Tool-IQA: Augmenting Image Quality Assessment with Simple Tools

Guanyi Qin[1,2,†], **Junjie Zhang**[3,†], **Chunming He**[4], **Yibing Fu**[1]
Jie Liang[2], **Tianhe Wu**[5], **Lei Zhang**[2,6,*]
[1]National University of Singapore, [2]OPPO Research Institute
[3]Nanyang Technological University, [4]Duke University
[5]City University of Hong Kong, [6]The Hong Kong Polytechnic University
guanyi.qin@u.nus.edu
https://github.com/narthchin/Tool-IQA

Abstract

Vision-Language Models (VLMs) have been increasingly adopted for Image Quality Assessment (IQA). However, current methods typically employ a static *one-shot* scoring paradigm, despite the fact that humans assess image quality through dynamic visual inspection, *e.g.*, selectively adjusting views to verify details and subtle artifacts. Specifically, relying solely on a single-pass observation introduces two primary limitations: first, perceiving the image only at a global scale restricts the assessment of finer local details; second, the original intensity distribution of the image may overwhelm the visibility, leading to insufficient inspection of image quality. To address these issues, we propose **Tool-IQA**, shifting the assessment mechanism from passive scoring to a tool-augmented workflow. In particular, we equip VLMs with simple yet effective view tools: a *Magnifier* to inspect local details, and a *Gamma Corrector* to uncover visibility and hidden artifacts. The assessment follows a structured pipeline that consists of an initial observation with rubric notes, a tool-augmented in-depth inspection, and a final quantification for calibrated quality score. Furthermore, to ensure efficient and purposeful tool callings, we introduce a batch-aware training strategy to reward tool interactions that can yield positive contributions rather than simply encouraging usage. Experiments on a variety of IQA benchmarks demonstrate that, with effective tool calling and calibrated assessment, our proposed Tool-IQA significantly outperforms existing state-of-the-art models, *e.g.*, it achieves a PLCC of 0.854 on the challenging CLIVE dataset.

1 Introduction

The goal of Image Quality Assessment (IQA) is to emulate the Human Visual System (HVS) in terms of perceptual quality [39, 46]. Conventionally, this task was performed using handcrafted statistical features [25, 52, 53, 26], followed by the adoption of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to learn data-driven representations [17, 18, 31, 34, 37, 54]. With the advances of large Vision-Language Models (VLMs) [32], which are pre-trained on massive data and possess profound image understanding capabilities, a significant transformation of IQA model training is underway [42, 61], not only in improving the scoring performance but also in the assessment patterns from scalar regression to more comprehensive, interpretable ones.

To align VLMs with IQA purpose, strategies ranging from prompt engineering to Supervised Fine-Tuning (SFT) have been employed [4, 41]. For example, Q-Align [42] reformulates IQA as a

*Corresponding author. †Equal Contributions.

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

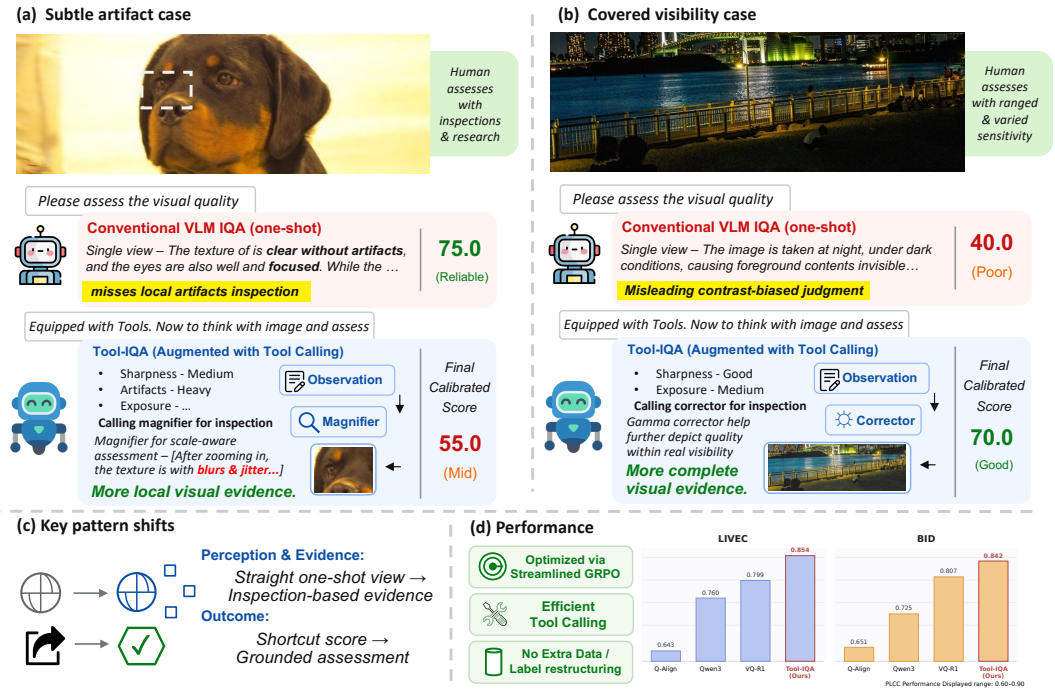


Figure 1: Existing VLM-based IQA methods typically maintain a one-shot pattern, which can lead to (a) *scale-limited failures* by overlooking artifacts and (b) *visibility-biased failures* by misattributing to poor quality. Subfigure (c) shows our proposed pattern shifts for IQA, and (d) shows that Tool-IQA leads to more satisfied performance by utilizing simple tools without extra data and annotation.

text-defined rating task and Q-Instruct [41] constructs instruction-following data for quality-aware understanding, respectively. Inspired by recent RLHF breakthroughs [10], reinforcement learning has also been introduced to optimize VLM-based quality assessment beyond SFT. A representative work is VisualQuality-R1 [45], where Group Relative Policy Optimization (GRPO) [10] with ranking rewards is employed to assess the multiple candidates sampled for each image-prompt pair and optimize the model according to human annotations. Such a reinforcement mechanism effectively improves scoring accuracy [20, 45].

Notwithstanding these advances, a fundamental problem persists in the assessment workflow: determining image quality from a single static glance is inherently insufficient. HVS [28, 27] mitigates ambiguities through a sophisticated visual inspection process by exploring and inspection: observers selectively adjust their views to access local details [36], while adapting visual sensitivity to inspect subtle artifacts hidden under various lighting conditions [1]. In contrast, current VLM-based frameworks mostly adopt a passive *one-shot* workflow [31, 45, 7] using a single-pass observation at a global scale and a fixed contrast. This limitation manifests as *scale-limited failure*, where subtle artifacts are overlooked or misidentified under the global scale, as shown in Fig. 1(a); and *visibility-biased failure*, where low-light or contrast-dominated scenes lead to misleading quality attribution, as shown in Fig. 1(b). Consequently, without the affordance for in-depth perception, these models are prone to exploiting shortcut cues or illusions for quality assessment.

To bridge this gap, we propose to align VLM-based assessment models with the HVS inspection process, equipping them with the interactive mechanisms to perform sequential observation, *e.g.*, observing the global context, characterizing regions that require closer and detailed inspection [59]. To this end, we propose employing simple yet effective view tools to supplement the initial input, *i.e.*, a ***Magnifier*** for local inspection and a ***Gamma Corrector*** for visibility-aware inspection to uncover more details, reformulating the assessment process from passive scoring to a tool-augmented workflow. Based on these principles, we propose **Tool-IQA**, a novel tool-augmented framework optimized via GRPO [10], where the inspection of complex visual attributes is explicitly externalized to and supplemented by auxiliary tools. It is worth mentioning that Tool-IQA achieves superior performance through effective adaptive perception, **without requiring additional datasets or restructuring**

existing annotations, resulting in stronger capability to characterize evidence and produce calibrated assessments, as illustrated and summarized in Fig. 1.

Specifically, Tool-IQA adopts a systematic workflow that encompasses *observation*, *inspection*, and *quantification*. During the initial *observation* phase, a rubric note protocol is employed. The model analyzes standard imaging attributes [39], *e.g.*, sharpness, artifacts, and exposure, recording its estimated levels and textual descriptions to establish a baseline of the global visual context. Subsequently, the model is prompted to determine the sufficiency of this initial assessment. Should further scrutiny be required, a tool-assisted *inspection* is initiated, selectively invoking a tool from the toolkit consisting of a Magnifier and a Gamma Corrector, to overcome the scale and contrast limitations of the initial observation. Finally, the model performs inspections upon these visual returns as auxiliary evidence for inspecting the original image and gives the assessment results.

Furthermore, to promote tool-related learning, we propose a Batch-Aware Tool-Efficiency (**BATE**) reward based on the Reinforcement Learning to Rank (RL2R) strategy [45, 10]. Different from conventional static incentives [15, 59], BATE dynamically modulates rewards according to global batch-wise usage rates while enforcing intra-group competition between tool-invoked and direct assessment trajectories. This tournament-style pairwise comparison mechanism effectively guides the model to discern the need of external assistance, fostering efficient and contextual tool-calling.

Our contributions are summarized as follows:

- We propose Tool-IQA, which facilitates the transition from *one-shot* scoring to a *tool-augmented* workflow for more effective IQA with VLMs.
- We introduce a specialized toolkit comprising a magnifier and a Gamma corrector, designed to overcome the limitations of shallow observations by providing complementary visual evidence.
- We design a BATE reward function within the GRPO framework, dynamically incentivizing appropriate tool-calling to enhance the efficiency.

Extensive experiments are conducted on multiple IQA benchmarks, where Tool-IQA significantly surpasses existing methods, without the need for additional annotations or SFT starting, validating the effectiveness of our tool-augmented design.

2 Related Work

VLM-based Image Quality Assessment. IQA has traditionally relied on dedicated regressors trained on Mean Opinion Scores (MOS) from subjective studies [31, 44, 60]. Recently, VLMs have been explored as general-purpose *quality judges*, capable of describing and reasoning about low-level degradations in natural language [40, 58]. This line of work has been advanced through instruction tuning with human feedback, language-grounded descriptive or comparative assessment, distribution-aware score regression, and objectives for better cross-dataset consistency [41, 51, 42, 43, 61, 50]. Further efforts incorporate region-level grounding and semantic guidance to localize distortions and assess AI-generated images with complex artifacts [2, 38]. Despite these advances, most existing VLM-based IQA methods remain passive and one-shot, deriving judgments from a single observation, which may miss subtle distortions or suffer from scene-level visibility and contrast biases.

Tool-Use in VLMs. Recent studies improve VLMs by augmenting them with active perception, external tools, and modular visual experts, enabling iterative evidence gathering and decision making [59, 12, 57]. By casting multimodal reasoning as executable programs or reasoning-and-action trajectories, models can decompose visual tasks and invoke modules, code, or external APIs for additional evidence [11, 35, 49]. Subsequent work strengthens this paradigm through tool-trace distillation, recursive decomposition, self-trained visual program synthesis, and robustness-oriented testing [14, 8, 19, 29]. More broadly, modular tool-use and decoupled perception–reasoning frameworks show that external augmentation can complement native VLM perception, especially for tasks requiring iterative inspection, region-level evidence, or auxiliary computation [33, 30]. In this paper, we introduce tool-use into VLM-based IQA, enabling active inspection for complementary visual evidence and more calibrated quality assessment.

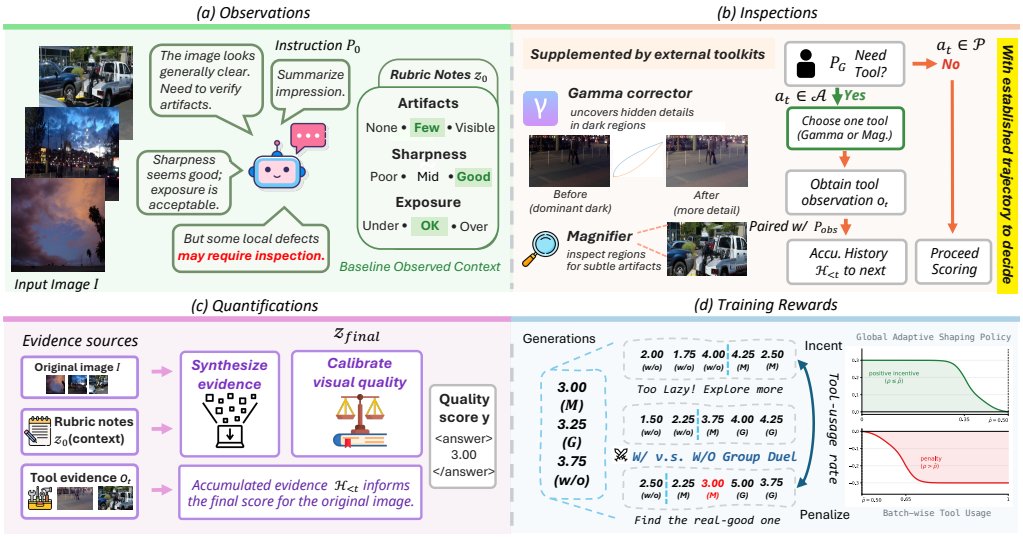


Figure 2: Overview of Tool-IQA, which involves specific stages, including (a) observation with rubric notes, (b) tool-augmented inspection, and (c) quantification. In addition, a (d) BATE reward is designed to improve tool-calling efficiency.

3 Tool-IQA: Methodology

3.1 Framework Overview

Problem Formulation. We formulate Tool-IQA as a structured tool-augmented process for IQA. Different from conventional VLM-based IQA methods that rely on one-shot scoring, Tool-IQA enables the VLM to inspect quality-related cues using external tools and incorporate them as additional evidence for assessment. As illustrated in Fig. 2, given an input image I and an initial task instruction P_0, the model first generates a rubric note z_0, which summarizes visual attributes, *e.g.*, sharpness, exposure, contrast, and artifacts, thereby establishing the baseline. The model then performs the assessment through a trajectory $\tau = (z_0, a_1, o_1, \ldots, a_T, o_T, z_{\text{final}}, y)$. At each step t, a tool-selection prompt P_G guides the model to select an operation $a_t \sim \pi_\theta(a_t \mid \mathcal{H}_{<t}, P_G)$, where $\mathcal{H}_{<t}$ denotes the accumulated history of previous assessment notes, selected operations, and observations. The operation space consists of external tool calls $\mathcal{A}$, instantiated as either the magnifier or the Gamma corrector, and a final assessment option $\mathcal{P}$. When $a_t \in \mathcal{A}$, an external tool is invoked to supplement additional visual evidence o_t, which is paired with an observation prompt P_{obs} and provided to the VLM to support closer inspection. When $a_t \in \mathcal{P}$, the model summarizes the accumulated tool evidence together with the initial rubric note, figures the final assessment rationale z_{final}, and outputs a calibrated quality score y for the original image I. In this way, Tool-IQA shifts VLM-based IQA from one-shot scoring to evidence-aware quality assessment, where tool observations are explicitly incorporated to calibrate the quality score of the given image.

Learning Objective. We adopt GRPO [10] to train the policy π_θ without requiring manually annotated assessment trajectories. For a given input (I, P_0), a group of G trajectories $\{\tau_1, \tau_2, \ldots, \tau_G\}$ is sampled from the old policy $\pi_{\theta_{\text{old}}}$. The optimization objective $J(\theta)$ is formulated as:

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta_{\text{old}}}} \left[\frac{\pi_\theta(\tau)}{\pi_{\theta_{\text{old}}}(\tau)} A(\tau) - \beta \cdot \mathcal{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \right], \tag{1}$$

where the relative advantage $A(\tau)$ is computed by normalizing the trajectory rewards $R(\tau)$ within each sampled group. In Tool-IQA, $R(\tau)$ combines the rank reward [45] and the BATE reward introduced in Sec. 3.4. The rank reward encourages consistency between the predicted quality score y and the annotation, while the BATE reward promotes tool calls that contribute positively to the final assessment rather than simply rewarding the usage. Through these settings, GRPO assigns advantages to trajectories that achieve reliable final quantification and effective tool usage. As a result, the model is encouraged to coordinate external tools and final scoring within successful trajectories, allowing purposeful tool-calling capabilities for quality assessment to emerge through training.

3.2 Workflow Specifics

Initial Rubric Observation (z_0). As illustrated in Fig. 2(a), the assessment workflow starts with a structured decomposition of perceptual attributes. Conditioned on the initial prompt P_0, the model performs an attribute assessment over well-established quality-related dimensions [39], *i.e.*, **S**harpness, **E**xposure, **N**oise, **A**rtifacts, and **V**isibility, and records the result as a structured rubric note z_0. Rather than directly producing a score, z_0 contains categorical attribute levels and brief observed notes, following the format:

```
Attribute: [Level] - Brief Note.
```

This rubric analysis establishes a perceptual baseline for the subsequent assessment by eliciting and organizing the rationales into structured notes, as shown in Fig. 2. These notes summarize and open up quality-related entries to be enriched by external tool observations, thereby benefiting final score calibration. The prompt template is detailed in **Appendix A**.

Tool-Augmented Inspection (a_t). Following the rubric analysis, Tool-IQA introduces a decision step in which the model chooses whether to acquire additional tool-assisted observations before final scoring, as illustrated in Fig. 2(b). Guided by the tool-selection prompt P_G, the model is instructed to choose one of the following operations:

```
Output one of: <tool_call> or <answer>.
```

Under such, the model may select a tool call ($a_t \in \mathcal{A}$) to perform further inspections, or select the quality assessment option ($a_t \in \mathcal{P}$) to provide the final calibrated score. Generated by the policy π_θ conditioned on the accumulated history, this decision determines how the assessment moves from inspection to quantification. In this way, Tool-IQA enables purposeful tool-calling behaviors to emerge during training, allowing the assessment process to adapt to the available visual evidence.

Quantification (z_{final}, y). When the policy emits the quality assessment option ($a_t \in \mathcal{P}$), or when the maximum interaction depth is reached, the model directly provides the final calibrated score, as illustrated in Fig. 2(c). Specifically, the model synthesizes the accumulated assessment trajectory by combining the initial rubric note z_0 with any supplementary tool observations $\{o_t\}$ gathered during the inspection. It then generates the final assessment rationale z_{final} and outputs a scalar quality score y for the original image I. In this way, the final prediction is calibrated with both the initial assessment baseline and external tool observations.

3.3 Perceptual Toolkit

To support observation supplementation, we introduce two optional external tools to provide complementary views: the *Magnifier* enables scale-aware inspection by re-concentrating representation on regions, while the *Gamma Corrector* provides visibility views.

Magnifier (Scale-aware Inspection). In the original view, local distortions may be under-emphasized due to inherently limited spatial extent or interference from contents, *e.g.*, subtle blur, texture degradation, blocking artifacts, and small structural distortions. To supplement the observations, the model issues a magnifier call with a candidate region specified by a bounding box $[x_1, y_1, x_2, y_2]$, conditioned on the accumulated assessment context. The corresponding region is cropped from the original image and presented as an additional view for inspection. Such a view changes the effective inspection scale, making regional structures, textures, and artifacts more salient to the VLM. The resulting observation provides additional evidence supporting the final score calibration.

Gamma Corrector (Visibility Observation). The Gamma Corrector is also available as an optional external tool for generating visibility-referenced views under different intensity conditions. When selected by the model, it supports the inspection of content visibility in low-contrast regions, especially for extremely dark or bright images, where details may be difficult to assess from the original view. To obtain such observations, the Gamma Corrector applies a non-linear intensity transformation, denoted as $I_{\text{out}} = I_{\text{in}}^{\gamma}$, where I_{in} is normalized to $[0, 1]$ and γ controls the intensity remapping. By varying γ, the tool produces alternative views that emphasize different intensity ranges, with generated views treated as supplementary visual references that help the model assess visibility-referenced cues and calibrate the final quality score of the original image.

3.4 Batch-Aware Tool-Efficiency Reward

To encourage purposeful tool calling, we propose a **B**atch-**A**ware **T**ool-**E**fficiency Reward, denoted as **BATE**. Instead of treating tool usage as a binary objective, BATE evaluates the ***marginal utility*** of tool calls relative to tool-free scoring, encouraging external tools to be invoked only when they provide positive contributions to the final assessment. As shown in Fig. 2(d), BATE operates at two levels: a local (*i.e.*, within the generation group) pairwise comparison of solution fidelity, and a global (*i.e.*, across all groups) batch-level tool-calling rate regulation.

Tool-Free Duels. To promote efficiency and discourage indiscriminate tool calling, BATE evaluates tool usage by its ***gain*** over tool-free scoring. Rather than assigning a bonus to any tool call, the generated trajectories for each input are partitioned into two patterns: a tool-invoked set $\mathcal{T}$ and a tool-free set $\mathcal{N}$. The utility of tool usage is then estimated through pairwise duels between these two cohorts, following a Thurstonian preference formulation.

Specifically, for a tool-invoked trajectory $t \in \mathcal{T}$, the base reward is derived by comparing it against the local ensemble of tool-free trajectories $\mathcal{N}$ generated for the same input:

$$R_{\text{base}}^{(t)} \propto \frac{1}{|\mathcal{N}|} \sum_{n \in \mathcal{N}} \sqrt{\Phi\left(\frac{s_t - s_n}{\tau}\right)}, \tag{2}$$

where Φ denotes the cumulative distribution function of the normal distribution, τ is a temperature parameter, and s_t and s_n denote the fidelity rank scores of the tool-invoked and tool-free trajectories, respectively, computed following [45]. By grounding the reward in such group-wise duels, BATE reinforces tool usage only when it brings a positive accuracy margin over tool-free scoring.

Global Adaptive Shaping. During training, the model may fall into degenerate patterns, *e.g.*, indiscriminate tool calling, leading to insufficient exploration or unnecessary computation and inspection on high-quality images with few visible degradations. To address this, we compute the global tool utilization rate ρ over all valid samples in the current batch and use it to modulate the reward through two schedules: an *incentive curve* and a *penalty curve*. When tool utilization is sparse ($\rho \leq \hat{\rho}$), the incentive term encourages tool exploration. Conversely, when the utilization is saturated ($\rho > \hat{\rho}$), the penalty term makes tool calling more selective by suppressing low-margin usage. This encourages tool-invoked trajectories to retain a sufficient advantage over their tool-free counterparts. The total reward is defined as:

$$R_{\text{total}}^{(t)} = R_{\text{base}}^{(t)} + \begin{cases} \mathcal{S}_{\text{inc}}(\rho, s_t) & \text{if } \rho \leq \hat{\rho}, \\ -\mathcal{S}_{\text{pen}}(\rho) & \text{if } \rho > \hat{\rho}, \end{cases} \tag{3}$$

where $\mathcal{S}_{\text{inc}}$ and $\mathcal{S}_{\text{pen}}$ denote smoothed sigmoidal piecewise functions for incentive and penalty, respectively. By batch-level regulation, BATE encourages a balanced and purposeful tool-calling policy, where external tools are favored when they provide a factual contribution to the final assessment.

4 Experiments

4.1 Experimental Settings

Experiment Setup. Following previous protocols [20, 45], we train Tool-IQA on 60% of the synthetic KADID-10K dataset [22], which contains 10,125 images with 25 distortion types. Without further fine-tuning, we evaluate it on four authentic blind IQA benchmarks, including SPAQ [6], CLIVE [9], BID [5], and KonIQ-10K [13], as well as three algorithm-oriented datasets, Dehaze [24], SRIQA [3], and Deblur [23]. These datasets cover diverse real-world distortions, capture conditions, and task-specific degradations, enabling a comprehensive evaluation of Tool-IQA.

Evaluation Metrics. We evaluate performance using Spearman's rank-order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC). SRCC measures monotonic consistency with human judgments, while PLCC measures linear correlation after score regression. For both metrics, higher values indicate better performance and alignment with human annotations.

Implementation Details. *Qwen3-VL-8B-Instruct* was used as the foundation VLM. Training was conducted with GRPO for 680 steps on 16 NVIDIA A800 GPUs, where 4 GPUs were used for vLLM-based on-policy sampling and 12 for parameter updates, taking about 16 hours. During the

Table 1: Results of models trained on KADID-10K. Gains are against the best competing method. VisualQuality-R1 is reproduced with Qwen3. The top twos are in **bold** and underline, respectively.

Methods	*Scenarios* - SRCC/PLCC							
	BID	CLIVE	KonIQ	SPAQ	Deblur	SRIQA	Dehaze	Avg
Handcrafted								
NIQE [26]	.515/.527	.450/.494	.421/.439	.676/.683	.360/.376	.557/.587	.343/.482	.475/.513
BRISQUE [25]	.522/.528	.314/.362	.385/.400	.614/.624	.389/.380	.482/.556	.242/.468	.421/.474
Discriminative Deep-Learning-based								
UNIQUE [55]	.412/.385	.470/.472	.649/.590	.751/.708	.669/.654	.649/.668	.577/.578	.597/.579
MUSIQ [18]	.327/.280	.284/.325	.473/.435	.720/.666	.656/.563	.404/.441	.458/.455	.475/.452
MANIQA [48]	.420/.512	.487/.571	.213/.257	.745/.753	.726/.728	.263/.243	.608/.663	.495/.532
DEIQT [31]	.534/.539	.511/.528	.498/.493	.740/.735	.734/.759	.412/.438	.613/.652	.577/.592
VLM-based								
LIQE [56]	.677/.680	.719/.726	.684/.652	.815/.814	.797/.712	.743/.775	.646/.661	.726/.717
Q-Align [42]	.576/.651	.554/.643	.573/.612	.767/.779	.761/.802	.684/.713	.455/.525	.624/.675
DeQA-Score [50]	.702/.743	.743/.795	.677/.703	.852/.858	.785/.838	.710/.763	.643/.688	.730/.770
Qwen2.5-VL-7B [32]	.711/.725	.733/.760	.754/.810	.848/.854	.820/.852	.603/.653	.458/.553	.704/.744
Qwen3-VL-8B-Ins [47]	.779/.725	.760/.772	.729/.769	.856/.849	.810/.842	.694/.797	.603/.652	.747/.772
Q-Insight [20]	.784/.796	.761/.795	.806/.829	.872/.872	.831/.857	.724/.798	.601/.669	.768/.802
VisualQuality-R1 [45]	.788/.807	.792/.799	.788/.811	.872/.856	.811/.859	.709/.814	.662/.721	.775/.810
Tool-IQA (Ours)	**.828/.842**	**.833/.854**	**.825/.848**	**.898/.898**	**.848/.878**	**.751/.847**	**.730/.797**	**.816/.852**
Gains	.040/.035	.041/.055	.019/.019	.026/.026	.017/.019	.008/.033	.068/.076	.042/.042

Table 2: Replacement study on VQ-R1.

Method	CLIVE	KonIQ	SPAQ
VQ-R1	.792/.799	.788/.811	.872/.856
# Replaced	476	2357	6243
VQ-R1 w/ Tools	.794/.807	.784/.808	.870/.858
Tool-IQA (Ours)	**.833/.854**	**.825/.848**	**.898/.898**
Gains	.039/.047	.041/.040	.028/.040

Table 3: Augmented methods trained on KonIQ.

Method	SPAQ	KADID	CLIVE	PIPAL [16]
VisualQuality-R1	.892/.889	.712/.703	.827/.856	.441/.451
Tool-augmented				
Zoom-IQA [21]	.900/.902	.700/.701	.870/.887	.465/.468
Tool-IQA (Ours)	**.905/.906**	**.721/.737**	**.889/.902**	**.479/.483**
Gains	.005/.004	.021/.036	.019/.015	.014/.015

Table 4: Ablation study on the tool-calling pattern regarding training.

Methods	Training	*Scenarios*		
		CLIVE	KonIQ	SPAQ
SRCC				
Qwen3-VL-8B-Ins	×	0.760	0.729	0.856
VisualQuality-R1	✓	0.792	0.788	0.872
Tool-IQA	×	0.794	0.732	0.866
Tool-IQA	✓	**0.833**	**0.825**	**0.898**
PLCC				
Qwen3-VL-8B-Ins	×	0.772	0.769	0.849
VisualQuality-R1	✓	0.799	0.811	0.856
Tool-IQA	×	0.828	0.781	0.872
Tool-IQA	✓	**0.854**	**0.848**	**0.898**

Table 5: Ablation study on the effectiveness of tools. **Mag.** and **Cor.** denote the Magnifier and Gamma Corrector, respectively.

Methods	Mag.	Cor.	*Scenarios*		
			CLIVE	KonIQ	SPAQ
SRCC					
VisualQuality-R1	×	×	0.792	0.788	0.872
Tool-IQA	✓	×	0.808	0.779	0.880
Tool-IQA	✓	✓	**0.833**	**0.825**	**0.898**
PLCC					
VisualQuality-R1	×	×	0.799	0.811	0.856
Tool-IQA	✓	×	0.833	0.809	0.882
Tool-IQA	✓	✓	**0.854**	**0.848**	**0.898**

first 340 steps, we used a learning rate of 2×10^{-6}, KL coefficient $\beta = 0.04$, temperature 0.9, 4 candidates per input, warmup ratio 0.05, batch size 12, and sequence-level importance sampling. In the last 340 steps, only β was increased to 0.06. The model was allowed at most two tool-use rounds, with one call per round. Magnifier coordinates were parsed from `<tool_call>`, while the Gamma Corrector provided darkening and brightening options. Invalid tool calls were replaced with failure prompts, and the random seed was set to 42.

The training objective is a combination of BATE, the rank reward [45], and a format constraint. For BATE, the target tool utilization rate was $\hat{\rho} = 0.50$. Batch-level utilization ρ was shaped with incentives for $\rho \in [0.35, 0.50)$ and penalties for $\rho \in (0.50, 0.65]$, with threefold strength when ρ fell outside $[0.35, 0.65]$. The shaping signals were sigmoid-smoothed, rank-modulated, and applied only to successful tool calls.

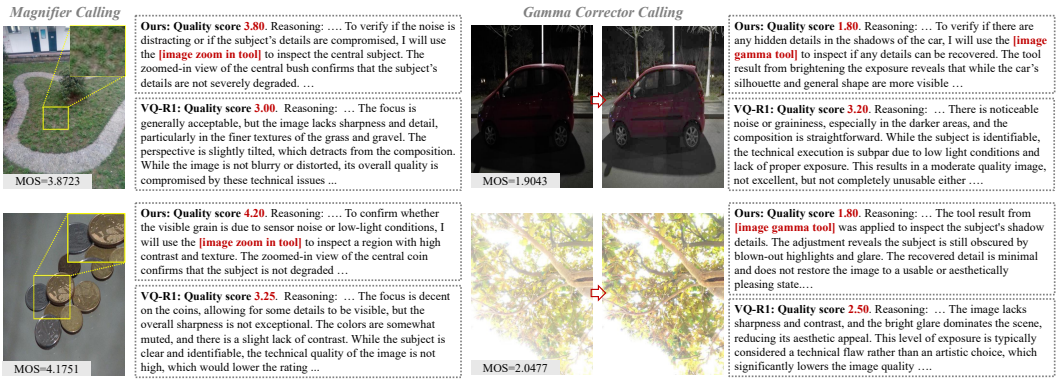


Figure 3: Visualization of the tool-calling mechanism of Tool-IQA. Tool-IQA demonstrates higher score prediction consistency with the ground-truths while providing better assessment process.

4.2 Performance Comparison

Quantitative Comparison. As shown in Tab. 1, we compare Tool-IQA with representative handcrafted [26, 25], discriminative learning-based [55, 18, 48, 31], and VLM-based methods [56, 42, 50, 47, 45, 20]. We see that conventional models show limited cross-dataset generalization, whereas VLM-based methods achieve much more competitive results. GRPO-based methods generally outperform SFT-based ones, suggesting the effectiveness of RLHF for VLM-based IQA. Among them, Tool-IQA consistently obtains superior performance, especially on authentic datasets with complex contents and capture conditions, such as CLIVE and SPAQ. In particular, Tool-IQA achieves a PLCC of 0.854 on CLIVE. The gains mainly come from the tool-augmented workflow, which provides supplementary visual observations for better score calibration. To ensure a fair comparison, we further equip VisualQuality-R1 (VQ-R1) [45] with the tool outputs of Tool-IQA and let it analyze the replaced samples. The results are shown in Tab. 2. We see that VQ-R1 w/ Tools brings only marginal gains, indicating limited ability to leverage external tools. In addition, we present more results of Tool-IQA on LIVEFB and TID datasets in **Appendix B**.

Comparison with Tool-augmented Methods. We further compare Tool-IQA with a recently developed tool-augmented method, Zoom-IQA [21]. Since the source code and model of Zoom-IQA are not publicly available, we re-train our Tool-IQA on KonIQ according to the experimental settings of Zoom-IQA, and compare with its officially reported results. As shown in Tab. 3, Tool-IQA achieves consistently better performance without requiring additional cold-starting annotations, demonstrating the effectiveness of our design, including the toolkits and BATE reward.

Visualizations and Qualitative Analysis. As illustrated in Fig. 3, Tool-IQA learns to invoke tools within the assessment trajectory according to the input image. The Magnifier is typically used to inspect details, *e.g.*, to examine textures, or artifacts. In contrast, the Gamma Corrector provides visibility-aware views for images with ill-exposed, or low-contrast regions, specifically, to make structures and visibility-related cues more discernible for final score calibration. This behavior resembles human visual inspection, where observers adjust their viewing focus or sensitivity according to image content. Compared with VQ-R1, Tool-IQA produces predictions that are more consistent with human annotations, which is also reflected in the quantitative results.

4.3 Ablation Studies

Tool-calling Patterns. The proposed tool-augmented workflow and the training pipeline are further examined in Tab. 4. Even without training, Tool-IQA achieves better correlations than the baseline, suggesting that the structured workflow can elicit useful tool-assisted observations from the VLM. Further gains are obtained after training with the proposed reward functions, consistently outperforming VQ-R1 across all scenarios, and confirming that robust quality assessment benefits from the combination of augmented observation supplementation and purposeful tool-calling optimization.

Tool Effectiveness. The effectiveness of the proposed tools, *i.e.*, the Magnifier and Gamma Corrector, is validated in Tab. 5. The integration of such tools consistently improves performance over the baseline after training, with the best results achieved when both tools are available. Rather than relying

Table 6: Comparative analysis of VQ-R1 and Tool-IQA on SPAQ **w/ training**.

VQ-R1		Tool-IQA				Gain (Δ)	
SRCC	PLCC	$\mathcal{F}(\%)$	$\mathcal{T}$	SRCC	PLCC	SRCC	PLCC
Mechanism Analysis							
0.872	0.856	100.0	—	**0.898**	**0.898**	+0.026	+0.042
0.501	0.492	43.9	×	**0.586**	**0.574**	+0.085	+0.083
0.752	0.723	56.1	✓	**0.790**	**0.777**	+0.039	+0.054
Specific Tool Effectiveness							
0.762	0.731	47.4	$\mathcal{C}$	**0.800**	**0.782**	+0.038	+0.051
0.677	0.678	8.7	$\mathcal{M}$	**0.695**	**0.740**	+0.018	+0.063

Table 7: Analysis of tool-use capability of Qwen3-VL and Tool-IQA on SPAQ **w/o training**.

Qwen3-VL		Tool-IQA (w/o Train)				Gain (Δ)	
SRCC	PLCC	$\mathcal{F}(\%)$	$\mathcal{T}$	SRCC	PLCC	SRCC	PLCC
Mechanism Analysis							
0.857	0.850	100.0	—	**0.866**	**0.872**	+0.010	+0.022
0.249	0.395	11.1	×	0.172	**0.423**	-0.076	+0.028
0.840	0.825	88.9	✓	**0.849**	**0.852**	+0.009	+0.026
Specific Tool Effectiveness							
0.843	0.827	75.7	$\mathcal{C}$	**0.847**	**0.854**	+0.004	+0.028
0.600	0.689	13.5	$\mathcal{M}$	**0.603**	**0.715**	+0.003	+0.027

Table 8: Efficiency comparison.

Method	Speed (it/s)	OTokens	Tool Calls
VQ-R1	6.35	192.85	-
w/o BATE	2.28	587.62	1.77
Tool-IQA	4.10	350.51	0.39

Table 9: Ablation study on components.

Components			CLIVE		SPAQ	
Rubric	Tools	BATE	SRCC	PLCC	SRCC	PLCC
×	×	×	0.792	0.799	0.872	0.856
✓	×	×	0.796	0.814	0.878	0.873
✓	✓	×	0.812	0.837	0.888	0.890
✓	✓	✓	**0.832**	**0.854**	**0.898**	**0.898**

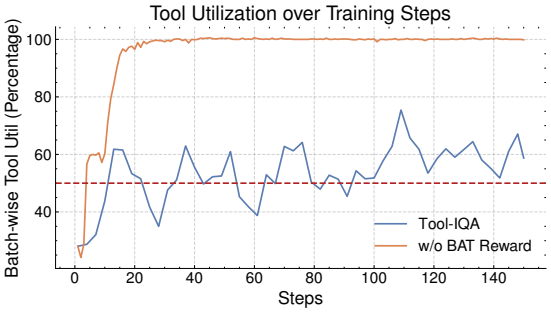


Figure 4: Curves of batch-wise tool utilization over training steps.

solely on the original image view, the Magnifier and Gamma Corrector provide complementary auxiliary views for scale-aware inspection and visibility-aware observation, respectively. These observations complement the initial analysis and contribute to more reliable final score calibration. Results of Qwen3 under zero-shot settings could be found in **Appendix C**.

Tool Utilization & Efficiency. Tab. 6 summarizes tool-use behavior on SPAQ, where $\mathcal{F}$ denotes frequency, $\mathcal{T}$ tool usage, and $\mathcal{C}/\mathcal{M}$ Gamma Corrector/Magnifier. With BATE, Tool-IQA learns selective tool invocation, reducing unnecessary computation while improving over VQ-R1 when tools are used. The Gamma Corrector yields the largest gain, with a Δ of +0.038, while gains on tool-free samples ($\mathcal{T} = \times$) indicate that the rubric-note assessment and BATE learning also benefit the tool-free pathway. Tab. 7 shows that the base VLM tends to call tools before training, with a calling rate of 89%. This indicates basic instruction-following ability for tool use, but also reveals poor IQA-specific calibration, calling for training and the BATE promotion. Furthermore, we evaluate the inference latency of full pipeline, including tool generations and feeding, as shown in Tab. 8. Tool-IQA further improves speed and reduces output tokens and tool calls compared with w/o BATE, demonstrating better efficiency and supporting the component gains in Tab. 9.

Modular Study. Tab. 9 dissects the contribution of each core component, showing consistent gains from the rubric-note protocol, tool-augmented workflow, and BATE reward. As shown in Fig. 4, removing BATE causes excessive tool calling, while BATE promotes selective tool use by favoring calls with positive marginal contributions.

5 Conclusion

In this work, we presented Tool-IQA, which reformulated VLM-based IQA from *one-shot* prediction into a *tool-augmented* assessment workflow. Specifically, we introduced two external tools, including a Magnifier for scale-aware inspection and a Gamma Corrector for visibility-aware observation, to supplement with complementary visual evidence. In this way, Tool-IQA produced calibrated and better-aligned quality scores for the input image. Furthermore, with the proposed BATE reward under GRPO, Tool-IQA learned purposeful tool-calling behaviors via principled exploration. Extensive

experiments on widely used IQA benchmarks demonstrated that Tool-IQA significantly outperformed other VLM-based IQA models, without requiring additional datasets or annotation reconstruction.

Limitations. Despite promising results, our work faces some constraints. First, the framework relies on the *capabilities* of the base VLM. Second, training efficiency is limited by the *rollout sampling speed*. To push the boundary of VLM-based IQA, foundation development needs to be pursued.

Acknowledgments and Disclosure of Funding

The authors received no third-party funding or in-kind support for this work. The authors declare no competing interests.

References

[1] Matteo Carandini and David J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, Jan 2012.

[2] Chaofeng Chen, Sensen Yang, Haoning Wu, Liang Liao, Zicheng Zhang, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Q-ground: Image quality grounding with large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 486–495, New York, NY, USA, 2024. Association for Computing Machinery.

[3] Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang. Toward generalized image quality assessment: Relaxing the perfect reference quality assumption. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12742–12752, 2025.

[4] Zewen Chen, Haina Qin, Juan Wang, Chunfeng Yuan, Bing Li, Weiming Hu, and Liang Wang. Promptiqa: Boosting the performance and generalization for no-reference image quality assessment via prompts. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 247–264, Cham, 2025. Springer Nature Switzerland.

[5] Alexandre Ciancio, André Luiz N Targino Targino da Costa, Eduardo A. B. da Silva, Amir Said, Ramin Samadani, and Pere Obrador. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on Image Processing*, 20(1):64–75, 2011.

[6] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3674–3683, 2020.

[7] Bohan Fu, Guanyi Qin, Fazhan Zhang, Zihao Huang, Mingxuan Li, and Runze Hu. Dr.experts: Differential refinement of distortion-aware experts for blind image quality assessment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(5):3985–3993, Mar. 2026.

[8] Jiaxin Ge, Sanjay Subramanian, Baifeng Shi, Roei Herzig, and Trevor Darrell. Recursive visual programming. In *European Conference on Computer Vision (ECCV)*, 2024.

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

[10] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, September 2025.

[11] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14953–14962, 2023.

[12] Jack Hong, Chenxiao Zhao, Chenglin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. In *International Conference on Learning Representations (ICLR)*, 2026. Poster.

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

[14] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9590–9601, 2024.

[15] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*, 2025.

[16] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S. Ren, and Dong Chao. Pipal: A large-scale image quality assessment dataset for perceptual image restoration. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 633–651, Cham, 2020. Springer International Publishing.

[17] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014.

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

[19] Zaid Khan, Vijay Kumar BG, Samuel Schulter, Yun Fu, and Manmohan Chandraker. Self-training large language models for improved visual program synthesis with visual reinforcement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14344–14353, 2024.

[20] Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin ZHANG, Junlin Li, Li zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[21] Guoqiang Liang, Jianyi Wang, Zhonghua Wu, and Shangchen Zhou. Zoom-iqa: Image quality assessment with reliable region-aware reasoning, 2026.

[22] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10K: A large-scale artificially distorted IQA database. In *IEEE International Conference on Quality of Multimedia Experience*, pages 1–3, 2019.

[23] Yiming Liu, Jue Wang, Sunghyun Cho, Adam Finkelstein, and Szymon Rusinkiewicz. A no-reference metric for evaluating the quality of motion deblurring. *ACM Trans. Graph.*, 32(6), November 2013.

[24] Xiongkuo Min, Guangtao Zhai, Ke Gu, Yucheng Zhu, Jiantao Zhou, Guodong Guo, Xiaokang Yang, Xinping Guan, and Wenjun Zhang. Quality evaluation of image dehazing methods using synthetic hazy images. *IEEE Transactions on Multimedia*, 21(9):2319–2333, 2019.

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

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

[27] Jiri Najemnik and Wilson S. Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, Mar 2005.

[28] Jonathan J. Nassi and Edward M. Callaway. Parallel processing strategies of the primate visual system. *Nature Reviews Neuroscience*, 10(5):360–372, May 2009.

[29] Artemis Panagopoulou, Honglu Zhou, Silvio Savarese, Caiming Xiong, Chris Callison-Burch, Mark Yatskar, and Juan Carlos Niebles. Viunit: Visual unit tests for more robust visual programming. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[30] Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A framework for decoupling and assessing the capabilities of VLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[31] Guanyi Qin, Runze Hu, Yutao Liu, Xiawu Zheng, Haotian Liu, Xiu Li, and Yan Zhang. Data-efficient image quality assessment with attention-panel decoder. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2091–2100, Jun. 2023.

[32] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report, 2025.

[33] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: solving ai tasks with chatgpt and its friends in hugging face. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[34] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3664–3673, 2020.

[35] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11854–11864, 2023.

[36] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[37] Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2555–2563, Jun. 2023.

[38] Puyi Wang, Wei Sun, Zicheng Zhang, Jun Jia, Yanwei Jiang, Zhichao Zhang, Xiongkuo Min, and Guangtao Zhai. Large multi-modality model assisted ai-generated image quality assessment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 7803–7812, New York, NY, USA, 2024. Association for Computing Machinery.

[39] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[40] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *The Twelfth International Conference on Learning Representations*, 2024.

[41] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25490–25500, 2024.

[42] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, et al. Q-align: teaching lmms for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[43] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, Xiaohong Liu, Guangtao Zhai, et al. Towards open-ended visual quality comparison. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 360–377, Cham, 2025. Springer Nature Switzerland.

[44] Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang. A comprehensive study of multimodal large language models for image quality assessment. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 143–160, Cham, 2025. Springer Nature Switzerland.

[45] Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[46] Fengyang Xiao, Peng Hu, Lei Xu, XingE Guo, Guanyi Qin, Yuqi Shen, Chengyu Fang, Rihan Zhang, Chunming He, and Sina Farsiu. Beyond ground-truth: Leveraging image quality priors for real-world image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29897–29908, June 2026.

[47] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, et al. Qwen3 technical report, 2025.

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

[49] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting ChatGPT for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

[50] Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14483–14494, 2025.

[51] Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVII*, page 259–276, Berlin, Heidelberg, 2024. Springer-Verlag.

[52] Lin Zhang, Lei Zhang, and Alan C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015.

[53] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.

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

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

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

[57] Zhehao Zhang, Ryan A. Rossi, Tong Yu, Franck Dernoncourt, Ruiyi Zhang, Jiuxiang Gu, Sungchul Kim, Xiang Chen, Zichao Wang, and Nedim Lipka. Vipact: Visual-perception enhancement via specialized vlm agent collaboration and tool-use. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[58] Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. 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*, 46(12):10404–10418, 2024.

[59] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026. Poster.

[60] Hantao Zhou, Rui Yang, Longxiang Tang, Guanyi Qin, Runze Hu, and Xiu Li. Gamma: Toward generic image assessment with mixture of assessment experts. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 8815–8824, 2025.

[61] Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang. Adaptive image quality assessment via teaching large multimodal model to compare. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc.

Appendix

In this appendix, we provide the following materials:

- **A** Details of prompts used in Tool-IQA, including those providing tools to VLMs.
- **B** More results of Tool-IQA on LIVEFB and TID datasets.
- **C** More studies on the effectiveness of tool callings and relations to rubric notes.

A Prompt Listings

Rubric Observation. We use the following prompts for the rubric turn that conditions the model.

Rubric Reasoning Prompt

```
You are doing the image quality assessment task.
Instead of giving a score immediately, first analyze the image based on the
    following key dimensions:

1. **Sharpness & Detail**: Is the image blurry or sharp? Are textures well-
    defined?
2. **Noise & Grain**: Is there visible noise (luminance/color)? Is it
    distracting?
3. **Artifacts**: Are there compression artifacts (JPEG blocks), ringing, or
    ghosting?
4. **Exposure & Color**: Is it over/underexposed? Are colors natural?
5. **Saliency Visibility**: Is the MAIN SUBJECT clearly visible? Or is it
    obscured by shadows/highlights? Focus primarily on the subject's clarity,
    treating background visibility as secondary.

Output Format:
<thinking>
...
</thinking>
<answer>
Sharpness: [Low/Medium/High] - Brief note
Noise: [None/Visible/Severe] - Brief note
Artifacts: [None/Minor/Severe] - Brief note
Exposure: [Poor/Good/Excellent] - Brief note
Saliency Visibility: [Subject Clear/Obscured] -  Brief Note
</answer>
```

Action Gated - Tool or Answer. We use the following prompt template to choose actions.

Rubric Reasoning Prompt

```
Single-Action per Turn: choose EXACTLY ONE
- A) Final answer (1.00 - 5.00)
- B) Tool call

Output one of:

A) Final Answer:
<thinking>...</thinking>
<answer>...</answer>

B) Tool call:
<thinking>...</thinking>
<tool_call>...</tool_call>
```

Providing Tools. Following are the prompt template for tool providing, as per Qwen3 [47]. Models are required to answer with JSON to call tools.

Rubric Reasoning Prompt

```
{
  "name": "<FUNCTION_NAME>",
  "description": "<BRIEF_DESCRIPTION>",
  "parameters": {
    "<PARAM_NAME>": {
      "type": "<TYPE>",
      "range": [<MIN>, <MAX>],
      "description": "<Usage info or constraints>"
    },
    "<OPTIONAL_PARAM>": {
      "type": "string",
      "description": "<Optional note>"
    }
  },
  "required": ["<PARAM_NAME>"]
}
```

Table 10: Supplemental results.

Method	LIVEFB	TID
VLM-based		
Qwen3-VL-8B-Ins [47]	.357/.408	.656/.706
VisualQuality-R1 [45]	.423/.479	.700/.752
Tool-IQA (Ours)	**.468/.548**	**.733/.780**
Gains	.045/.069	.033/.028

Table 11: Ablation study ***without training***.

Methods	Mag.	Cor.	*Scenarios* CLIVE	KonIQ	SPAQ
SRCC					
Qwen3-VL-8B-Ins	×	×	0.760	0.729	0.856
Tool-IQA	✓	×	0.784	0.714	**0.867**
Tool-IQA	✓	✓	**0.794**	**0.732**	0.866
PLCC					
Qwen3-VL-8B-Ins	×	×	0.772	0.769	0.849
Tool-IQA	✓	×	0.802	0.758	0.871
Tool-IQA	✓	✓	**0.828**	**0.781**	**0.872**

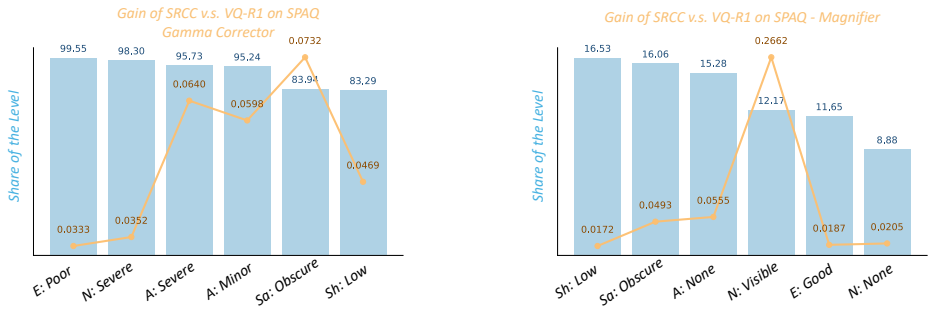


Figure 5: The distribution of tool usage and corresponding gains against induced rubric notes. Zoom in for better view.

B Supplemental Performance Study

We further evaluate Tool-IQA on LIVEFB and TID to verify its generalization ability across diverse IQA benchmarks. As shown in Tab. 10, Tool-IQA consistently outperforms VisualQuality-R1 on these two datasets, demonstrating the effectiveness of tool-augmented visual quality assessment.

C Tool Effectiveness Study

We supplement more ablation studies in Tab. 11. Robust gains are also observed under the *w/o training* setting, indicating that tool-assisted observations can benefit the VLM even without additional optimization. In addition, Fig. 5 reveals distinct tool-usage patterns. We see that the model focuses on the six rubric attribute levels with the highest tool-utilization rates (by share within each attribute level), where the Gamma Corrector tool is predominantly invoked for lower-quality inputs to address visual impairments. Conversely, the magnifier function is utilized more for high-quality images to facilitate detailed inspection, a strategy that is fully consistent with the rubric-guided analysis.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Claims are reflected within proposed method.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Limitations are discussed within section Conclusion.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: Each theoretical reference is within the section of problem formulation.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Information for reproducibility is represented within the section of implementation details.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We provide a clear algorithm description and implementation details, which is conducive to reproducing, and the code and data will be open.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Specified results could be referred to the section of experiments.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The results of both our ablation experiments and the comparison experiments effectively demonstrate the validity of our method and claims.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We provide the GPU model and quantity used by the computers in section of implementation details.

   Guidelines:

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the thesis complies with the NeurIPS Code of Ethics in every respect.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: There is no broader societal impacts of this work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We cite all the work covered in this paper and follow their license.

    Guidelines:

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

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: Our code, data, and model will be open.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: Our user study does not involve any potential risks.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: We just use LLM for editing the content for clarification.

Guidelines:

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