*, Wei Wei[1]✉, Yingying Yan[1], Jianmin Chen[1], Botong Geng[1], Lei Zhang[1], and Qifeng Chen[2]✉

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an, China
wujinjian@mail.nwpu.edu.cn weiweinwpu@nwpu.edu.cn
[2] The Hong Kong University of Science and Technology, Hong Kong SAR, China
jtang092@connect.ust.hk cqf@ust.hk

Abstract. Image Quality Assessment (IQA) in open-world environments remains challenging due to limited generalization and interpretability. Recent approaches based on multimodal large language models (MLLMs) introduce textual reasoning for quality prediction, yet their judgments rely heavily on semantically biased internal representations, making them insensitive to low-level perceptual degradations. We propose IQA-T1, a tool-based visual evidence reasoning framework that augments MLLM reasoning with explicit perceptual observations. During inference, the model autonomously invokes specialized analysis tools to generate structured visual evidence, such as noise residual maps, gradient statistics, and frequency spectra, which are progressively integrated into the reasoning process. To support this paradigm, we construct Q-Tool, a dataset containing 11k multimodal reasoning chains grounded in tool-generated evidence. Extensive experiments on seven IQA benchmarks show that IQA-T1 achieves the best overall performance across datasets while producing interpretable and evidence-grounded quality assessments. Code and dataset are available at https://github.com/zibuyu-02/IQA-T1.

Keywords: Image Quality Assessment · Multimodal Large Language Models · Tool-based Reasoning

1 Introduction

Image Quality Assessment (IQA) aims to predict perceptual image quality in alignment with human subjective judgment. In image restoration [13, 17, 26, 38], enhancement [2, 32, 40, 52], compression [10, 28, 54], and various downstream vision applications [42], image quality assessment not only plays an important role in algorithm performance but also directly affects the usability and robustness of vision systems in open-world environments.

With the rapid development of multimodal large language models (MLLMs) [1, 20, 39, 43, 53], researchers have begun to explore leveraging their reasoning capabilities to shift IQA from score regression toward language-driven frameworks.

* Equal contribution. ✉ Correspondence to: Wei Wei and Qifeng Chen.

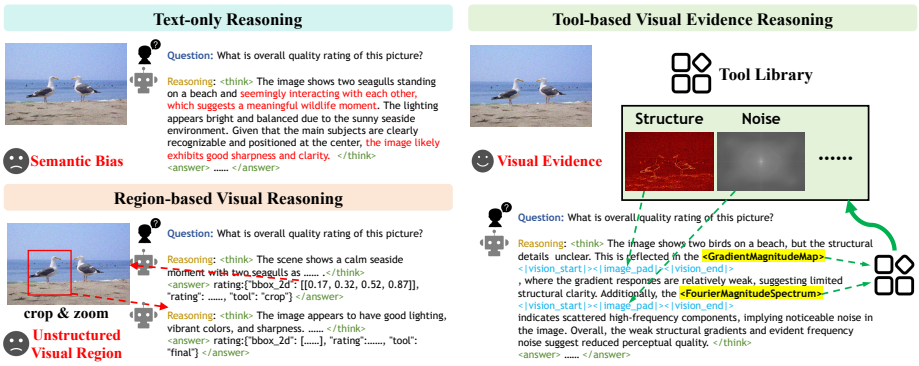

Fig. 1: Motivation of our visual evidence reasoning framework. Existing MLLM-based IQA methods typically follow two reasoning paradigms: text-only reasoning, which is prone to semantic bias, and region-based visual reasoning using cropped image patches that lack explicit perceptual interpretation. In contrast, our framework introduces a tool-based visual evidence paradigm that generates structured visual evidence through specialized analysis tools, enabling perceptually grounded and evidence-referenced quality reasoning.

Although this transition promises enhanced interpretability through textual rationales and improved generalization via semantic priors, we identify a fundamental limitation shared by existing MLLM-based IQA methods: their quality assessments rely on internally represented, semantically biased visual features, rather than explicit, verifiable evidence of low-level degradation. This limitation manifests across two predominant reasoning modalities, as illustrated in Fig. 1. (1) **Text-only Reasoning.** These methods [21, 51] generate structured quality descriptions to interpret and predict image quality. However, the reasoning process relies on the MLLM's internal representations and pre-trained semantic priors. (2) **Region-based Visual Reasoning.** To address the lack of visual grounding in text-only reasoning, recent methods [22, 24] introduce cropped or zoomed regions of the original image during multi-step reasoning, aiming to guide the model to focus on potentially degraded areas. However, these regions are essentially unstructured pixel patches that lack explicit association with specific perceptual attributes, and thus cannot form visual evidence to support the reasoning process.

To address this fundamental gap, we introduce IQA-T1, a novel framework that instantiates a *visual evidence reasoning* paradigm for IQA. The core insight is to augment the MLLM's reasoning process with explicit and structured visual evidence generated by external specialized tools. During inference, the model autonomously invokes tools from a predefined library, where each tool is designed to isolate a specific perceptual attribute, to generate intermediate visual representations such as noise residual maps, gradient magnitude distributions, and Fourier magnitude spectra. These evidence images are then incorporated into the reasoning chain as additional visual observations. This process trans-

forms quality assessment from a purely semantic inference task into a visually grounded and evidence-driven procedure.

Realizing this paradigm requires equipping the model with two complementary capabilities: understanding *how* to invoke tools and learning *when* to do so strategically. We address this through a two-stage training pipeline. First, supervised fine-tuning (SFT) instills foundational behaviors: structured reasoning, standardized tool invocation syntax, and the ability to associate visual evidence with quality judgments. Second, reinforcement learning (RL) with a carefully designed reward function optimizes the tool invocation policy. The overall training reward consists of a format reward, a scoring reward, a tool usage reward, and a tool repetition penalty, which together improve scoring accuracy while suppressing redundant or improper tool calls.

Since no existing IQA dataset supports visual evidence reasoning, we construct Q-Tool, containing 11k multimodal reasoning chains grounded in tool-generated visual evidence. The dataset is built through a three-stage pipeline: tool library construction, visual evidence reasoning chain generation, and reasoning chain verification.

We conduct comprehensive evaluations on seven diverse IQA benchmarks. Experimental results show that IQA-T1 achieves the best overall performance, with an average PLCC / SRCC of 0.795 / 0.784, surpassing traditional deep learning-based approaches and recent MLLM-based methods. Further ablation studies demonstrate that incorporating structured visual evidence significantly improves the stability and accuracy of quality prediction, validating the effectiveness of the visual evidence reasoning framework in enhancing both assessment reliability and interpretability. In summary, our contributions are threefold:

- We propose IQA-T1, a tool-based visual evidence reasoning framework for image quality assessment. It enables MLLMs to autonomously invoke specialized tools, dynamically constructing a structured evidence base that grounds quality assessment in explicit perceptual representations.
- We introduce Q-Tool, containing over 11k high-quality multimodal reasoning chains with standardized tool invocation formats and rigorous verification.
- We conduct extensive evaluations of IQA-T1 on multiple IQA benchmarks, demonstrating its SOTA performance compared to both traditional IQA approaches and recent MLLM-based baselines.

2 Related Work

2.1 Image Quality Assessment

Image Quality Assessment (IQA) aims to predict perceptual image quality in alignment with human subjective judgments. Traditional methods model quality degradation and output scores under either full-reference (FR-IQA) [48] or no-reference (NR-IQA) [31,33,34] settings. With the rapid advancement of multimodal large language models (MLLMs) [1,20,43,53], researchers have leveraged their cross-modal understanding to extend IQA from discriminative regression

toward more generalizable, language-driven frameworks. Existing MLLM-based IQA research can be taxonomized into three evolving paradigms: (1) **Score-Oriented Paradigm.** Early MLLM-based methods formulate quality prediction as classification, distribution regression, or contrastive learning to improve numerical accuracy [29, 45, 50, 56, 63]. While these approaches achieve strong correlation with human judgments, their outputs are limited to numerical values, offering no interpretability regarding why an image receives a particular score. This opacity limits their utility in applications requiring explainable decisions. (2) **Description-Oriented Paradigm.** To enhance interpretability, description-oriented methods [5, 49, 57, 58] generate structured or fine-grained natural language quality analyses. By producing textual descriptions of perceived distortions, these models provide insight into their quality judgments. However, they typically rely on supervised fine-tuning with text descriptions constructed from synthetic distortions, which fail to capture the diversity and complexity of real-world degradation patterns. Moreover, their training objectives prioritize textual coherence over numerical accuracy, often resulting in unstable or imprecise quality scores. (3) **Textual Reasoning Paradigm.** Recent studies have introduced explicit reasoning processes, forming a "reason-then-score" joint optimization framework. Exemplified by Q-Insight [21] and VisualQuality-R1 [51], these methods generate step-by-step textual rationales before predicting quality scores, often incorporating reinforcement learning to align reasoning with scoring. Despite this advancement, their quality judgments remain grounded in the MLLM's internal representations, which are inherently biased toward high-level semantics. As argued in our introduction, this semantic bias leaves these methods ill-equipped to capture low-level degradation cues, and their purely textual reasoning lacks verifiable visual anchors, rendering the generated rationales potentially disconnected from actual visual evidence.

2.2 Thinking with Images

The recognition that text-only reasoning lacks visual grounding has motivated a broader exploration of paradigms where images actively participate in the reasoning process. Termed "Thinking with Images" [36], this emerging direction transforms visual information from passive input into a dynamic cognitive workspace. Recent work can be categorized into three progressive levels [36]: (1) tool-driven visual exploration, where models invoke external tools such as cropping and zooming to acquire fine-grained visual evidence [27, 44, 47, 59, 62]; (2) programmatic visual manipulation, where models generate executable code to perform structured edits on images [7, 30, 46]; and (3) intrinsic visual imagination, where models generate new visual content as intermediate reasoning steps [4, 12, 18, 61]. These explorations mark the evolution from "thinking about images" to "thinking with images."

Thinking with Images for IQA. Applying this paradigm to IQA represents a natural progression toward addressing the lack of visual evidence in text-only reasoning. Methods such as Zoom-IQA [24] and Q-Probe [22] iteratively crop or zoom into evidence regions during inference, guiding the model to focus on

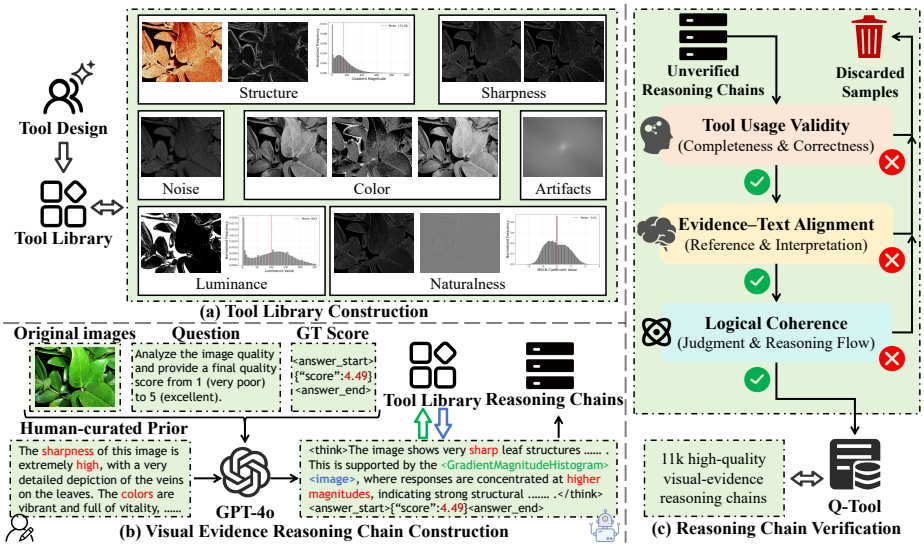

Fig. 2: Overview of the Q-Tool dataset construction pipeline, consisting of three key stages: (a) Perceptual tool library construction for visual evidence generation; (b) Visual evidence-grounded reasoning chain synthesis guided by human-curated priors; (c) Multi-level verification of reasoning chains.

locally degraded areas. While these approaches represent important steps, they remain fundamentally limited: the introduced regions are unstructured pixel patches that still rely on the model's semantic interpretation. They do not generate explicit, structured visual evidence that directly highlights specific perceptual attributes such as noise characteristics, frequency anomalies, or structural coherence. Consequently, they fail to bridge the representational gap between semantic reasoning and low-level degradation, which motivates our proposed IQA-T1 framework based on tool-generated visual evidence.

3 Dataset

Existing MLLM-based IQA methods suffer from a fundamental limitation: their quality assessments rely on semantically biased internal representations rather than explicit, verifiable evidence of low-level degradation. Addressing this gap requires a dataset that explicitly models the relationship between perceptual degradation and structured visual evidence within a reasoning framework. However, no such dataset currently exists. To enable the training of visual evidence reasoning models, we introduce Q-Tool —the first dataset specifically designed for tool-based visual evidence reasoning in IQA. It comprises 11k high-quality multimodal reasoning chains, each integrating explicit tool-generated visual evidence with human-aligned textual rationales.

The construction of Q-Tool is guided by three requirements: (1) the dataset must include explicit visual evidence that captures low-level perceptual attributes; (2) the reasoning chains must systematically integrate this evidence with textual analysis; and (3) the overall structure must support supervised fine-tuning of MLLMs to learn evidence-grounded reasoning behaviors. To meet these requirements, we design a systematic three-stage construction pipeline, illustrated in Fig. 2. The following sections describe each stage in detail.

3.1 Tool Library Construction

The foundation of visual evidence reasoning is the ability to generate explicit representations of low-level perceptual attributes. However, as argued in Sec. 1, MLLM internal representations are inherently biased toward high-level semantics and fail to capture degradation cues such as noise, blur, or frequency anomalies. To overcome this limitation, we construct a library of specialized perceptual tools, each designed to isolate and visualize a specific quality-relevant attribute.
Design rationale. We identify seven perceptual attributes critical for IQA, including structure, sharpness, noise, artifacts, luminance, color, and naturalness. Based on these attributes, we design 15 corresponding types of visual evidence. Detailed descriptions are provided in the supplementary material. Each tool T_k maps an input image I to a structured visual representation $e_k = \psi(T_k(I))$ that highlights degradation patterns invisible to standard MLLM representations. For example, the *NoiseResidualMap* isolates sensor or compression noise by subtracting a denoised version from the original; the *FourierMagnitudeSpectrum* reveals periodic artifacts through frequency-domain analysis; and *GradientOrientationCoherenceMap* exposes structural inconsistencies caused by blur or over-smoothing.
Implementation considerations. To ensure stable training and inference, all tools share three properties. First, they adopt unified invocation interfaces, allowing the MLLM to call any tool using a consistent syntax. Second, outputs are moderately downsampled to balance spatial structure preservation with visual token efficiency. Third, all tools are deterministic, guaranteeing identical outputs across data construction, supervised fine-tuning, and reinforcement learning stages—a critical requirement for consistent policy learning.

This tool library provides the perceptual foundation for subsequent reasoning chain construction, enabling the explicit representation of degradation cues that semantic MLLM representations inherently lack.

3.2 Visual Evidence Reasoning Chain Construction

With the tool library established, we next construct multimodal reasoning chains that systematically integrate visual evidence into quality assessment. A reasoning chain consists of (i) explicit tool invocations at appropriate analytical stages, (ii) interpretation of the generated visual evidence, and (iii) a final quality judgment logically derived from the accumulated evidence.

Challenge of naive generation. Directly prompting large language models to generate such chains autonomously leads to semantic hallucinations, logical discontinuities, and improper tool usage. The resulting chains lack the structured, verifiable relationship between evidence and judgment required for effective supervision.

Human-curated reasoning templates. To address this, we first develop human-curated reasoning templates that encode expert knowledge about perceptual analysis. These templates define: the logical decomposition of quality assessment into stages (e.g., "examine structure," "analyze noise," "assess color fidelity"); the mapping from perceptual attributes to appropriate tools; and the causal relationships between observed evidence and final quality judgment. These templates serve as semantic constraints, ensuring that generated chains follow human-aligned perceptual logic.

Constrained generation with GPT-4o. Under these constraints, we employ GPT-4o to generate complete multimodal reasoning chains. For each image, the model receives: the image itself, the corresponding reasoning template, and access to the tool library. At each designated stage, the model inserts explicit tool invocations, and the resulting visual evidence is integrated into subsequent analysis. This process transforms purely textual reasoning into an evidence-grounded procedure where each inference step relies on observable visual signals.

By combining human perceptual priors with controlled model generation, we produce reasoning chains that are both logically structured and grounded in verifiable evidence.

3.3 Reasoning Chain Verification

The reliability of Q-Tool as a training resource depends on the quality of its constituent reasoning chains. We therefore conduct systematic multi-level verification to ensure consistent alignment among tool usage, visual evidence, and reasoning logic.

Verification protocol. Each generated chain is examined along three dimensions. (1) **Tool appropriateness**: Are tool invocations aligned with the intended reasoning stage and analytical purpose? We verify that tools are neither redundant nor improperly used. (2) **Evidence grounding**: Is the generated visual evidence correctly referenced, interpreted, and integrated into the textual reasoning? We ensure that each evidence image serves a genuine analytical function rather than appearing as a decorative element. (3) **Logical consistency**: Is the final quality judgment sufficiently supported by the preceding evidence accumulation? We reject chains where conclusions lack adequate evidentiary support or exhibit reasoning discontinuities.

Quality refinement. Chains failing any verification criterion are removed. A subset of borderline cases undergoes manual refinement to correct minor inconsistencies while preserving the overall structure. This rigorous verification process yields a final dataset of 11k high-quality reasoning chains, each exhibiting: (i) appropriate tool utilization, (ii) coherent integration of visual evidence, and (iii) logically sound quality judgments.

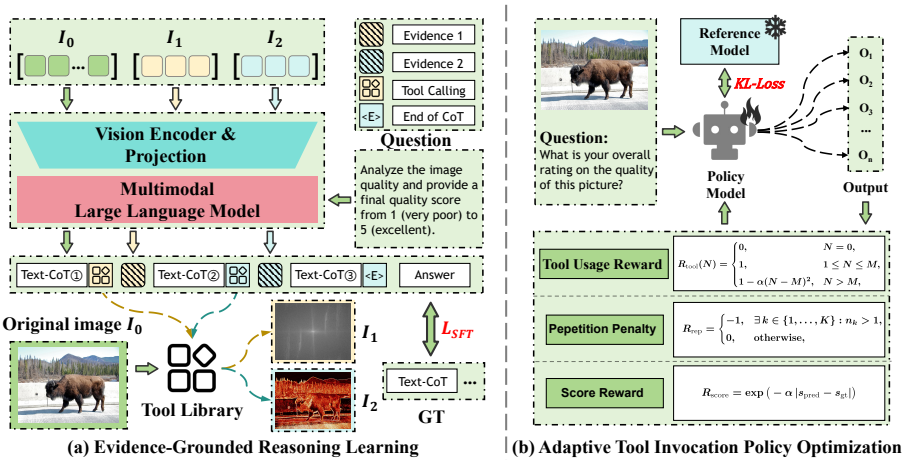

Fig. 3: Overview of the proposed IQA-T1 framework. The model operates in two stages: (1) Supervised fine-tuning (SFT) on the Q-Tool dataset to learn evidence-grounded reasoning and tool invocation syntax; (2) Reinforcement learning (RL) with a tailored reward function to optimize the adaptive tool invocation policy. During inference, the model autonomously invokes tools to acquire visual evidence $\mathbf{E}_r$, which is integrated into the reasoning chain to ground quality assessment.

The resulting Q-Tool dataset provides the necessary supervision for learning evidence-grounded reasoning behaviors in the subsequent two-stage training pipeline (Sec. 4.3 and Sec. 4.3), serving as the foundation for the visual evidence reasoning paradigm introduced in this work.

4 Methodology

In this section, we first formalize the problem of MLLM-based IQA and highlight the semantic bias inherent in standard visual representations (Sec. 4.1). We then introduce the proposed visual evidence reasoning framework (Sec. 4.2) and describe the two-stage training pipeline comprising supervised fine-tuning (Sec. 4.3) and reinforcement learning (Sec. 4.3) to equip the model with both foundational reasoning capabilities and an adaptive tool invocation policy.

4.1 Problem Formulation

MLLM-based IQA as Sequence Generation. We formalize image quality assessment using a multimodal large language model (MLLM) as a conditional sequence generation problem. Let $I \in \mathbb{R}^{H\times W\times 3}$ denote the input image and P the textual prompt (e.g., "Assess the quality of this image"). The target output is a sequence $Y = \{y_1, y_2, \ldots, y_T\}$ that includes both a reasoning chain and a final quality score $s \in \mathbb{R}$ (typically normalized to $[1, 5]$). A standard MLLM

comprises a visual encoder $\mathcal{E}_\phi$, a modality projector $\mathcal{M}$, and a large language model (LLM) $\mathcal{G}_\theta$. The image is first encoded and projected into a sequence of visual tokens aligned with the textual embedding space: $\mathbf{z}_v = \mathcal{M}\big(\mathcal{E}_\phi(I)\big) \in \mathbb{R}^{L\times d}$, where L is the number of visual tokens and d the embedding dimension. The inference process then models the posterior distribution of the output sequence conditioned on the multimodal inputs:

$$p(Y \mid I, P) = \prod_{t=1}^{T} p_\theta(y_t \mid \mathbf{z}_v, \mathbf{h}_p, y_{<t}), \tag{1}$$

with $\mathbf{h}_p$ denoting the prompt embeddings and $y_{<t}$ the previously generated tokens.

Semantic Bias in Visual Representations. A growing body of work [3, 23] has shown that the visual representation $\mathbf{z}_v$ produced by MLLMs is heavily biased toward high-level semantic content, while offering limited sensitivity to low-level perceptual degradations critical for IQA, such as noise, blur, or compression artifacts. Formally, consider an image I and its perceptually degraded counterpart I_d (e.g., with added noise). Although human observers would assign significantly different quality scores $s(I)$ and $s(I_d)$, the corresponding visual representations often remain nearly identical:

$$\|\mathbf{z}_v(I) - \mathbf{z}_v(I_d)\|_2 \approx 0, \qquad |s(I) - s(I_d)| \gg \epsilon, \tag{2}$$

where ϵ is a small tolerance. This observation reveals that $\mathbf{z}_v$ alone is insufficient to capture the degradation cues necessary for accurate quality assessment, motivating the introduction of complementary perceptual evidence.

4.2 Visual Evidence Reasoning

To compensate for the information loss in $\mathbf{z}_v$, we propose a framework that augments the MLLM's reasoning process with explicit, structured visual evidence generated by external tools. Let $\mathcal{T} = \{T_k\}_{k=1}^{K}$ denote a library of specialized perceptual analysis tools, each mapping an image to a degradation-oriented feature space: $T_k : \mathbb{R}^{H\times W\times 3} \to \mathcal{D}_k$, where $\mathcal{D}_k$ is the output domain (e.g., a residual map, histogram, or spectrum). The set of all possible visual evidence for image I is defined as

$$\mathbf{E} = \{e_k \mid e_k = \psi(T_k(I)),\ k = 1, \ldots, K\}, \tag{3}$$

with $\psi(\cdot)$ projecting tool outputs into the LLM's context space (e.g., downsampling and tokenization).

During inference, the model does not access the entire set $\mathbf{E}$ upfront. Instead, it dynamically constructs a subset $\mathbf{E}_r \subseteq \mathbf{E}$ by invoking relevant tools at appropriate reasoning steps. The generation process thus conditions on both the original visual tokens and the progressively acquired evidence:

$$\hat{Y} = \arg\max_{Y} \sum_{t=1}^{T} \log p_\theta(y_t \mid \mathbf{z}_v, \mathbf{h}_p, \mathbf{E}_r, y_{<t}). \tag{4}$$

The evidence $\mathbf{E}_r$ serves as a dynamically acquired perceptual complement to $\mathbf{z}_v$, enabling the model to ground each reasoning step in verifiable, low-level visual cues. This formulation transforms IQA from a purely semantic inference task into an evidence-driven reasoning process.

4.3 Two-Stage Training Pipeline

Realizing the visual evidence reasoning paradigm requires equipping the MLLM with two complementary capabilities: (i) understanding *how* to invoke tools and interpret the resulting evidence, and (ii) learning *when* to invoke tools strategically to maximize assessment accuracy while avoiding redundancy. We address these via a two-stage pipeline: supervised fine-tuning (SFT) on the Q-Tool dataset establishes foundational reasoning behaviors, followed by reinforcement learning (RL) that optimizes the adaptive tool invocation policy.

Stage I: Evidence-Grounded Reasoning Learning

In this stage, we train the model on the Q-Tool dataset to learn structured, evidence-grounded reasoning chains. Each training sample consists of the original image I, a set of tool-generated visual evidence V (corresponding to the ground-truth invocations), and a textual reasoning chain T that integrates the evidence and culminates in a quality score s_{gt}.

Crucially, the model is not required to generate the visual evidence itself, as the evidence is deterministically produced by the tools and provided as input. During SFT, we apply loss masking to the tokens representing the evidence images, as they are not prediction targets. The objective is to maximize the likelihood of the reasoning tokens conditioned on the complete context:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{L} \log P_\theta(y_t \mid y_{<t}, I, V, T), \tag{5}$$

where y_t denotes the t-th ground-truth token of the reasoning sequence (including the final score) and L is its length. Optimizing Eq. (5) instills three essential behaviors: (i) the ability to follow structured reasoning templates, (ii) correct syntax for tool invocation, and (iii) the capacity to associate visual evidence with quality judgments. This stage initializes a policy π_{SFT} that serves as the foundation for subsequent RL fine-tuning.

Stage II: Adaptive Tool Invocation Policy Optimization

While SFT establishes correct reasoning patterns, it does not explicitly optimize the *strategy* of when to invoke which tool. To learn an adaptive invocation policy, we employ reinforcement learning with a reward function designed to balance prediction accuracy, evidence efficiency, and reasoning stability.

Reinforcement Learning Setup. We adopt Group Relative Policy Optimization (GRPO) [9], a variant of PPO that estimates advantages from within-group comparisons, eliminating the need for a separate value network [41]. Given a prompt x, the current policy π_θ samples a group of n reasoning chains $\{y_1, \ldots, y_n\}$. The objective is:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D},\, y_i\sim\pi_\theta(\cdot|x)}\Big[\hat{A}_i \log \pi_\theta(y_i|x)\Big] + \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\,, \tag{6}$$

where $\hat{A}_i = (R_i - \mu_R)/\sigma_R$ is the normalized advantage within the group, π_{ref} is the reference policy from Stage I, and β controls the KL penalty to prevent catastrophic forgetting.

Reward Design. The reward function $R(x, y)$ is a weighted combination of four components:

(1) **Format reward** R_{fmt}: enforces that the output adheres to the expected structured format (e.g., contains tool invocation tags and a final score). It is 1 if the format is correct, 0 otherwise.

(2) **Scoring reward** R_{score}: measures accuracy of the predicted quality score s_{pred} relative to the ground truth s_{gt}:

$$R_{\text{score}} = \exp\big(-\alpha\,|s_{\text{pred}} - s_{\text{gt}}|\big), \tag{7}$$

with $\alpha > 0$ controlling the sensitivity.

(3) **Tool usage reward** R_{tool}: encourages efficient evidence acquisition by rewarding an appropriate number of tool invocations. Let $N = |\mathbf{E}_r|$ be the number of distinct evidence items collected. Then

$$R_{\text{tool}}(N) = \begin{cases} 0, & N = 0, \\ 1, & 1 \le N \le M, \\ 1 - \gamma(N - M)^2, & N > M, \end{cases} \tag{8}$$

where M is a predefined maximum (set to 4 in our experiments) and γ controls the penalty for excessive invocations.This reward is intentionally designed as a weak constraint rather than a strict tool-level supervision signal, since real-world IQA images often involve mixed degradations and may admit multiple valid tool combinations. Therefore, R_{tool} encourages the model to acquire sufficient visual evidence while preserving flexibility in adaptive tool selection.

(4) **Repetition penalty** R_{rep}: prevents degenerate behavior where the model repeatedly invokes the same tool without gaining new information:

$$R_{\text{rep}} = \begin{cases} -1, & \exists\, k \in \{1, \ldots, K\} : n_k > 1, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where n_k is the number of invocations of tool T_k.

The overall reward is

$$R = \lambda_{\text{fmt}} R_{\text{fmt}} + \lambda_{\text{score}} R_{\text{score}} + \lambda_{\text{tool}} R_{\text{tool}} + \lambda_{\text{rep}} R_{\text{rep}}, \tag{10}$$

with weights λ balancing the contributions (set empirically as $\lambda_{\text{fmt}} = 1$, $\lambda_{\text{score}} = 5$, $\lambda_{\text{tool}} = 1$, $\lambda_{\text{rep}} = 2$). Through this reward formulation, the RL stage refines the tool invocation policy to maximize scoring accuracy while maintaining efficient and non-redundant use of perceptual evidence, ultimately yielding a model whose reasoning is both accurate and interpretable.

Table 1: PLCC/SRCC comparison on score regression tasks between our method and other competitive IQA methods. All methods except handcrafted ones are trained on the KonIQ dataset. AVG. denotes the average PLCC/SRCC across all evaluation datasets. The best result is bolded and the second best is underlined.

Methods	Wild Images			Synthetic Distortion			AI-Generated	AVG.
	KonIQ [11]	SPAQ [6]	LiveW [8]	KADID [25]	PIPAL [14]	CSIQ [16]	AGIQA [19]	
Handcrafted Methods								
NIQE [34]	0.533 / 0.530	0.679 / 0.664	0.493 / 0.449	0.468 / 0.405	0.195 / 0.161	0.718 / 0.628	0.560 / 0.533	0.521 / 0.481
BRISQUE [33]	0.225 / 0.226	0.490 / 0.406	0.361 / 0.313	0.429 / 0.356	0.267 / 0.232	0.740 / 0.556	0.541 / 0.497	0.436 / 0.369
Non-MLLM Methods								
NIMA [37]	0.896 / 0.859	0.838 / 0.856	0.814 / 0.771	0.532 / 0.535	0.390 / 0.399	0.695 / 0.649	0.715 / 0.654	0.697 / 0.675
HyperIQA [35]	0.917 / 0.906	0.791 / 0.788	0.772 / 0.749	0.506 / 0.468	0.410 / 0.403	0.752 / 0.717	0.702 / 0.640	0.693 / 0.667
DBCNN [60]	0.884 / 0.875	0.812 / 0.806	0.773 / 0.755	0.497 / 0.484	0.384 / 0.381	0.586 / 0.572	0.730 / 0.641	0.667 / 0.645
MUSIQ [15]	0.924 / 0.929	0.868 / 0.863	0.789 / 0.830	0.575 / 0.556	0.431 / 0.431	0.771 / 0.710	0.722 / 0.630	0.726 / 0.707
ManIQA [55]	0.849 / 0.834	0.768 / 0.758	0.849 / 0.832	0.499 / 0.465	0.457 / 0.452	0.623 / 0.627	0.723 / 0.636	0.681 / 0.658
MLLM-based Methods (w/o reasoning)								
CLIP-IQA+ [45]	0.909 / 0.895	0.866 / 0.864	0.832 / 0.805	0.653 / 0.654	0.427 / 0.419	0.772 / 0.719	0.736 / 0.685	0.742 / 0.720
C2Score [63]	0.923 / 0.910	0.867 / 0.860	0.786 / 0.772	0.500 / 0.453	0.354 / 0.342	0.735 / 0.705	0.777 / 0.671	0.706 / 0.673
Q-Align [50]	0.941 / 0.940	0.886 / 0.887	0.853 / 0.860	0.674 / 0.684	0.403 / 0.419	0.671 / 0.737	0.772 / 0.735	0.743 / 0.752
DeQA [56]	**0.953** / **0.941**	0.895 / 0.896	**0.892** / **0.879**	0.694 / 0.687	0.472 / 0.478	0.787 / 0.744	0.809 / 0.729	0.786 / 0.765
MLLM-based Methods (w/ reasoning)								
Q-Insight [21]	0.918 / 0.895	0.903 / 0.903	0.870 / 0.839	0.702 / 0.702	0.458 / 0.435	0.685 / 0.640	0.816 / **0.766**	0.765 / 0.740
VisualQuality-R1 [51]	0.910 / 0.896	0.889 / 0.892	0.856 / 0.827	**0.703** / **0.712**	0.451 / 0.441	0.768 / 0.707	**0.817** / 0.760	0.771 / 0.748
Zoom-IQA [24]	0.938 / 0.922	0.902 / 0.900	0.887 / 0.870	0.701 / 0.700	0.468 / 0.465	0.797 / 0.754	0.816 / 0.765	0.787 / 0.768
IQA-T1	0.942 / 0.925	**0.905** / **0.908**	0.888 / 0.860	0.686 / 0.708	**0.480** / **0.490**	**0.850** / **0.840**	0.816 / 0.754	**0.795** / **0.784**

5 Experiments

5.1 Experimental Setup

Datasets and Metrics. To evaluate the robustness of IQA-T1 under diverse degradations, we follow prior works by training the model on KonIQ [11] and evaluating its cross-distribution generalization on six IQA benchmarks. KonIQ [11], SPAQ [6], and LIVE-Wild [8] represent real-world distortions; KADID [25] and CSIQ [16] contain synthetic distortions; and PIPAL [14] focuses on algorithm-induced degradations. We further include AGIQA [19] to evaluate AI-generated content. Following common IQA practice, performance is measured using PLCC and SRCC.

Implementation Details. We build our framework upon Qwen3-VL-4B and adopt a two-stage training pipeline. In the first stage, we perform SFT on our Q-Tool datasets using cross-entropy loss. Training is conducted with a batch size of 1, 16 gradient accumulation steps, a learning rate of 1×10^{-5}, and a warm-up ratio of 0.05 for two epochs. In the second stage, we apply GRPO to further align the model with perceptual quality preferences. The GRPO training initializes from the SFT checkpoint and uses a global batch size of 12 with a micro-batch size of 1 per GPU. The learning rate is set to 1×10^{-6}, and each prompt generates $N = 8$ rollout samples. Training is performed on six NVIDIA RTX 4090 GPUs, and the KL-penalty coefficient is fixed at 1×10^{-2}.

Baseline Methods. We evaluate IQA-T1 on the image quality score regression task against four categories of baselines. (1) Traditional handcrafted IQA

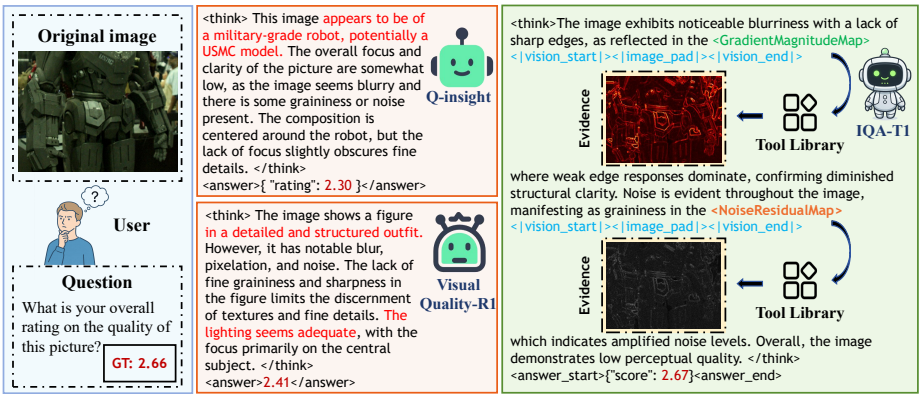

Fig. 4: Qualitative comparison of IQA-T1 with competing methods. We highlight: incorrect descriptions, and the visual evidence reasoning unique to our model.

metrics, including NIQE [34] and BRISQUE [33]. (2) Deep learning-based IQA models such as NIMA [37], HyperIQA [35], DBCNN [60], MUSIQ [15], and ManIQA [55]. (3) MLLM-based score prediction methods, including CLIP-IQA+ [45], C2Score [63], Q-Align [50], and DeQA [56]. (4) MLLM-based frameworks with reasoning, such as Q-Insight [21], VisualQuality-R1 [51], and Zoom-IQA [24].

5.2 Main Results

Table 1 presents a comprehensive comparison of PLCC and SRCC across all methods and datasets. Our proposed IQA-T1 achieves the highest average PLCC (0.795) and SRCC (0.784) across the seven datasets.

Comparison with MLLM-based methods without reasoning. Methods such as Q-Align [50] and DeQA [56] focus exclusively on score regression and achieve strong performance, particularly on in-distribution datasets like KonIQ [11]. However, they lack the ability to explain their judgments. IQA-T1 not only remains highly competitive in score accuracy but also generates interpretable, evidence-grounded reasoning chains, offering a unique combination of precision and transparency.

Comparison with MLLM-based reasoning methods. Among methods that produce both scores and rationales, IQA-T1 exhibits strong cross-distribution generalization (Qualitative examples are shown in Fig. 4). It demonstrates robust performance on real-world images (e.g., SPAQ [6]), establishes new highest results on challenging datasets for typical synthetic and algorithm-induced distortions (CSIQ [16] and PIPAL [14], respectively), and remains highly competitive on AI-generated content (AGIQA [19]).

These results provide strong empirical evidence for the central thesis of this work: by augmenting MLLM reasoning with explicit, tool-generated visual evidence, IQA-T1 effectively reduces the semantic bias inherent in standard MLLM representations.

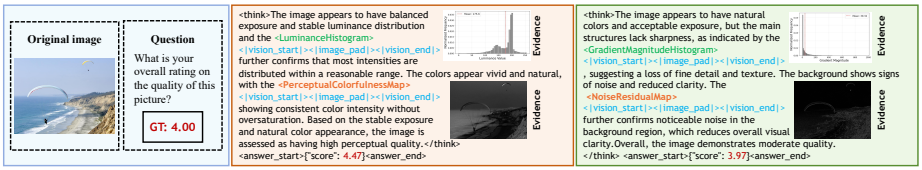

Fig. 5: Qualitative comparison of IQA-T1 with competing methods.

Table 2: Ablation studies of individual components measured by PLCC and SRCC, where all models are trained using the KonIQ dataset.

Tool	SFT	R_{tool}	R_{rep}	KonIQ [11]	SPAQ [6]	LiveW [8]	KADID [25]	PIPAL [14]	CSIQ [16]	AGIQA [19]	AVG.
×	✓	×	×	0.843 / 0.817	0.848 / 0.836	0.789 / 0.742	0.621 / 0.603	0.412 / 0.418	0.675 / 0.640	0.747 / 0.701	0.705 / 0.680
✓	✓	×	×	0.915 / 0.901	0.891 / 0.890	0.826 / 0.774	0.660 / 0.705	0.464 / 0.455	0.768 / 0.730	0.770 / 0.712	0.756 / 0.738
✓	✓	✓	×	0.933 / 0.917	0.903 / 0.904	0.881 / 0.858	0.684 / 0.704	0.479 / 0.481	0.841 / 0.835	0.811 / 0.739	0.790 / 0.777
✓	✓	×	✓	0.935 / 0.919	0.902 / 0.902	0.874 / 0.840	**0.689** / **0.708**	**0.481** / 0.485	0.838 / 0.832	0.804 / 0.742	0.789 / 0.775
✓	✓	✓	✓	**0.942** / **0.925**	**0.905** / **0.908**	**0.888** / **0.860**	0.686 / **0.708**	0.480 / **0.490**	**0.850** / **0.840**	**0.816** / **0.754**	**0.795** / **0.784**

5.3 Ablation Studies and Analysis

We provide a set of complementary analyses to better understand the effectiveness and generality of IQA-T1. Table 2 isolates the contribution of individual components, Table 3 compares different tool invocation strategies and their inference costs, and Table 4 evaluates the robustness of the proposed framework across different MLLM backbones.

Effectiveness of Visual Evidence. We first examine the effect of tool-generated visual evidence by comparing two SFT variants: one where tool calls return only placeholders without actual evidence (Table 2 Row 1), and another where visual evidence is available during reasoning (Table 2 Row 2). Since both variants share identical training objectives and reasoning templates, the difference isolates the impact of visual evidence. The results show consistent improvements across all datasets, indicating that structured visual evidence effectively compensates for the low-level information loss in $\mathbf{z}_v$ and mitigates the semantic bias discussed in Sec. 4.1.

Effectiveness of Dynamic Tool Invocation. We compare dynamic tool invocation with all, fixed, and random tool selection strategies in Table 3. Compared with the base Qwen3-VL-4B, tool-generated evidence brings clear gains with moderate inference overhead. However, more evidence is not necessarily better, as irrelevant tools may introduce redundant visual tokens and distracting cues. In contrast, dynamic invocation adaptively selects relevant evidence for each image, achieving the best performance with fewer tool calls. The qualitative case study in Fig. 5 further shows that tools aligned with dominant degradations provide stronger support for quality judgment, indicating that IQA-T1 benefits from evidence reasoning rather than simple score fitting.

Effectiveness of Reward Design. We next analyze the contribution of the RL-stage reward components. Starting from the full model, we ablate either the tool usage reward R_{tool} or the repetition penalty R_{rep} in Table 2. Removing

Table 3: Ablation study of different tool invocation strategies on the KonIQ dataset.

Method	PLCC	SRCC	Avg. Tools	Avg. Time(s)	Peak Mem.(GB)
Qwen3-VL-4B	0.785	0.744	0.00	**4.62**	**8.52**
IQA-T1 (All Tools)	0.865	0.841	15.00	6.06	9.13
IQA-T1 (Fixed-K)	0.895	0.869	3.00	5.32	8.70
IQA-T1 (Random-K)	0.871	0.854	3.00	5.42	8.74
IQA-T1 (Dynamic)	**0.942**	**0.925**	2.34	5.09	8.67

Table 4: Generalization of IQA-T1 across different MLLM backbones. Results are reported as PLCC/SRCC.

Method	KonIQ [11]	SPAQ [6]	LiveW [8]	KADID [25]	PIPAL [14]	CSIQ [16]	AGIQA [19]
IQA-T1 (Qwen3-VL-2B)	0.929 / 0.910	0.892 / 0.898	0.851 / 0.800	0.654 / 0.700	0.471 / 0.470	0.812 / 0.773	0.792 / 0.723
IQA-T1 (InternVL3.5-4B)	0.898 / 0.897	0.881 / 0.875	0.842 / 0.821	**0.701** / **0.719**	0.443 / 0.437	0.749 / 0.733	**0.822** / **0.774**
IQA-T1 (Qwen3-VL-4B)	**0.942** / **0.925**	**0.905** / **0.908**	**0.888** / **0.860**	0.686 / 0.708	**0.480** / **0.490**	**0.850** / **0.840**	0.816 / 0.754

either component leads to degraded performance, indicating that both rewards are necessary for stable tool-based reasoning. Specifically, R_{tool} encourages the model to acquire sufficient perceptual evidence, while R_{rep} suppresses redundant or repetitive tool calls. Their combination enables the model to balance evidence sufficiency and evidence diversity, leading to more reliable quality prediction.

Generality across Backbones. To examine whether the proposed framework depends on a specific MLLM architecture, we instantiate IQA-T1 with different backbones, including Qwen3-VL-2B, InternVL3.5-4B, and Qwen3-VL-4B. As shown in Table 4, IQA-T1 maintains strong performance across different model scales and architectures. Although the absolute performance varies with backbone capacity and pretraining characteristics, the overall trend remains consistent across real-world, synthetic, algorithm-induced, and AI-generated image quality benchmarks. These results suggest that tool-based visual evidence reasoning is a general mechanism for enhancing MLLM-based IQA, rather than a backbone-specific design.

6 Conclusion

In this work, we introduced IQA-T1, a tool-based visual evidence reasoning framework that grounds image quality assessment in explicit perceptual evidence rather than semantically biased internal representations. By enabling MLLMs to autonomously invoke specialized analysis tools, our approach transforms IQA from purely semantic inference into an evidence-driven reasoning process. Supported by the proposed Q-Tool dataset and a two-stage training strategy combining supervised learning and reinforcement learning, IQA-T1 achieves the best average performance across multiple IQA benchmarks while providing interpretable reasoning grounded in visual evidence. We hope this work encourages future research on integrating structured visual evidence into multimodal reasoning systems, advancing more reliable and interpretable perception in multimodal large language models.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62472359, No. 62372379), in part by Xi'an's Key Industrial Chain Core Technology Breakthrough Project: AI Core Technology Breakthrough under Grand 24ZDCYJSGG0003, and in part by the Hong Kong University of Science and Technology under Grant No. WEB26EG02.

References

1. 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)
2. An, S., Xu, L., Deng, Z., Zhang, H.: Hfm: A hybrid fusion method for underwater image enhancement. Engineering applications of artificial intelligence **127**, 107219 (2024)
3. Chen, B., Pan, S., Wu, D., Xie, L., Sui, X., Zhu, L., Zhu, H.: Mitigating perception bias: A training-free approach to enhance lmm for image quality assessment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 2760–2768 (2026)
4. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
5. Chen, Z., Zhang, X., Li, W., Pei, R., Song, F., Min, X., Liu, X., Yuan, X., Guo, Y., Zhang, Y.: Grounding-iqa: Multimodal language grounding model for image quality assessment. arXiv preprint arXiv:2411.17237 **10** (2024)
6. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3677–3686 (2020)
7. Fu, X., Liu, M., Yang, Z., Corring, J., Lu, Y., Yang, J., Roth, D., Florencio, D., Zhang, C.: Refocus: Visual editing as a chain of thought for structured image understanding. arXiv preprint arXiv:2501.05452 (2025)
8. Ghadiyaram, D., Bovik, A.C.: Live in the wild image quality challenge database. Online: http://live. ece. utexas. edu/research/ChallengeDB/index. html [Mar, 2017] **2**(5), 6 (2015)
9. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
10. He, D., Yang, Z., Peng, W., Ma, R., Qin, H., Wang, Y.: Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5718–5727 (2022)
11. Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Transactions on Image Processing **29**, 4041–4056 (2020)
12. Jiang, D., Guo, Z., Zhang, R., Zong, Z., Li, H., Zhuo, L., Yan, S., Heng, P.A., Li, H.: T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. arXiv preprint arXiv:2505.00703 (2025)

13. Jiang, J., Zuo, Z., Wu, G., Jiang, K., Liu, X.: A survey on all-in-one image restoration: Taxonomy, evaluation and future trends. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
14. Jinjin, G., Haoming, C., Haoyu, C., Xiaoxing, Y., Ren, J.S., Chao, D.: Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In: European conference on computer vision. pp. 633–651. Springer (2020)
15. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021)
16. Larson, E.C., Chandler, D.M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. Journal of electronic imaging **19**(1), 011006–011006 (2010)
17. Li, B., Zhao, H., Wang, W., Hu, P., Gou, Y., Peng, X.: Mair: A locality-and continuity-preserving mamba for image restoration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7491–7501 (2025)
18. Li, C., Wu, W., Zhang, H., Xia, Y., Mao, S., Dong, L., Vulić, I., Wei, F.: Imagine while reasoning in space: Multimodal visualization-of-thought. arXiv preprint arXiv:2501.07542 (2025)
19. Li, C., Zhang, Z., Wu, H., Sun, W., Min, X., Liu, X., Zhai, G., Lin, W.: Agiqa-3k: An open database for ai-generated image quality assessment. IEEE Transactions on Circuits and Systems for Video Technology **34**, 6833–6846 (2023)
20. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
21. Li, W., Zhang, X., Zhao, S., Zhang, Y., Li, J., Zhang, L., Zhang, J.: Q-insight: Understanding image quality via visual reinforcement learning. arXiv preprint arXiv:2503.22679 (2025)
22. Li, X., Li, X., Wang, Y., He, X., Hu, Z., Yu, W., Xie, C.: Q-probe: Scaling image quality assessment to high resolution via context-aware agentic probing. arXiv preprint arXiv:2601.15356 (2026)
23. Li, Y., Sun, Z., Chen, Y.J., Nishida, S.: Investigate the low-level visual perception in vision-language based image quality assessment. arXiv preprint arXiv:2512.09573 (2025)
24. Liang, G., Wang, J., Wu, Z., Zhou, S.: Zoom-iqa: Image quality assessment with reliable region-aware reasoning. arXiv preprint arXiv:2601.02918 (2026)
25. Lin, H., Hosu, V., Saupe, D.: Kadid-10k: A large-scale artificially distorted iqa database. In: 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX). pp. 1–3. IEEE (2019)
26. Lin, Y., Lin, Z., Chen, H., Pan, P., Li, C., Chen, S., Wen, K., Jin, Y., Li, W., Ding, X.: Jarvisir: Elevating autonomous driving perception with intelligent image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22369–22380 (2025)
27. Liu, D., Wang, Z., Ruan, M., Luo, F., Chen, C., Li, P., Liu, Y.: Thinking with visual abstract: Enhancing multimodal reasoning via visual abstraction. arXiv preprint arXiv:2505.20164 (2025)
28. Liu, J., Sun, H., Katto, J.: Learned image compression with mixed transformer-cnn architectures. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14388–14397 (2023)
29. Liu, K., Zhang, Z., Li, W., Pei, R., Song, F., Liu, X., Kong, L., Zhang, Y.: Dog-iqa: Standard-guided zero-shot mllm for mix-grained image quality assessment. arXiv preprint arXiv:2410.02505 (2024)

30. Liu, Z., Zang, Y., Zou, Y., Liang, Z., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual agentic reinforcement fine-tuning. arXiv preprint arXiv:2505.14246 (2025)
31. Ma, C., Yang, C.Y., Yang, X., Yang, M.H.: Learning a no-reference quality metric for single-image super-resolution. Computer Vision and Image Understanding **158**, 1–16 (2017)
32. Ma, L., Jin, D., An, N., Liu, J., Fan, X., Luo, Z., Liu, R.: Bilevel fast scene adaptation for low-light image enhancement. International Journal of Computer Vision **133**(11), 8234–8252 (2025)
33. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012)
34. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)
35. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., 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. pp. 3667–3676 (2020)
36. Su, Z., Xia, P., Guo, H., Liu, Z., Ma, Y., Qu, X., Liu, J., Li, Y., Zeng, K., Yang, Z., et al.: Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. arXiv preprint arXiv:2506.23918 (2025)
37. Talebi, H., Milanfar, P.: Nima: Neural image assessment. IEEE transactions on image processing **27**(8), 3998–4011 (2018)
38. Tang, J., Chen, J., Zhai, Y., Wei, W., Liu, R., Zhao, M., Wu, X., Xiao, Q., Chen, Q.: Robust-u1: Can mllms self-recover corrupted visual content for robust understanding? In: Proceedings of the 43rd International Conference on Machine Learning (ICML) (2026)
39. Tang, J., Lu, H., Wu, R., Xu, X., Ma, K., Fang, C., Guo, B., Lu, J., Chen, Q., Chen, Y.C.: Hawk: Learning to understand open-world video anomalies. Advances in Neural Information Processing Systems **37**, 139751–139785 (2024)
40. Tang, J., Wu, R., Xu, X., Hu, S., Chen, Y.C.: Learning to remove wrinkled transparent film with polarized prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24987–24996 (June 2024)
41. Tang, J., Xia, Y., Wu, Y.F., Hu, Y., Yuhui, C., Chen, Q.G., Xu, X., Wu, X., Lu, H., Ma, Y., et al.: Lpo: Towards accurate gui agent interaction via location preference optimization. In: Findings of the Association for Computational Linguistics: ACL 2026. pp. 14617–14628 (2026)
42. Tang, J., Yan, Y., Wang, Q., Xia, Y., Geng, B., Chen, J., Ma, K., Zhai, Y., He, Q., Shao, W., Sun, Y., Dai, J., Chen, C., Xu, X., Yao, K., Zhang, L., Wei, W., Chen, Q., Plaza, A., Zhang, Y.: Intelligent remote sensing agents: A survey (2026), `https://github.com/PolyX-Research/Awesome-Remote-Sensing-Agents`
43. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
44. Wang, H., Su, A., Ren, W., Lin, F., Chen, W.: Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. arXiv preprint arXiv:2505.15966 (2025)
45. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2555–2563 (2023)

46. Wang, Q., Ding, R., Zeng, Y., Chen, Z., Chen, L., Wang, S., Xie, P., Huang, F., Zhao, F.: Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. arXiv preprint arXiv:2505.22019 (2025)
47. Wang, Y., Wang, S., Cheng, Q., Fei, Z., Ding, L., Guo, Q., Tao, D., Qiu, X.: Visuothink: Empowering lvlm reasoning with multimodal tree search. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 21707–21719 (2025)
48. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
49. 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 (2024)
50. 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 lmms for visual scoring via discrete text-defined levels. arXiv preprint arXiv:2312.17090 (2023)
51. Wu, T., Zou, J., Liang, J., Zhang, L., Ma, K.: Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. arXiv preprint arXiv:2505.14460 (2025)
52. Yan, Q., Feng, Y., Zhang, C., Pang, G., Shi, K., Wu, P., Dong, W., Sun, J., Zhang, Y.: Hvi: A new color space for low-light image enhancement. In: Proceedings of the computer vision and pattern recognition conference. pp. 5678–5687 (2025)
53. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
54. Yang, R., Mandt, S.: Lossy image compression with conditional diffusion models. Advances in Neural Information Processing Systems **36**, 64971–64995 (2023)
55. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., 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. pp. 1191–1200 (2022)
56. You, Z., Cai, X., Gu, J., Xue, T., 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 (2025)
57. You, Z., Gu, J., Cai, X., Li, Z., Zhu, K., Dong, C., Xue, T.: Enhancing descriptive image quality assessment with a large-scale multi-modal dataset. IEEE Transactions on Image Processing **34**, 8201–8215 (2025)
58. You, Z., Li, Z., Gu, J., Yin, Z., Xue, T., 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)
59. Zhang, G., Zhong, T., Xia, Y., Liu, M., Yu, Z., Li, H., He, W., She, D., Wang, Y., Jiang, H.: Cmmcot: Enhancing complex multi-image comprehension via multi-modal chain-of-thought and memory augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 12430–12438 (2026)
60. Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology **30**(1), 36–47 (2018)
61. Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al.: Cot-vla: Visual chain-of-thought reasoning for vision-language-

action models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1702–1713 (2025)
62. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing" thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025)
63. Zhu, H., Wu, H., Li, Y., Zhang, Z., Chen, B., Zhu, L., Fang, Y., Zhai, G., Lin, W., Wang, S.: Adaptive image quality assessment via teaching large multimodal model to compare. Advances in Neural Information Processing Systems **37**, 32611–32629 (2024)