EAGLE: GROUNDING MULTIMODAL LLM JUDGES FOR AI-GENERATED IMAGE ASSESSMENT VIA EVIDENCE ANCHORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) provide a promising foundation for AI-generated image assessment, yet existing MLLM-based evaluators typically derive assessment scores from holistic multimodal representations. This overlooks a fundamental distinction: different judgments require different evidence dependencies, *e.g.*, text-image semantic correspondence depends on cross-modal evidence, whereas perceptual quality or authenticity relies primarily on visual evidence. To adaptively characterize these distinct evidence dependencies across different assessment tasks, we introduce **evidence anchors** that explicitly ground different assessment judgments in their relevant evidence. Building on this idea, we propose **EAGLE** (Evidence-Anchored Grounding for Multimodal LLM Evaluation), an MLLM judging framework that couples holistic MLLM scoring with *anchor-guided evidence decomposition*. EAGLE learns a set of latent queries divided into shared and anchor-specific groups: shared queries capture evidence reusable across judgments, while anchor-specific queries adaptively extract visual or cross-modal evidence to ground their corresponding judgments. This design preserves the broad assessment capability of pretrained MLLMs while providing each judgment with evidence tailored to its underlying dependency. Extensive experiments on popular benchmarks show that EAGLE achieves substantially improved within-dataset and cross-dataset generalization, validating the importance of explicitly grounding MLLM judges in their underlying evidence dependencies.

1 INTRODUCTION

Recent advances in generative AI have substantially improved the fidelity and controllability of text-to-image synthesis, attracting considerable attention from both academia and industry (Rombach et al., 2022; Saharia et al., 2022; Podell et al., 2024; OpenAI, 2025) in AI-generated images. As generated content continues to increase in scale and diversity, manual evaluation becomes increasingly difficult to apply consistently across models, datasets, and assessment dimensions, creating a growing need for reliable automatic assessment. However, this remains challenging because human judgments span a diverse set of perceptual, semantic, and authenticity-related criteria that can rely on different evidence. For instance, the text–image correspondence criterion in image quality assessment requires reasoning over the relationship between the conditioning text prompt and the generated image, whereas perceptual quality and authenticity criteria rely primarily on the visual characteristics of the image itself. As illustrated in Fig. 1(a), the same AI-generated image may therefore require different sources of evidence depending on the assessment objective.

Most existing AI-generated image assessment methods are developed for individual assessment criteria, with dedicated models separately designed for image quality, text–image alignment, authenticity, and other related dimensions (Ke et al., 2021; Yang et al., 2022; Saha et al., 2023; Peng et al., 2024). While effective within their intended assessment dimension, such models are often specialized for single assessment tasks and training distributions, which can restrict their generalization across datasets, assessment tasks, and evaluation scenarios. More recently, evaluators based on multimodal large language models (MLLMs) offer a more flexible alternative by combining broad multimodal knowledge with strong instruction-following capabilities (Wu et al., 2024; You et al., 2025; Cui et al., 2025). Different assessment objectives can be accommodated by these MLLM judges

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

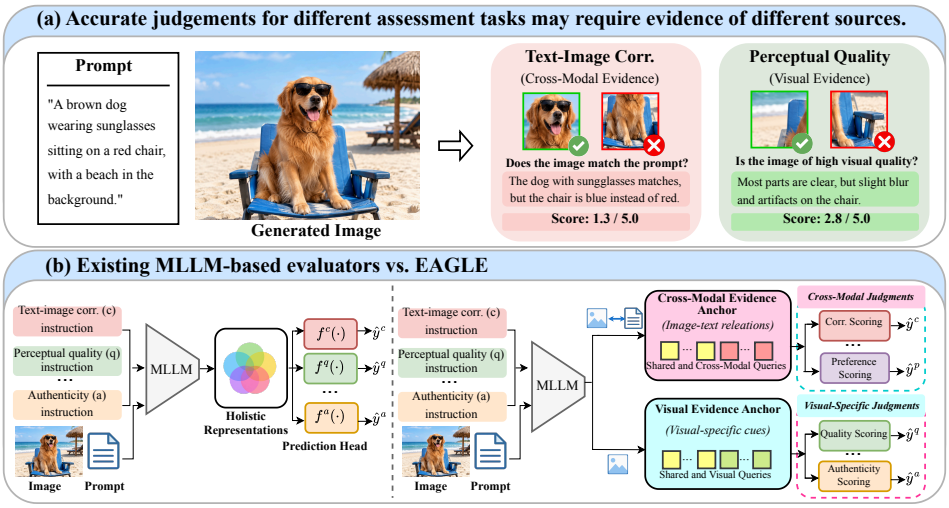

Figure 1: **(a)** Different assessment judgments may rely on different evidence for the same generated image. **(b) Left:** Existing MLLM-based evaluators derive different assessment scores from a holistic multimodal representation; **Right:** EAGLE grounds different judgments in cross-modal or visual evidence anchors and uses shared and anchor-specific queries to acquire the corresponding evidence.

through task-specific instructions and prediction modules while sharing a common multimodal backbone. However, as illustrated in Fig. 1(b) (left), the evidence supporting these predictions typically remains implicit in holistic multimodal representations. Consequently, differences in how individual judgments leverage cross-modal relationships or visual-specific cues for image assessment are only weakly reflected in the representation and scoring process, leading to under-utilization of the broad knowledge learned in MLLMs and suboptimal assessment results.

To address this limitation, we introduce **evidence anchors** to characterize the distinct evidence dependencies required for different assessment judgments. We consider two representative evidence anchors: i) a *cross-modal evidence anchor*, where the judgment is grounded in relationships between the conditioning prompt and the generated image, as required for text–image correspondence; and ii) a *visual evidence anchor*, where the judgment is grounded primarily in information within the visual modality, as required for perceptual quality and authenticity. This motivation shifts the focus from merely *distinguishing what score should be predicted* to *modeling what evidence supports each judgment and how that evidence should be organized*. By doing so, knowledge shared across judgments can be learned jointly to strengthen the multimodal representation, while explicit evidence grounding allows each judgment to focus on the most relevant assessment dependencies.

Building on this idea, we propose **EAGLE** (**E**vidence-**A**nchored **G**rounding for Multimodal **LL**M **E**valuation), an MLLM-based judging framework that couples holistic MLLM scoring with anchor-guided evidence decomposition. The holistic component retains the broad assessment capability of the pretrained MLLM, while the decomposition component introduces latent queries to capture and organize the supporting evidence underlying different judgments. Specifically, the queries are divided into **shared queries**, which capture evidence reusable across judgments, and **anchor-specific queries**, which specialize in evidence associated with the corresponding cross-modal or visual anchor. As illustrated in Fig. 1(b) (right), this structured query organization enables different judgments to acquire and organize supporting evidence according to their corresponding evidence anchors. Together, the holistic and decomposition components combine broad multimodal assessment with structured evidence grounding for different judgments within a unified framework.

Overall, our contributions are summarized as follows:

- We introduce **evidence anchors**, providing a new perspective on characterizing how different judgments rely on distinct forms of evidence in a unified MLLM judging framework for AI-generated image assessment. We consider two representative anchors: cross-modal evidence anchor, capturing relationships between the conditioning text and generated image, and visual evidence anchor, focusing primarily on information within visual modality.

- We further propose **EAGLE**, a novel MLLM-based evaluator framework that augments holistic MLLM representations with anchor-guided evidence decomposition. The key insight is to learn and organize latent queries into shared and anchor-specific groups, with the shared queries grounded by generic, reusable evidence clues across assessment tasks while the anchor-specific queries grounded by task-specific evidence.
- We conduct extensive experiments across a diverse set of AI-generated image assessment benchmarks and tasks. EAGLE achieves substantially better performance on both within-dataset and cross-dataset assessment tasks, supporting the importance of grounding different assessment judgments in their relevant evidence.

2 Related Work

AI-generated image assessment is an evaluation task extending image quality assessment (IQA) (Chen et al., 2025; Shen et al., 2025; Li et al., 2025a) beyond perceptual quality to generation-specific criteria such as text–image correspondence and authenticity. Conventional IQA methods, such as MUSIQ (Ke et al., 2021), MANIQA (Yang et al., 2022), and Re-IQA (Saha et al., 2023), learn perceptual quality representations from visual content, while LIQE (Zhang et al., 2023) further introduces vision–language representations for quality assessment. Compared to them, representative AI-generated image assessment methods leverage prompt conditioning, vision–language representations, and multimodal priors to assess generated content (Fang et al., 2024; Xia et al., 2025; Peng et al., 2024; Li et al., 2026a). Despite their effectiveness, these approaches generally rely on specialized representations or evaluation mechanisms tailored to particular assessment dimensions.

Recent advances further introduce MLLMs into IQA and AI-generated image assessment, enabling more flexible quality assessment through the use of MLLMs and related recent techniques (Li et al., 2025b; Liang et al., 2026; Chen et al., 2026; Wu et al., 2026; Liu et al., 2026; Wen et al., 2025; Li et al., 2026b; Tang et al., 2026; Fan et al., 2026). Specifically, *MLLM-based AI-generated image assessment* methods include Q-Align (Wu et al., 2024) with language-defined quality levels, DeQA-Score (You et al., 2025) with score distributions, and M3-AGIQA (Cui et al., 2025) for multi-aspect assessment. However, existing methods primarily formulate this task as score prediction, while the explicit evidence used to ground each judgment remains largely unexplored. Compared to them, our work explicitly models such task-dependent evidence formation, rather than treating different assessment criteria through a uniformly holistic representation.

3 Methodology

3.1 Problem Formulation for AI-Generated Image Assessment

Given an AI-generated image x and its conditioning prompt p, we consider a set of assessment criteria with different evidence anchors. Let τ denote an assessment criterion (*e.g.*, text–image correspondence and perceptual quality), with each criterion associated with either a cross-modal or visual-evidence anchor according to its underlying evidence dependency. Let $y^\tau \in [1, 5]$ denote the corresponding mean opinion score (MOS) obtained from human ratings, then our goal is to learn a scoring function

$$\hat{y}^\tau = F(x, p, \tau), \tag{1}$$

where $\hat{y}^\tau$ denotes the predicted score for criterion τ from the scoring function F. The learning objective is to estimate the corresponding human-rated score y^τ as accurately as possible. This formulation accommodates both single-criterion assessment and joint learning across criteria with different evidence anchors when two or more assessment annotations are available.

3.2 Overview of EAGLE

Through joint learning across assessment criteria with different evidence anchors, EAGLE learns holistic representations that are better tailored to image assessment than those of existing holistic representation-based methods, leading to more faithful assessment scores. Moreover, unlike existing MLLM evaluators that directly predict scores from holistic representations, EAGLE explicitly learns evidence representations associated with cross-modal and visual evidence anchors, grounding

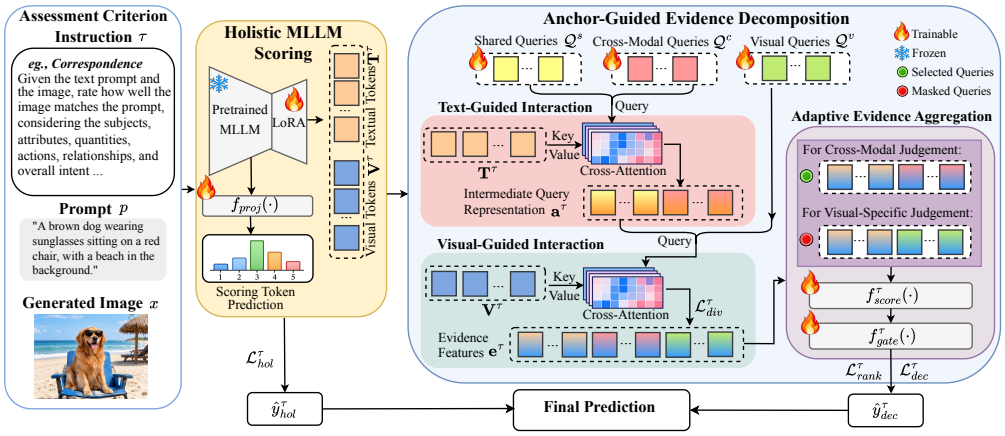

Figure 2: **Overview of EAGLE**. It combines holistic MLLM scoring with anchor-guided evidence decomposition. The decomposition branch organizes latent queries into shared and anchor-specific groups to respectively acquire reusable evidence across judgments and cross-modal/visual evidence anchor, followed by adaptive aggregation of these grounded judgments for criterion-specific scoring. The decomposed prediction score is further combined with the holistic score for final assessment.

its assessment in the evidence relevant to each judgment and enabling more effective modeling of text–image semantic correspondence and vision-specific semantics, respectively.

Specifically, as illustrated in Fig. 2, EAGLE builds on a pretrained MLLM to support assessment criteria with different evidence anchors within a unified framework. For each criterion τ, a criterion-specific instruction conditions the MLLM on the corresponding assessment objective (see `Appendix` B.1 for details). EAGLE consists of two complementary components: *Holistic MLLM Scoring* and *Anchor-Guided Evidence Decomposition*. The holistic component predicts scores from task-conditioned multimodal representations, while the decomposition component introduces latent evidence queries that interact with textual and visual token representations. These queries are organized into shared and anchor-specific groups to capture reusable and anchor-related evidence, respectively. At inference, the predictions from the two components are combined to obtain the final assessment score, benefiting from both improved holistic MLLM scoring and evidence-grounded assessment for better alignment with human ratings.

3.3 Holistic MLLM Scoring

The holistic branch shown in Fig. 2 is designed to retain the broad assessment capability of the pretrained MLLM while adapting its representation to human-rated image assessment. Rather than fully fine-tuning the backbone, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) to efficiently adapt selected attention projections. For a pretrained weight matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$, LoRA parameterizes the adapted weight as $\mathbf{W}' = \mathbf{W} + \frac{\alpha}{r}\mathbf{BA}$, where $\mathbf{A} \in \mathbb{R}^{r\times k}$ and $\mathbf{B} \in \mathbb{R}^{d\times r}$ are trainable low-rank matrices, r denotes the rank, and α controls the update scale. We apply these low-rank updates to the query, key, value, and output projections of the attention layers while keeping the pretrained backbone parameters fixed.

For each assessment criterion τ, the adapted MLLM produces a task-conditioned multimodal hidden sequence $\mathbf{H}^\tau$ at its final layer. We use the hidden representation $\mathbf{h}^\tau$ at the score-prediction position in $\mathbf{H}^\tau$ for holistic MLLM scoring. Meanwhile, the hidden states corresponding to the textual tokens and visual tokens in $\mathbf{H}^\tau$ are extracted as the textual representation $\mathbf{T}^\tau$ and visual representation $\mathbf{V}^\tau$, respectively, and are subsequently used for evidence decomposition.

Following prior work (Wu et al., 2024; You et al., 2025), we formulate holistic assessment as score-token prediction, aligning the scoring process with the native token-prediction mechanism of the MLLM while retaining a distribution over possible rating levels. The assessment instruction prompts the MLLM to output a single rating from 1 to 5. Let $\mathcal{S} = \{1, 2, 3, 4, 5\}$ denote the corresponding score-token set. We first transform $\mathbf{h}^\tau$ through a lightweight score-space projection and then pass it to the pretrained language-model head. The logits corresponding to the score tokens in $\mathcal{S}$ are

normalized to form the score distribution:

$$\pi_s^\tau = \frac{\exp\big(\mathbf{w}_s^\top f_{\text{proj}}(\mathbf{h}^\tau)\big)}{\sum_{j\in\mathcal{S}} \exp\big(\mathbf{w}_j^\top f_{\text{proj}}(\mathbf{h}^\tau)\big)}, \qquad s \in \mathcal{S}, \tag{2}$$

where $f_{\text{proj}}(\cdot)$ denotes the scoring projection and $\mathbf{w}_s$ denotes the output-head weight vector associated with score token s in the pretrained language model. Since human supervision is provided as a continuous MOS, directly using a hard categorical target would discard its relative proximity to neighboring rating levels. We therefore represent the MOS as a Gaussian-shaped soft distribution over the same score space:

$$q_s^\tau = \frac{\exp\Big(-\frac{(s-y^\tau)^2}{2\sigma^2}\Big)}{\sum_{j\in\mathcal{S}} \exp\Big(-\frac{(j-y^\tau)^2}{2\sigma^2}\Big)}, \qquad s \in \mathcal{S}, \tag{3}$$

where y^τ denotes the MOS for criterion τ and σ controls the spread of the target distribution. We then optimize the holistic branch by minimizing the KL divergence between the target and predicted score distributions:

$$\mathcal{L}_{\text{hol}}^\tau = D_{\text{KL}}\left(\mathbf{q}^\tau \,\|\, \boldsymbol{\pi}^\tau\right). \tag{4}$$

3.4 Anchor-Guided Evidence Decomposition

The holistic branch derives assessment scores from a unified task-conditioned MLLM representation. However, different assessment judgments can rely on different forms of evidence, which are not explicitly distinguished within such holistic representations. As illustrated in Fig. 2, we therefore introduce an anchor-guided decomposition component that organizes judgment-relevant evidence through a compact set of learnable latent queries.

Anchor-Structured Evidence Acquisition. Different assessment judgments may rely on broadly shared information while requiring distinct evidence associated with their underlying modality dependencies. To capture both common and anchor-specific evidence, we organize a compact set of learnable latent queries into shared and anchor-specific groups, without assigning predefined semantic meanings to individual queries. Formally, let $\mathcal{Q} = \{\mathbf{q}_1, \ldots, \mathbf{q}_K\}$ denote K learnable latent queries for evidence decomposition. We partition them into three disjoint groups:

$$\mathcal{Q} = \mathcal{Q}^s \cup \mathcal{Q}^c \cup \mathcal{Q}^v, \tag{5}$$

where $\mathcal{Q}^s$ denotes the shared query group, while $\mathcal{Q}^c$ and $\mathcal{Q}^v$ are anchor-specific queries associated with the cross-modal and visual evidence anchors, respectively.

For criterion τ, we activate the shared queries together with the anchor-specific group corresponding to its evidence anchor, i.e., $\mathcal{Q}^s \cup \mathcal{Q}^c$ for a cross-modal anchor and $\mathcal{Q}^s \cup \mathcal{Q}^v$ for a visual anchor. We denote the resulting active query set by $\mathcal{Q}_\tau$. The active queries then interact with the MLLM representations to acquire supporting evidence. Let $\mathbf{T}^\tau$ and $\mathbf{V}^\tau$ denote the textual and visual token representations, respectively, then for each active query $\mathbf{q}_k \in \mathcal{Q}_\tau$, we obtain an intermediate representation as

$$\mathbf{a}_k^\tau = \begin{cases} \text{CrossAttn}(\mathbf{q}_k, \mathbf{T}^\tau, \mathbf{T}^\tau), & \mathbf{q}_k \in \mathcal{Q}^s \cup \mathcal{Q}^c, \\ f_{\text{v}}(\mathbf{q}_k), & \mathbf{q}_k \in \mathcal{Q}^v. \end{cases} \tag{6}$$

where $f_{\text{v}}(\cdot)$ projects the visual-specific queries into the visual interaction space. The shared and cross-modal queries are first contextualized by $\mathbf{T}^\tau$, enabling them to incorporate textual context before retrieving visual evidence. In contrast, visual-specific queries bypass this textual interaction and directly enter the visual pathway.

All queries in $\mathcal{Q}_\tau$ then retrieve spatial evidence from the generated image:

$$\mathbf{e}_k^\tau = \text{CrossAttn}(\mathbf{a}_k^\tau, \mathbf{V}^\tau, \mathbf{V}^\tau), \tag{7}$$

where $\mathbf{a}_k^\tau$ serves as the query and $\mathbf{V}^\tau$ provides the keys and values.

Adaptive Evidence Aggregation. For each assessment criterion τ, prediction is formed from the query subset $\mathcal{Q}_\tau$ determined by its evidence anchor. While this subset specifies which evidence can contribute to the judgment, the importance of individual queries may vary across samples. We

therefore separate query-level scoring from sample-adaptive contribution estimation, allowing each query to provide a scoring signal whose influence is determined dynamically for the current sample.

Specifically, for each $\mathbf{q}_k \in \mathcal{Q}_\tau$, we combine its intermediate representation and retrieved visual evidence as $\mathbf{u}_k^\tau = [\mathbf{a}_k^\tau; \mathbf{e}_k^\tau]$. Each assessment criterion is equipped with its own scoring and gating functions. Within criterion τ, a query-specific scoring head $f_{\text{score},k}^\tau(\cdot)$ predicts a scalar scoring signal, while a criterion-specific gating network $f_{\text{gate}}^\tau(\cdot)$ estimates the sample-dependent importance of each query:

$$s_k^\tau = f_{\text{score},k}^\tau(\mathbf{u}_k^\tau), \qquad g_k^\tau = f_{\text{gate}}^\tau(\mathbf{u}_k^\tau). \tag{8}$$

The gating scores are normalized over $\mathcal{Q}_\tau$ to obtain the aggregation weights:

$$\alpha_k^\tau = \frac{\exp(g_k^\tau)}{\sum_{\mathbf{q}_j \in \mathcal{Q}_\tau} \exp(g_j^\tau)}. \tag{9}$$

The resulting criterion-level logit is obtained by weighted aggregation:

$$z^\tau = \sum_{\mathbf{q}_k \in \mathcal{Q}_\tau} \alpha_k^\tau s_k^\tau. \tag{10}$$

Learning Decomposed Evidence. To align the aggregated prediction from decomposition component with the absolute human rating, we first supervise it using the criterion-level MOS. Since z^τ is defined in logit space, we normalize the MOS $y^\tau \in [1, 5]$ to $[0, 1]$ and supervise it using

$$\mathcal{L}_{\text{dec}}^\tau = \text{BCEWithLogits}\left(z^\tau, \frac{y^\tau - 1}{4}\right). \tag{11}$$

Absolute score supervision alone does not directly constrain the relative ordering between samples. We therefore introduce a pairwise ranking objective to preserve the ordering implied by human judgments. Let $\mathcal{P}^\tau$ denote the set of sample pairs (m, n) for criterion τ satisfying $y_m^\tau > y_n^\tau$. The ranking loss is defined as

$$\mathcal{L}_{\text{rank}}^\tau = \frac{1}{|\mathcal{P}^\tau|} \sum_{(m,n)\in\mathcal{P}^\tau} \left[\delta - (z_m^\tau - z_n^\tau)\right]_+, \tag{12}$$

where δ denotes the ranking margin and $[u]_+ = \max(0, u)$. Different latent queries may nevertheless encode redundant evidence, limiting the benefit of decomposing the assessment into multiple query representations. To encourage complementary evidence across queries, we further regularize the features $\mathbf{u}_k^\tau$ before the scoring and gating modules:

$$\mathcal{L}_{\text{div}}^\tau = \frac{1}{|\mathcal{R}_\tau|} \sum_{(i,j)\in\mathcal{R}_\tau} \text{sim}\left(\mathbf{u}_i^\tau, \mathbf{u}_j^\tau\right), \tag{13}$$

where $\mathcal{R}_\tau$ denotes the set of distinct query pairs in $\mathcal{Q}_\tau$.

3.5 Training Objective and Inference

The overall training objective combines the holistic and decomposition objectives over all assessment criteria included in the training set:

$$\mathcal{L} = \sum_{a\in\mathcal{A}} \left(\mathcal{L}_{\text{hol}}^\tau + \mathcal{L}_{\text{dec}}^\tau + \mathcal{L}_{\text{rank}}^\tau + \mathcal{L}_{\text{div}}^\tau\right), \tag{14}$$

where $\mathcal{A}$ denotes the set of assessment criteria available in the current training setting.

At inference, the holistic score is recovered from the predicted score-token distribution, while the decomposition logit is mapped to the same $[1, 5]$ rating range:

$$\hat{y}_{\text{hol}}^\tau = \sum_{s\in\mathcal{S}} s\,\pi_s^\tau, \qquad \hat{y}_{\text{dec}}^\tau = 1 + 4\ \text{sigmoid}(z^\tau). \tag{15}$$

The final prediction is obtained by averaging the two estimates:

$$\hat{y}^\tau = \frac{\hat{y}_{\text{hol}}^\tau + \hat{y}_{\text{dec}}^\tau}{2}. \tag{16}$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We conduct experiments on six AI-generated image assessment benchmarks covering diverse assessment criteria and evidence dependencies. For the *cross-modal evidence anchor*, we evaluate text–image correspondence on AGIQA-3K (Li et al., 2023), AIGCIQA2023 (Wang et al., 2023), T2I-CompBench (Huang et al., 2023), and GenAI-Bench (Li et al., 2024), and human preference on HPDv2 (Wu et al., 2023). For the *visual-specific evidence anchor*, we evaluate perceptual quality on AGIQA-3K and AIGCIQA2023, authenticity on AIGCIQA2023, and naturalness on AGIN (Chen et al., 2024).

We consider two evaluation scenarios. In **standard evaluation**, all annotated criteria in AGIQA-3K and AIGCIQA2023 are jointly optimized within each dataset, with two criteria for AGIQA-3K and three for AIGCIQA2023, and evaluated on the corresponding test sets. Since CompBench, GenAI-Bench, AGIN, and HPDv2 involve distinct assessment tasks, EAGLE is separately tuned on each dataset. We further evaluate **cross-dataset generalization** between AGIQA-3K and AIGCIQA2023 by training on one dataset and directly testing on the other without target-domain adaptation. Additional dataset details are provided in `Appendix` A.

Competing Methods and Performance Metrics. We compare EAGLE with representative state-of-the-art (SOTA) image quality assessment methods across different modeling paradigms, including conventional and representation-based methods, MUSIQ (Ke et al., 2021), MANIQA (Yang et al., 2022), and Re-IQA (Saha et al., 2023); vision-language (VL)-based judges, LIQE (Zhang et al., 2023), IPCE (Peng et al., 2024), and DPGF-Net (Li et al., 2026a); and MLLM-based judges, DeQA-Score (You et al., 2025) and M3-AGIQA (Cui et al., 2025). We also include Qwen2.5-VL-7B (Bai et al., 2025) as the MLLM backbone baseline.

We report official results when available. For AGIQA-3K and AIGCIQA2023, all results are taken from the corresponding original works except DeQA-Score, for which such results are unavailable. For the other four benchmarks, we evaluate all methods using the same data splits as EAGLE for fair comparison; the same protocol is used for cross-dataset evaluation. Missing entries (*i.e.*, '–') indicate settings where evaluation is infeasible due to unavailable required inputs or resources; see `Appendix` B.3 for details. Following prior studies (Li et al., 2023; Wu et al., 2024; Cui et al., 2025), we report Spearman's rank-order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC) to measure rank consistency and linear correlation with human ratings, respectively. We also provide a computational efficiency comparison in `Appendix` C.1.

Implementation Details. EAGLE uses Qwen2.5-VL-7B-Instruct (Bai et al., 2025) as the frozen backbone and trains LoRA adapters with rank 128 and scaling factor 256. The evidence decomposition module uses $K = 8$ latent queries: two shared, three cross-modal, and three visual-specific. See `Appendix` B.2 for details.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Standard Evaluation. As shown in Table 1, EAGLE consistently achieves superior performance across assessment criteria grounded in different evidence dependencies. For the baselines, traditional IQA methods (MUSIQ, MANIQA, and Re-IQA) remain competitive on visual-specific criteria but struggle when cross-modal evidence is required. VL-based methods (LIQE and IPCE) and more recent VL/MLLM-based evaluators (DeQA-Score, DPGF-Net, and M3-AGIQA) generally improve cross-modal and overall assessment, yet notable gaps remain on challenging correspondence benchmarks such as GenAI-Bench. Notably, M3-AGIQA further relies on external online MLLMs to enhance aspect-specific evaluation. In contrast, EAGLE requires no external MLLM-generated supervision and remains strong across both cross-modal and visual-specific evidence anchors. In particular, EAGLE improves over the strongest competing results by up to 4.3 SRCC and 9.3 PLCC points. These results validate that EAGLE is capable of generating the most faithful scores via explicit modeling of the two key assessment evidence anchors.

Cross-Dataset Generalization. Table 2 evaluates cross-dataset generalization using correspondence and perceptual quality, representing cross-modal and visual-specific evidence anchors, respectively. Most existing methods exhibit noticeable degradation due to the cross-dataset distribution shifts, particularly for correspondence assessment. In contrast, EAGLE consistently achieves

Table 1: Comparison results (SRCC, PLCC) with state-of-the-art IQA, VL, and MLLM methods across nine visual/cross-modal image assessment tasks. Best and second-best results are highlighted in **bold** and underline, respectively. '–' denotes the methods are not applicable.

Method	AGIQA-3K		AIGCIQA2023			CompBench	GenAI-Bench	AGIN	HPDv2
	Corr.	Quality	Corr.	Quality	Authenticity	Corr.	Corr.	Naturalness	Preference
MUSIQ	(62.9, 78.4)	(83.5, 88.6)	(76.2, 75.3)	(84.2, 86.0)	(69.5, 69.0)	(62.9, 61.5)	(39.0, 35.4)	(72.8, 71.9)	(76.8, 76.6)
MANIQA	(76.0, 84.9)	(87.2, 91.0)	(76.5, 75.6)	(85.7, 88.1)	(79.9, 78.7)	(76.5, 76.7)	(49.8, 48.3)	(80.3, 79.6)	(82.8, 82.8)
Re-IQA	(63.7, 78.8)	(81.9, 88.0)	(64.3, 63.6)	(81.4, 83.2)	(60.2, 59.7)	(40.2, 38.3)	(77.5, 77.2)	(76.5, 77.3)	(78.8, 78.1)
LIQE	(65.2, 74.0)	(86.6, 90.0)	(77.1, 75.9)	(85.1, 87.0)	(79.0, 77.5)	(75.3, 74.7)	(48.1, 48.1)	(81.1, 80.6)	(79.4, 79.4)
IPCE	(77.0, 87.3)	(88.1, 92.1)	(79.8, 78.9)	(86.4, 87.9)	(79.0, 78.3)	(84.4, 84.4)	(69.7, 69.6)	–	(84.4, 84.7)
DPGF-Net	(80.1, 87.0)	(90.1, 92.9)	(82.8, 81.2)	(88.1, 89.8)	(80.9, 79.8)	(84.1, 84.1)	(74.4, 74.2)	(81.6, 82.5)	(85.1, 85.1)
Qwen2.5-VL-7B	(68.7, 65.7)	(71.4, 69.3)	(67.0, 71.3)	(64.1, 71.6)	(50.6, 47.3)	(59.9, 56.5)	(66.5, 65.8)	(41.8, 42.9)	(54.7, 47.5)
M3-AGIQA	(85.2, 91.4)	(**90.5**, 93.2)	(80.6, 79.7)	(86.2, 89.2)	(**82.8**, **81.7**)	–	–	–	–
DeQA-Score	(69.9, 81.2)	(90.3, 92.7)	(70.7, 70.6)	(86.8, 88.9)	(79.0, 78.6)	(73.2, 71.8)	(60.3, 60.5)	(81.7, 82.1)	(83.6, 83.9)
EAGLE	(**87.1**, **91.5**)	(90.4, **93.3**)	(**85.2**, **84.6**)	(**88.2**, **89.9**)	(80.5, 80.4)	(**88.7**, **88.6**)	(**83.4**, **83.5**)	(**83.3**, **84.6**)	(**85.3**, **85.5**)

Table 2: Cross-dataset generalization comparison. Results for Qwen2.5-VL-7B are zero-shot.

Method	AGIQA-3K → AIGCIQA2023		AIGCIQA2023 → AGIQA-3K	
	Corr.	Quality	Corr.	Quality
MUSIQ	(52.5, 51.5)	(65.0, 64.3)	(39.4, 43.7)	(63.5, 67.3)
MANIQA	(57.8, 55.9)	(73.1, 73.0)	(55.0, 59.3)	(71.5, 75.8)
Re-IQA	(47.9, 48.4)	(65.4, 65.4)	(24.3, 15.4)	(47.3, 35.2)
LIQE	(49.3, 51.9)	(68.6, 71.0)	(50.0, 48.3)	(70.8, 70.8)
IPCE	(61.1, 60.5)	(76.9, 76.5)	(65.2, 72.3)	(71.4, 75.2)
DPGF-Net	(62.3, 61.4)	(76.8, 78.4)	(65.4, 73.3)	(77.1, 82.4)
Qwen2.5-VL-7B	(64.1, 71.6)	(50.6, 47.3)	(68.7, 65.7)	(71.4, 69.3)
M3-AGIQA	(74.2, 72.5)	(74.9, 74.6)	(73.6, 77.2)	(74.3, 75.4)
DeQA-Score	(58.7, 59.2)	(77.8, 79.3)	(51.5, 58.2)	(75.0, 79.4)
EAGLE	(**80.6**, **79.6**)	(**81.8**, **82.0**)	(**81.7**, **86.3**)	(**82.0**, **86.9**)

Table 3: Loss ablation (SRCC, PLCC).

$\mathcal{L}_{\text{rank}}$	$\mathcal{L}_{\text{div}}$	AGIQA-3K		AIGCIQA2023 → AGIQA-3K	
		Corr.	Quality	Corr.	Quality
×	×	(84.4, 90.2)	(89.0, 92.1)	(80.1, 84.6)	(79.2, 85.2)
✓	×	(86.3, 90.6)	(89.4, 92.3)	(80.5, 85.3)	(80.0, 85.2)
×	✓	(86.4, 90.9)	(89.9, 92.6)	(80.3, 85.0)	(81.2, 85.5)
✓	✓	(**87.1**, **91.5**)	(**90.4**, **93.3**)	(**81.7**, **86.3**)	(**82.0**, **86.9**)

Table 4: Evidence sensitivity analysis.

Type	Intervention	Corr. $\lvert\Delta\rvert$	Quality $\lvert\Delta\rvert$
Prompt Intervention	Prompt Swap	**1.095**	0.125
	Null Prompt	**0.471**	0.057
Visual Degradation	Gaussian Blur	0.282	**0.669**
	Gaussian Noise	0.287	**0.578**
	JPEG Compression	0.188	**0.325**

the best results in both transfer directions and across both evidence anchors, surpassing the second best method, M3-AGIQA (which remains highly competitive on cross-modal assessment benefiting from additional MLLM-generated reasoning supervision), by up to 7.4 SRCC and 8.8 PLCC points. This consistent advantage suggests that the evidence dependencies learned by EAGLE generalize beyond dataset-specific assessment patterns due to the complementary holistic MLLM scoring and anchor-guided evidence decomposition.

4.3 Ablation Study

Loss Ablation. Table 3 studies the effects of the ranking and diversity objectives. Each objective individually improves the baseline across most settings, while their combination consistently performs best under both standard and cross-dataset evaluations. $\mathcal{L}_{\text{rank}}$ brings clear gains in correspondence assessment, whereas $\mathcal{L}_{\text{div}}$ provides particularly consistent improvements in perceptual quality. Their complementary benefits are more evident under cross-dataset transfer, jointly preserving relative assessment consistency and promoting diverse evidence modeling.

Holistic MLLM Scoring and Evidence Decomposition. Table 5 evaluates the key components of EAGLE. We use Qwen2.5-VL-7B-Instruct with LoRA as the '**Baseline**', which adapts only the holistic MLLM representation for image assessment. We then retain only holistic MLLM scoring ('**EAGLE**$_H$**'**) or evidence decomposition ('**EAGLE**$_D$') in our model. '**EAGLE**$_H$' consistently outperforms the '**Baseline**', with particularly clear gains under cross-dataset evaluation, indicating that joint learning in EAGLE strengthens the holistic representation. '**EAGLE**$_D$' yields comparable improvements, while combining both components in the full '**EAGLE**' brings further gains, highlighting their complementary roles in holistic assessment and structured evidence grounding.

Ablation within EAGLE$_D$. We further ablate the decomposition design. Removing anchor-dependent interaction (**w/o Anchor Interaction**) or sharing one query set across all criteria (**w/o Structured Queries**) generally degrades performance, supporting the benefits of anchor-aware evidence acquisition and shared/anchor-specific query modeling.

Joint-Criteria vs. Individual-Criterion Training. Training each criterion independently (**EAGLE**$_I$) also underperforms the full **EAGLE**, indicating that joint optimization across multiple criteria improves both MLLM representations and evidence-anchor learning.

Table 5: Ablation study of EAGLE under standard evaluation and cross-dataset generalization.

Variant	Standard Evaluation				Cross-Dataset Generalization			
	AGIQA-3K		AIGCIQA2023		AGIQA-3K → AIGCIQA2023		AIGCIQA2023 → AGIQA-3K	
	Corr.	Quality	Corr.	Quality	Corr.	Quality	Corr.	Quality
Baseline	(85.9, 89.6)	(87.4, 90.7)	(82.1, 80.7)	(85.0, 86.6)	(77.7, 77.1)	(79.7, 79.8)	(75.2, 81.6)	(75.9, 81.8)
EAGLE$_H$	(86.8, 91.3)	(89.7, 91.7)	(82.6, 81.2)	(85.7, 87.3)	(80.4, 77.1)	(81.7, 81.5)	(77.8, 82.9)	(77.0, 82.9)
EAGLE$_D$	(86.8, 91.0)	(89.5, 91.9)	(82.5, 81.9)	(84.2, 85.4)	(79.6, 77.7)	(80.4, 81.0)	(76.8, 82.2)	(79.5, 85.0)
w/o Anchor Interaction	(84.4, 89.3)	(89.9, 92.6)	(81.6, 81.2)	(87.5, 89.3)	(79.4, 78.7)	(81.4, 81.1)	(76.2, 80.9)	(79.1, 84.5)
w/o Structured Queries	(85.5, 90.2)	(90.2, 92.6)	(84.5, 83.9)	(86.8, 88.5)	(80.0, 78.9)	(81.4, 81.2)	(80.8, 85.5)	(80.1, 85.1)
EAGLE$_I$	(85.5, 90.5)	(89.8, 92.5)	(83.9, 83.4)	(87.2, 88.8)	(79.6, 78.4)	(80.4, 80.3)	(81.2, 86.2)	(81.4, 86.4)
EAGLE	(**87.1**, **91.5**)	(**90.4**, **93.3**)	(**85.2**, **84.6**)	(**88.2**, **89.9**)	(**80.6**, **79.6**)	(**81.8**, **82.0**)	(**81.7**, **86.3**)	(**82.0**, **86.9**)

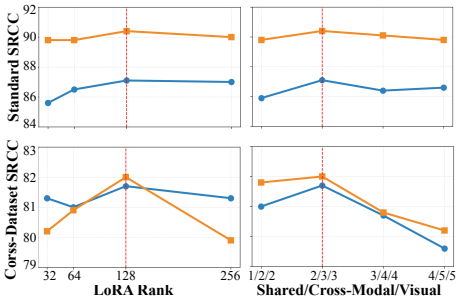

Figure 3: Sensitivity analysis of the LoRA rank and evidence query configuration.

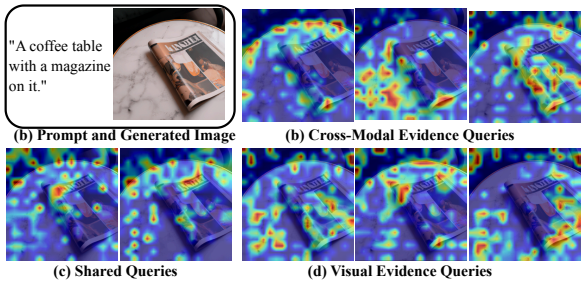

Figure 4: Visualization of evidence attention across shared, cross-modal, and visual-specific query groups.

4.4 Further Analysis of EAGLE

Hyperparameter Sensitivity Analysis. Fig. 3 examines the effects of the LoRA rank and evidence query configuration. EAGLE remains relatively stable across different LoRA ranks, with $r = 128$ achieving the best performance under both standard and cross-dataset evaluation. Similarly, the default $(2, 3, 3)$ query configuration performs best across the reported settings, while either reducing or substantially increasing the number of queries leads to performance degradation. These results suggest that a moderate query capacity is sufficient for effective evidence modeling.

Evidence Sensitivity under Controlled Interventions. We examine how correspondence and quality scores respond when perturbing either textual or visual input while keeping the other modality fixed. As shown in Table 4, prompt replacement changes the correspondence score much more than the quality score (1.095 vs. 0.125), whereas image degradation via blur, noise, or JPEG compression causes larger changes in quality than correspondence. This indicates that correspondence is more sensitive to cross-modal evidence, while quality is more sensitive to visual evidence, consistent with our evidence-anchor formulation. Additional analysis and results are provided in `Appendix` C.2.

Visualization of Evidence Attention. Fig. 4 visualizes the spatial attention patterns of different latent query groups in the generated image for the input prompt "*A coffee table with a magazine on it.*" The **shared queries** exhibit relatively broad attention over the main scene structure, covering both the magazine and surrounding table regions. The **cross-modal queries** focus more on prompt-relevant semantic evidence, including the magazine, tabletop, and their spatial relation, whereas the **visual-specific queries** respond more strongly to local appearance cues such as object boundaries, printed regions, and surface details. These distinct attention patterns are consistent with different evidence roles assigned to the three query groups.

5 Conclusion

In this paper, we introduce **evidence anchors** to characterize distinct evidence dependencies in AI-generated image assessment and propose **EAGLE**, which combines holistic MLLM scoring with anchor-guided evidence decomposition. EAGLE organizes latent queries into shared and anchor-specific groups to acquire judgment-relevant evidence. Extensive experiments validate its effectiveness in both within-dataset evaluation and cross-dataset generalization. Looking forward, we view evidence anchors as a general perspective for building more grounded and interpretable MLLM evaluators by explicitly linking each judgment to its supporting evidence. Beyond AI-generated image assessment, this perspective may extend to broader multimodal generation and understanding tasks.

AI USE STATEMENT

In this work, we used generative AI tools solely for language editing, including grammar checking and improving the clarity and readability of the manuscript. We did not use generative AI tools to generate or process data, develop research ideas or conceptual frameworks, formulate mathematical claims, design the methodology or experiments, implement the proposed method, or analyze and interpret experimental results. All AI-assisted edits were reviewed and verified by the authors. We take full responsibility for the final content of this work, including all text, claims, results, and artifacts.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 7, 14, 15

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

Zheng Chen, Xun Zhang, Wenbo Li, Renjing Pei, Fenglong Song, Xiongkuo Min, Xiaohong Liu, Xin Yuan, Yong Guo, and Yulun Zhang. Grounding-iqa: Grounding multimodal language model for image quality assessment. In *International Conference on Learning Representations*, 2026. 3

Zijian Chen, Wei Sun, Haoning Wu, Zicheng Zhang, Jun Jia, Ru Huang, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Study of subjective and objective naturalness assessment of ai-generated images. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. doi: 10.1109/TCSVT.2024.3509032. 7, 13

Chuan Cui, Kejiang Chen, Zhihua Wei, Wen Shen, Weiming Zhang, and Nenghai Yu. M3-agiqa: Multimodal, multi-round, multi-aspect ai-generated image quality assessment. *arXiv preprint arXiv:2502.15167*, 2025. 1, 3, 7, 15

Kanglong Fan, Tianhe Wu, Wen Wen, Jianzhao Liu, Le Yang, Yabin Zhang, Yiting Liao, Junlin Li, and Li Zhang. ME-IQA: memory-enhanced image quality assessment via re-ranking. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision - ECCV 2026 - 19th European Conference, Malmö, Sweden, September 8-12, 2026, Proceedings, Part LI*, volume 17051 of *Lecture Notes in Computer Science*, pp. 336–355. Springer, 2026. doi: 10.1007/978-3-032-37356-4_19. URL https://doi.org/10.1007/978-3-032-37356-4_19. 3

Xi Fang, Weigang Wang, Xiaoxin Lv, and Jun Yan. Pcqa: A strong baseline for aigc quality assessment based on prompt condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 6167–6176, 2024. 3

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 4

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 7, 13

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021. 1, 3, 7, 15

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024. 7, 13

Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Agiqa-3k: An open database for ai-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. doi: 10.1109/TCSVT.2023.3319020. 7, 13

Chunyi Li, Yuan Tian, Xiaoyue Ling, Zicheng Zhang, Haodong Duan, Haoning Wu, Ziheng Jia, Xiaohong Liu, Xiongkuo Min, Guo Lu, Weisi Lin, and Guangtao Zhai. Image quality assessment: From human to machine preference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7570–7581, 2025a. 3

Tao Li, Xingran Liao, and Mingliang Zhou. Dpgf-net: Dual-prior guided fusion network for joint assessment of perceptual quality and semantic consistency in ai-generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19529–19538, 2026a. 3, 7, 15

Xudong Li, Zihao Huang, Yan Zhang, Yunhang Shen, Ke Li, Xiawu Zheng, Liujuan Cao, and Rongrong Ji. Few-shot image quality assessment via adaptation of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10442–10452, 2025b. 3

Zhuangzi Li, Jian Jin, Shilv Cai, and Weisi Lin. R4-cgqa: Retrieval-based vision language models for computer graphics image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9468–9477, 2026b. 3

Guoqiang Liang, Jianyi Wang, Zhonghua Wu, Shangchen Zhou, and Chen Change Loy. Zoom-iqa: Image quality assessment with reliable region-aware reasoning. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision - ECCV 2026 - 19th European Conference, Malmö, Sweden, September 8-12, 2026, Proceedings, Part VIII*, volume 17008 of *Lecture Notes in Computer Science*, pp. 618–636. Springer, 2026. doi: 10.1007/978-3-032-37026-6_34. URL https://doi.org/10.1007/978-3-032-37026-6_34. 3

Jingwei Liu, Bo Liu, Hongyan Li, Tianlin Zhang, Yifei Qian, Congyang Zhao, Junhong Liu, Yang Cai, and Ling Yang. Reinforcing vision-language models for image quality assessment with grounding process rewards. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision - ECCV 2026 - 19th European Conference, Malmö, Sweden, September 8-12, 2026, Proceedings, Part LXXVIII*, volume 17078 of *Lecture Notes in Computer Science*, pp. 351–368. Springer, 2026. doi: 10.1007/978-3-032-37393-9_21. URL https://doi.org/10.1007/978-3-032-37393-9_21. 3

OpenAI. Introducing our latest image generation model in the api. https://openai.com/index/image-generation-api/, April 2025. Accessed: 2026-09-25. 1

Fei Peng, Huiyuan Fu, Anlong Ming, Chuanming Wang, Huadong Ma, Shuai He, Zifei Dou, and Shu Chen. Aigc image quality assessment via image-prompt correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 6432–6441, 2024. 1, 3, 7, 15

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024. 1

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. 1

Avinab Saha, Sandeep Mishra, and Alan C. Bovik. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5846–5855, 2023. 1, 3, 7, 15

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022. 1

Wenhao Shen, Mingliang Zhou, Yu Chen, Xuekai Wei, Yong Feng, Huayan Pu, and Weijia Jia. Image quality assessment: Investigating causal perceptual effects with abductive counterfactual inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17990–17999, 2025. 3

Long Tang, Huiyu Duan, Guoquan Zheng, Jianbo Zhang, Jie Hao, and Liang Yuan. Life-iqa: Boosting blind image quality assessment through gcn-enhanced layer interaction and moe-based feature decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29866–29876, 2026. 3

Jiarui Wang, Huiyu Duan, Jing Liu, Shi Chen, Xiongkuo Min, and Guangtao Zhai. Aigciqa2023: A large-scale image quality assessment database for ai generated images: From the perspectives of quality, authenticity and correspondence. In *Artificial Intelligence – Third CAAI International Conference, CICAI 2023*, pp. 46–57. Springer, 2023. doi: 10.1007/978-981-99-9119-8_5. 7, 13

Wen Wen, Tianwu Zhi, Kanglong Fan, Yang Li, Xinge Peng, Yabin Zhang, Yiting Liao, Junlin Li, and Li Zhang. Self-evolving vision-language models for image quality assessment via voting and ranking. volume abs/2509.25787, 2025. doi: 10.48550/ARXIV.2509.25787. URL https://doi.org/10.48550/arXiv.2509.25787. 3

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. 1, 3, 4, 7

Jinjian Wu, Jiaqi Tang, Wei Wei, Yingying Yan, Jianmin Chen, Botong Geng, Lei Zhang, and Qifeng Chen. IQA-T1: tool-based visual evidence reasoning for image quality assessment. volume abs/2607.12375, 2026. doi: 10.48550/ARXIV.2607.12375. URL https://doi.org/10.48550/arXiv.2607.12375. 3

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 7, 13

Jili Xia, Lihuo He, Cheng Deng, Leida Li, and Xinbo Gao. Ai-generated image quality assessment based on task-specific prompt and multi-granularity similarity. *IEEE Transactions on Image Processing*, 34:8216–8228, 2025. doi: 10.1109/TIP.2025.3639984. 3

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1191–1200, 2022. 1, 3, 7, 15

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14483–14494, 2025. 1, 3, 4, 7, 15

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14071–14081, 2023. 3, 7, 15

A Dataset Details

We evaluate EAGLE on six AI-generated image assessment benchmarks covering different assessment criteria and evidence anchors. For the cross-modal evidence anchor, we consider text–image correspondence on AGIQA-3K (Li et al., 2023), AIGCIQA2023 (Wang et al., 2023), T2I-CompBench (Huang et al., 2023), and GenAI-Bench (Li et al., 2024), as well as human preference on HPDv2 (Wu et al., 2023). For the visual evidence anchor, we evaluate perceptual quality on AGIQA-3K and AIGCIQA2023, authenticity on AIGCIQA2023, and naturalness on AGIN (Chen et al., 2024).

AGIQA-3K contains 2,982 AI-generated images produced by six text-to-image generation models, with human ratings for perceptual quality and text–image correspondence. We follow the data partition used by DPGF-Net and jointly train the correspondence and quality criteria on the shared training images.

AIGCIQA2023 contains 2,400 AI-generated images produced by six text-to-image models, with subjective annotations for perceptual quality, authenticity, and text–image correspondence. We follow the DPGF-Net partition for correspondence and quality, while authenticity is trained using its corresponding annotation split.

T2I-CompBench is designed for compositional text-to-image evaluation and contains 6,000 prompts spanning color, shape, texture, spatial relations, non-spatial relations, and complex compositions. For our experiments, we use the released human-rated subset for correspondence assessment, comprising 2,400 image–text instances, of which 1,920 are used for training and 480 for testing.

GenAI-Bench contains 1,600 prompts and 9,600 generated images with human ratings of prompt–image correspondence. After preprocessing, our experimental set contains 9,598 valid instances, with 8,638 used for training and 960 for testing.

AGIN contains 6,049 AI-generated images collected from 18 generative model variants across five generation tasks, together with 200 real images, resulting in 6,249 images in total. We evaluate visual naturalness using a 7:3 train–test split, yielding 4,234 training images and 1,815 test images.

HPDv2 is a large-scale human preference dataset containing more than 430K generated images and approximately 798K pairwise preference annotations conditioned on textual prompts. For our experiments, we use a subset of 3,700 preference pairs grouped by prompt. We perform an 80:20 group-level split, resulting in 2,956 training pairs from 320 prompt groups and 744 test pairs from 80 prompt groups.

Table 6 summarizes the datasets, assessment criteria, and evaluation partitions used in our experiments. For AGIQA-3K and AIGCIQA2023, multiple annotated criteria are jointly optimized within each dataset. For T2I-CompBench, GenAI-Bench, AGIN, and HPDv2, we train the model separately for the corresponding assessment criterion.

Table 6: Training and testing statistics of the datasets used in our experiments.

Dataset	Data Modality	Assessment Criterion	Dataset Scale	Train Instances	Test Instances
AGIQA-3K	Cross-modal	Correspondence	2,982 images	4,764	593
	Visual	Quality		4,764	600
AIGCIQA2023	Cross-modal	Correspondence	2,400 images	3,838	480
	Visual	Quality		3,838	480
	Visual	Authenticity		1,920	480
T2I-CompBench	Cross-modal	Correspondence	2,400 instances	1,920	480
GenAI-Bench	Cross-modal	Correspondence	9,598 instances	8,638	960
AGIN	Visual	Naturalness	6,249 images	4,234	1,815
HPDv2	Cross-modal	Preference	3,700 pairs	2,956	744

B Implementation Details

B.1 Criterion-Specific Instruction Templates

For each assessment criterion τ, we use a criterion-specific instruction to condition the MLLM on the corresponding assessment objective. The instructions follow a consistent format and differ primarily in the evidence and judgment to be considered. For cross-modal criteria, the instruction explicitly refers to the relationship between the conditioning prompt and the generated image, whereas visual criteria focus on properties of the generated image itself. All score-based instructions use the same rating range from 1 to 5 to maintain a unified output space across criteria.

The complete instruction templates are provided below, with $\{p\}$ denoting the conditioning prompt.

Instruction Template (Text–Image Correspondence)
User: *Given the text prompt and the image, rate how well the image matches the prompt, considering the subjects, attributes, quantities, actions, relationships, and overall intent. Do not evaluate visual quality. Text prompt: "*$\{p\}$*" Respond with only one number from 1 to 5.*

Instruction Template (Perceptual Quality)
User: *Rate the overall perceptual quality of the image from 1 to 5, considering composition, clarity, detail, color, visual coherence, realism when appropriate, and noticeable visual flaws. Respond with only one number from 1 to 5.*

Instruction Template (Authenticity)
User: *Rate the authenticity of this image on a scale of 1 to 5 (1=obviously fake / heavily distorted, 5=looks completely real, no visible generation artifacts). Focus on whether objects look natural and undistorted, whether anatomy / geometry / physics are plausible, and whether textures and structures are free of AI-generation artifacts. Answer with a single digit.*

Instruction Template (Naturalness)
User: *Rate the naturalness of this image on a scale of 1 to 5 (1=obviously unnatural, either technically distorted or semantically implausible; 5=looks completely natural, as if it could exist in the real world). Consider both technical rendering (luminance, contrast, detail, blur, artifacts) and content rationality (whether the depicted objects plausibly exist, whether colors are believable, whether the layout and context make sense, and whether the scene is perceptually clear). Weight content rationality more heavily than low-level technical rendering. Answer with a single digit.*

Instruction Template (Human Preference)
User: *The image was generated from the text prompt: "*$\{p\}$*". Rate your preference for this image on a scale of 1 to 5, as if you had to pick among candidate images generated for this same prompt (1=you would definitely not pick this image, 5=you would strongly prefer this image over typical alternatives). Consider how well the image matches the prompt intent, the visual aesthetics, and how appealing the composition is compared to what a typical AI-generated image for such a prompt would look like. Answer with a single digit.*

B.2 Implementation and Model Configuration

EAGLE is built upon Qwen2.5-VL-7B-Instruct (Bai et al., 2025). We adapt the MLLM using LoRA with rank $r = 128$ and scaling factor $\alpha = 256$, applied to its attention projection layers. The ViT backbone is kept frozen, while the visual merger and the newly introduced decomposition modules

are fully trainable. Overall, EAGLE contains approximately 220M trainable parameters out of 8.6B total parameters.

For anchor-guided evidence decomposition, we employ $K = 8$ learnable latent queries, consisting of two shared queries $\mathcal{Q}^s$, three cross-modal queries $\mathcal{Q}^c$, and three visual-specific queries $\mathcal{Q}^v$. For criteria associated with a cross-modal evidence anchor, the shared and cross-modal queries are activated, whereas criteria associated with a visual evidence anchor use the shared and visual-specific queries. Shared and cross-modal queries first interact with the textual token representations before retrieving visual evidence, while visual-specific queries directly enter the visual interaction stage. The resulting query representations are subsequently processed by the criterion-specific scoring and gating modules for decomposed prediction.

We optimize EAGLE using AdamW with a learning rate of 2×10^{-5}, cosine learning-rate decay, a warmup ratio of 3%, and a batch size of 32. All experiments are conducted on four AMD Instinct MI308X GPUs using BF16 precision and DeepSpeed ZeRO-2. Input images are resized such that the shorter side has 504 pixels. Unless otherwise specified, each dataset is trained for 3 epochs.

B.3 Implementation of Comparison Methods

For comparison methods with publicly available official implementations, we follow the released codebases whenever they are compatible with our experimental environment. The only exception is MUSIQ, whose official implementation is released in JAX/Flax; we instead use the widely adopted PyTorch reproduction `anse3832/MUSIQ`. For the remaining methods, we follow their official implementations or released model checkpoints, retaining the original architectures, preprocessing pipelines, pretrained initialization, and optimization settings whenever applicable. We modify only the dataset interfaces when reproduction on our evaluation splits is required. Results directly reported by the original papers are used when they exactly match our datasets and assessment criteria; otherwise, we reproduce the corresponding methods. Table 7 summarizes the implementations and model sources used in our experiments.

Table 7: Implementations and model sources of the comparison methods used in our experiments.

Method	**Category**	**Implementation / Model Source**
MUSIQ* (Ke et al., 2021)	IQA	`anse3832/MUSIQ`
MANIQA (Yang et al., 2022)	IQA	`IIGROUP/MANIQA`
Re-IQA (Saha et al., 2023)	IQA	`avinabsaha/ReIQA`
LIQE (Zhang et al., 2023)	Vision–Language	`zwx8981/LIQE`
IPCE (Peng et al., 2024)	Vision–Language	`pf0607/IPCE`
DPGF-Net (Li et al., 2026a)	Vision–Language	`leeto221/AGIQA`
Qwen2.5-VL-7B (Bai et al., 2025)	MLLM Backbone	`Qwen/Qwen2.5-VL-7B-Instruct`
M3-AGIQA (Cui et al., 2025)	MLLM-based	`strawhatboy/M3-AGIQA`
DeQA-Score (You et al., 2025)	MLLM-based	`zhiyuanyou/DeQA-Score`

*Third-party PyTorch reproduction; all other methods use their official code or released model checkpoints.

Unavailable Evaluation Settings. Some methods cannot be directly evaluated on all benchmarks because their released formulations require additional inputs or supervision that are unavailable in our evaluation setting. Specifically, for IPCE, a consistent conditioning text prompt is required, which is not available for samples in AGIN; for M3-AGIQA, extending the method to additional benchmarks requires aspect-specific intermediate descriptions generated by an external MLLM. We therefore report "–" for these unsupported settings rather than introducing additional inputs or supervision.

C Additional Results

C.1 Model Complexity of EAGLE vs. SotA Methods

We compare the computational complexity of EAGLE with representative AIGC image assessment methods in Table 8, considering the number of trainable parameters, per-GPU training time (ms/sample), and inference time (ms/sample). Compared with lightweight vision-language methods such as IPCE and DPGF-Net, EAGLE incurs higher computational cost due to its MLLM backbone. Nevertheless, among MLLM-based evaluators, EAGLE requires fewer trainable parameters than DeQA-Score, achieves faster inference than M3-AGIQA, and exhibits lower training time than both DeQA-Score and M3-AGIQA. Considering the substantial performance gains across diverse assessment criteria, this additional computation provides a reasonable trade-off between assessment performance and computational cost.

Table 8: Comparison of model complexity and computational efficiency. Training and inference times are measured in milliseconds per sample on a single GPU.

Model Type	Method	Trainable Params. (M)	Inference (ms/sample)	Training (ms/sample)
Non-MLLM	IPCE	151.3	8.1	3.2
	DPGF-Net	153.6	8.9	3.7
MLLM	DeQA-Score	393.8	65.0	482.5
	M3-AGIQA	72.2	370.4	817.1
	EAGLE	219.8	252.9	426.5

C.2 Evidence Sensitivity under Controlled Interventions

To further examine whether different assessment judgments exhibit distinct evidence dependencies, we conduct controlled intervention experiments on the AGIQA-3K test set. Specifically, we measure how EAGLE's predictions change when either the textual or visual input is perturbed while the other modality remains unchanged. For each criterion τ, sensitivity is quantified by the mean absolute change between the original and intervened predictions:

$$\Delta^{\tau} = \frac{1}{N} \sum_{i=1}^{N} \left| \hat{y}^{\tau}_{i,\mathrm{int}} - \hat{y}^{\tau}_{i,\mathrm{orig}} \right|, \tag{17}$$

where $\hat{y}^{\tau}_{i,\mathrm{orig}}$ and $\hat{y}^{\tau}_{i,\mathrm{int}}$ denote the predicted scores before and after intervention, respectively. A larger Δ^{τ} indicates greater sensitivity of criterion τ to the perturbed evidence source.

Textual Interventions. We first perturb the conditioning prompt while keeping the generated image unchanged. We consider three types of interventions. *Paraphrase* rewrites the original prompt while preserving its semantic content; *Prompt Swap* replaces it with an unrelated prompt; and *Null Prompt* removes the textual content entirely. These interventions allow us to distinguish sensitivity to surface-form variation from sensitivity to changes in cross-modal semantics.

Table 9: Sensitivity to textual interventions on AGIQA-3K, measured by absolute score change Δ.

Intervention	Correspondence $\lvert\Delta\rvert$	Quality $\lvert\Delta\rvert$
Paraphrase	0.106	0.000
Prompt Swap	1.095	0.125
Null Prompt	0.471	0.057

Semantic-preserving paraphrasing induces only a small change in correspondence and a negligible change in quality. In contrast, replacing the prompt with unrelated content produces a substantially larger change in correspondence, while quality remains relatively stable. This contrast suggests that the model is sensitive primarily to changes in text–image semantic consistency rather than to variations in prompt wording alone.

Visual Interventions. We then keep the conditioning prompt unchanged and perturb the generated image using Gaussian blur, Gaussian noise, and JPEG compression. These interventions alter the visual appearance without changing the textual input, allowing us to examine sensitivity to visual evidence.

Table 10: Sensitivity to visual interventions on AGIQA-3K, measured by absolute score change Δ.

Intervention	**Correspondence** $\lvert\Delta\rvert$	**Quality** $\lvert\Delta\rvert$
Gaussian Blur	0.282	0.669
Gaussian Noise	0.287	0.578
JPEG Compression	0.188	0.325

Visual degradations consistently induce larger changes in quality than in correspondence, indicating that perceptual quality is more sensitive to visual evidence. Together, the textual and visual interventions reveal complementary sensitivity patterns consistent with the proposed cross-modal and visual evidence anchors.