DECOUPLING PERCEPTION AND CALIBRATION: LABEL-EFFICIENT IMAGE QUALITY ASSESSMENT FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent multimodal large language models (MLLMs) have demonstrated strong capabilities in image quality assessment (IQA) tasks. However, adapting such large-scale models is computationally expensive and still relies on substantial Mean Opinion Score (MOS) annotations. We argue that for MLLM-based IQA, the core bottleneck lies not in the quality perception capacity of MLLMs, but in MOS scale calibration. Therefore, we propose LEAF, a ***L**abel-**E**fficient Image Quality **A**ssessment **F**ramework* that distills perceptual quality priors from an MLLM teacher into a lightweight student regressor, enabling MOS calibration with minimal human supervision. Specifically, the teacher conducts dense supervision through point-wise judgments and pair-wise preferences, with an estimate of decision reliability. Guided by these signals, the student learns the teacher's quality perception patterns through joint distillation and is calibrated on a small MOS subset to align with human annotations. Experiments on both user-generated and AI-generated IQA benchmarks demonstrate that our method significantly reduces the need for human annotations while maintaining strong MOS-aligned correlations, making lightweight IQA practical under limited annotation budgets. The code will be released upon publication.

1 INTRODUCTION

Image Quality Assessment (IQA) is a key component of modern visual systems, supporting a wide range of applications including camera calibration, image enhancement, content filtering, streaming media optimization, and large-scale data management (Fang et al., 2020; Seufert et al., 2014). IQA facilitates efficient benchmarking and optimization of image enhancement and generative models by providing feedback that is consistent with human standards (Ding et al., 2022; Li et al., 2025; Zhang et al., 2023; Wang et al., 2023).

Multimodal large language models (MLLMs) have recently demonstrated outstanding capabilities in IQA tasks, often matching specialized IQA models by leveraging strong visual understanding and rich prior knowledge (Zhang et al., 2025; Wang et al., 2024; Wu et al., 2024). This advancement has led to an emerging practice, treating MLLMs as high-performing quality evaluators and adapting them for IQA through instruction tuning or task-specific fine-tuning (Wang et al., 2025a; Chen et al., 2024). However, despite the considerable accuracy of these MLLM-based methods, they still face two practical barriers that limit their usability in real-world IQA pipelines, as shown in Figure 1.

First, MLLM-based methods are computationally expensive throughout the pipeline, both during training and at deployment (Wu et al., 2024). For example, Wang et al. (2025a) and Xu et al. (2025) proposed MLLM-based IQA methods, which require training and inference on large memory GPUs. This makes large MLLMs difficult to deploy in applications requiring fast, low-cost, and scalable quality prediction, such as on-device assessment, large-scale data filtering, and real-time monitoring (Wu et al., 2024).

Second, current IQA methods mainly rely on supervised learning with human mean opinion scores (MOS) (Li et al., 2025; Wang et al., 2024). Although MOS can provide reliable subjective judgments under controlled protocols, MOS annotation is resource-intensive. For example, the KonIQ-10K (Hosu et al., 2020) dataset contains more than $10,000$ images and has collected a total of 1.2

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.

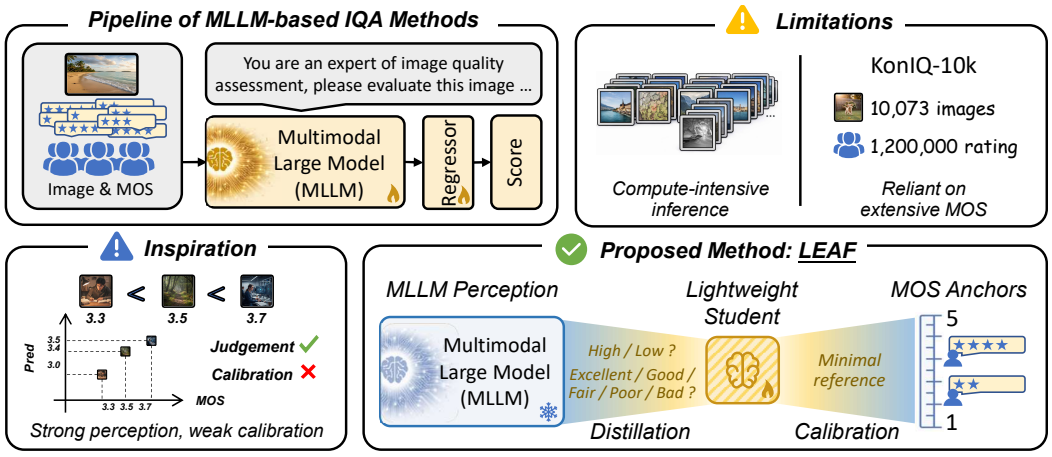

Figure 1: Using MLLMs for IQA is computationally expensive and annotation-intensive. MLLMs can perceive the key information of image quality. This work advocates decoupling quality perception from MOS calibration, thereby enabling efficient image quality assessment with minimal human supervision.

million opinion scores. Moreover, subjective quality perception drifts with the scenario. Therefore, MOS annotations need to be recollected as scenarios and time changes.

More importantly, for a pretrained MLLM, the primary limitation is often not the capability of quality perception, but the absence of explicit MOS scale calibration. Specifically, MLLMs perform well in capturing comparative and quality classification, such as distinguishing better versus worse images or assigning coarse quality levels. However, they struggle to map these perceptual judgments onto the MOS rating scales of specific datasets and specific scenarios.

Figure 2 illustrates this phenomenon on the AGIQA-3K dataset. Direct MLLM scoring preserves reliable ranking, but suffers from clear MOS scale miscalibration. Calibrating only a lightweight head with 10% MOS largely removes the residual bias and improves MOS alignment while keeping ranking performance (Fig. 2(b,c)). This suggests that the key challenge is MOS scale calibration rather than relearning perception. Consequently, directly fine-tuning MLLMs using large-scale MOS annotations is an inefficient strategy. Inspired by this observation, in this work, we address this challenge by proposing **LEAF**, a ***L**abel-**E**fficient Image Quality **A**ssessment **F**ramework* that transfers perceptual quality knowledge from an MLLM teacher into a lightweight student. The key idea is to use the teacher to provide dense supervision on a large unlabeled image set, and reserve human MOS labels only for final calibration.

Concretely, the teacher provides two complementary forms of supervision: (i) **point-wise** quality judgments for individual images, and (ii) **pair-wise** preferences that encode relative quality rankings, and also include estimates of decision reliability. Leveraging these signals, we formulate a Joint Teacher-Guided Distillation to enable the student to learn the teacher's quality perception. We then conduct Calibration Fine-Tuning on only a small MOS subset to align the student's scoring scale with human judgments. Our contributions are summarized as follows:

- We formulate a label-efficient IQA framework that decouples perceptual knowledge from MOS calibration, based on the insight that MLLMs are strong at quality perception but weak at dataset-specific MOS alignment.
- We propose a two-stage teacher-guided framework that transfers perceptual quality priors to a lightweight student by joint distillation of point-wise judgments and confidence-weighted pair-wise preferences, followed by calibration using a small MOS-labeled subset.
- Experiments on both UGC and AIGC benchmarks validate the proposed decoupling principle, showing that strong MOS-aligned correlation can be achieved with limited MOS supervision, while avoiding costly task-specific fine-tuning of MLLMs.

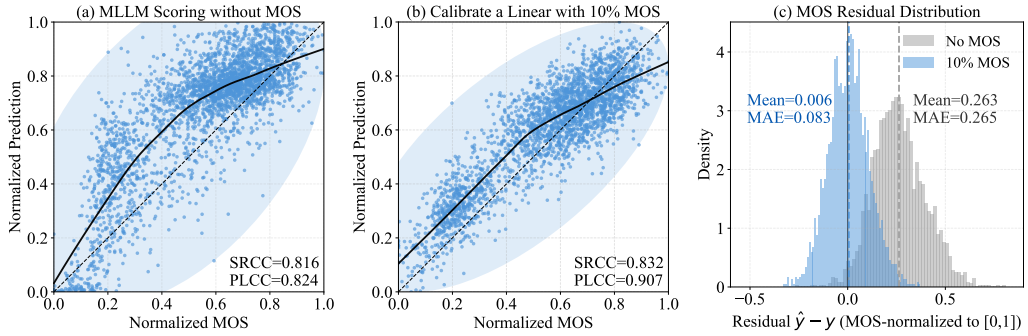

Figure 2: Direct MLLM scoring on AGIQA-3K exhibits strong monotonic correlation with MOS (SRCC=0.816) but suffers from pronounced scale bias and non-linear mapping (shown in (a)). Calibrating a Linear with only 10% MOS substantially improves MOS alignment (PLCC from 0.824 to 0.907) and reduces residual bias (mean residual from 0.263 to 0.006).

2 Related Work

2.1 Image Quality Assessment

2.1.1 Handcrafted IQA

Early IQA relied on handcrafted fidelity and perception priors, where SSIM-style structural comparison was widely adopted and extended for improved robustness (Wang et al., 2003), while phase- and gradient-based features offered complementary perceptual sensitivity (Zhang et al., 2011). No-reference IQA subsequently leveraged natural scene statistics, modeling distortions as deviations from natural regularities, as exemplified by DIIVINE (Moorthy & Bovik, 2011), BRISQUE (Mittal et al., 2012a), and NIQE (Mittal et al., 2012b).

2.1.2 Deep Learning–based IQA

Deep IQA learns quality-aware representations directly from data, with WaDIQaM (Bosse et al., 2016) as an early CNN-based blind model. Full-reference IQA similarly adopted deep feature distances, e.g., LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022), and PieAPP (Prashnani et al., 2018). Subsequent blind IQA methods enhanced adaptability via hypernetworks (Su et al., 2020), attention mechanisms (Chen et al., 2020), pseudo-reference hallucination (Lin & Wang, 2018), and deeper CNN interactions (Zhang et al., 2020), while transformers such as MUSIQ (Ke et al., 2021) and MANIQA (Yang et al., 2022) further enabled global multi-scale modeling.

2.1.3 Multi-modal Model-based IQA

Multimodal IQA leverages vision-language priors: LIQE (Zhang et al., 2023) fuses image and text embeddings, while CLIP-IQA (Wang et al., 2023) enables prompt-based zero-shot assessment. For AI-generated images, CLIP-based frameworks can explicitly model prompt consistency, and MLLMs further provide discrete judgments and continuous quality scores (Fu et al., 2024; Zhang et al., 2025). More recent AIGC-IQA works include MA-AGIQA (Wang et al., 2024), SEAGULL (Chen et al., 2024), EvoQuality (Wen et al., 2025), and the AGHI-QA benchmark (Li et al., 2025); however, most still depend on costly annotations, motivating label-free alternatives.

2.2 Label-Free IQA Methods

2.2.1 Classical NSS-based Model

Although supervised IQA achieves strong performance, collecting large-scale MOS labels is costly and difficult to scale to new domains and rapidly evolving AI-generated content. This motivates the emergence of label-free IQA, which estimates perceptual quality without explicit MOS supervision. Early methods follow the NIQE-style paradigm, they model pristine natural-scene statistics

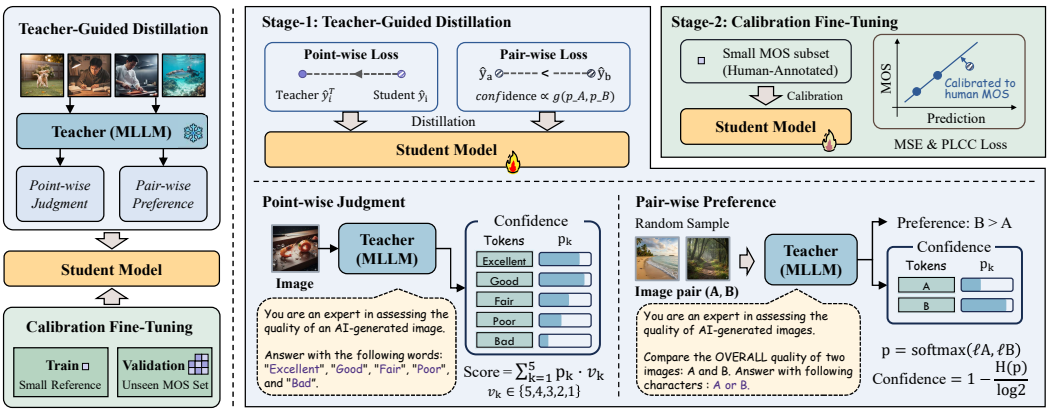

Figure 3: Overview of LEAF. Label-efficient IQA framework: Stage-1 jointly distills point-wise judgments and pair-wise preferences from an MLLM teacher into a lightweight student. Stage-2 calibrates the student to human MOS using a small annotated subset.

and quantify quality by the deviations from learned priors, e.g., NIQE (Mittal et al., 2012b), IL-NIQE (Zhang et al., 2015), and uBIQA (Xue et al., 2013). This idea has been further extended to domain-specific scenarios such as medical IQA (MSM-MIQ (Yuan et al., 2024)).

2.2.2 Deep Distribution Model

Beyond the hand-crafted NSS method, deep distribution modeling aims to better characterize high-quality image priors. MDFS (Ni et al., 2024) learns multi-scale deep feature distributions, while DSTS (Li et al., 2024b) separates shape- and texture-biased representations to capture various degradation scenarios. Generative and self-supervised variants learn quality-related latent spaces or features for deviation-based scoring, such as adversarial CVAE-based opinion-unaware perceptual IQA (Shukla et al., 2024) and NROUQA (Babu et al., 2023).

2.2.3 Pretraining and Quality Representation

QPT (Zhao et al., 2023) introduced quality-aware contrastive pretraining, ARNIQA (Agnolucci et al., 2024), and CONTRIQUE (Madhusudana et al., 2022) utilized synthetic degradation sequences and distortion-conditioned contrastive learning. Transformer-based designs further combine contrastive learning with quality regression (Shi et al., 2024), while Re-IQA (Saha et al., 2023) used a dual encoder to separate content and quality. DUBMA (Wang et al., 2025b) leverages a large distortion bank and FR-IQA evaluation metrics as pseudo-annotators for ranking supervision, and QualiCLIP (Agnolucci et al., 2025) aligned degraded images with quality-related antonymic prompts to learn multimodal embeddings while avoiding the use of MOS.

3 Method

3.1 Overview

To reduce the reliance on large-scale human MOS annotations while maintaining MOS-aligned IQA performance, we propose **LEAF**, a ***L**abel-**E**fficient Image Quality **A**ssessment **F**ramework* with two components: (i) *Teacher-Guided Distillation* that produces point-wise judgment and pair-wise preferences, and the student learns from dense teacher signals without using MOS in the distillation objective at stage 1, and (ii) *Calibration Fine-Tuning* that calibrates the student model using a small MOS subset at stage 2. An overview is shown in Figure 3.

Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ be a set of training images, and let only a small subset $\mathcal{D}_{\text{MOS}} = \{(x_i, y_i)\}_{i=1}^{M} \subset \mathcal{D}$ be annotated with human MOS $y_i \in \mathbb{R}$, where $M \ll N$. Our goal is to learn a lightweight student regressor $s_\theta : \mathcal{X} \to \mathbb{R}$ that predicts perceptual quality scores aligned with MOS.

3.1.1 TEACHER-GUIDED DISTILLATION

Given an image $x \in \mathcal{D}$, the MLLM teacher provides: (i) a point-wise judgment $\hat{y}^T(x)$, and (ii) a set of pair-wise supervision tuples $\mathcal{P} = \{(x_a, x_b, t_{ab}, \omega_{ab})\}$ with $(x_a, x_b) \in \mathcal{D} \times \mathcal{D}$. We construct $\mathcal{P}$ by uniformly sampling image pairs at random from $\mathcal{D} \times \mathcal{D}$. For each pair, we assign x_a as option A and x_b as option B in the teacher prompt. Here $t_{ab} \in \{0, 1\}$ indicates the teacher's preference, where $t_{ab} = 1$ if x_a is judged to have higher quality than x_b, and $\omega_{ab} \in [0, 1]$ measures the confidence of this decision.

During this stage, student s_θ is trained through dense teacher-induced supervision, without using MOS labels in the distillation objective. Concretely, the student is optimized to (i) match the teacher's point-wise judgment $\hat{y}^T(x)$ and (ii) preserve the pairwise ranking induced by the teacher, weighted by the teacher's confidence encoded in $\mathcal{P}$.

3.1.2 CALIBRATION FINE-TUNING

This stage uses the small MOS-labeled subset $\mathcal{D}_{\text{MOS}}$ to calibrate the student predictions to the MOS scale. We fine-tune the student on $\mathcal{D}_{\text{MOS}}$ with regression objective together with the correlation-based objective between predictions $\{s_\theta(x_i)\}$ and MOS labels $\{y_i\}$, so that the final scores are aligned with human perceptual judgments.

3.2 TEACHER-GUIDED DISTILLATION (TGD)

The teacher provides two complementary supervision signals: (1) *point-wise* quality judgments for individual images, and (2) *pair-wise* preferences with confidence for image pairs. We use fixed prompt templates to ensure constrained outputs. The complete prompts are provided in the appendix.

3.2.1 POINT-WISE JUDGMENT

For each image x, we prompt the teacher to judge its overall perceptual quality using a fixed set of discrete quality tokens $\mathcal{K} = \{Excellent, Good, Fair, Poor, Bad\}$. Instead of using the teacher's hard output, we extract token-level log-likelihoods $\{\ell_k(x)\}_{k \in \mathcal{K}}$ and convert them into a probability distribution:

$$p_k(x) = \frac{\exp(\ell_k(x))}{\sum_{j \in \mathcal{K}} \exp(\ell_j(x))}. \tag{1}$$

We then map the discrete tokens to ordinal scores $\{v_k\} = \{5, 4, 3, 2, 1\}$ (from $Excellent$ to Bad) and define a continuous teacher score by expectation:

$$\hat{y}^T(x) = \sum_{k \in \mathcal{K}} v_k \, p_k(x). \tag{2}$$

This produces a dense supervision signal for all images in $\mathcal{D}$.

3.2.2 PAIR-WISE PREFERENCE

To capture relative quality, we randomly sample pairs (x_a, x_b) and prompt the teacher to judge which image has higher quality, outputting a single token in $\{\texttt{A}, \texttt{B}\}$.

We obtain the log-likelihoods (ℓ_A, ℓ_B) at the decision position and compute:

$$p_A = \frac{e^{\ell_A}}{e^{\ell_A} + e^{\ell_B}}, \quad p_B = \frac{e^{\ell_B}}{e^{\ell_A} + e^{\ell_B}}. \tag{3}$$

The hard preference label is

$$t_{ab} = \begin{cases} 1, & \text{if teacher prefers } A \ (p_A \geq p_B), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Not all teacher pair-wise decisions are equally reliable. We quantify confidence through the normalized entropy of the two-class distribution:

$$H(p) = -\sum_{c \in \{A,B\}} p_c \log p_c, \qquad \omega_{ab} = 1 - \frac{H(p)}{\log 2} \in [0, 1], \tag{5}$$

where ω_{ab} is larger when teacher's decision is more certain. In practice, we discard low-confidence pairs with $\omega_{ab} < \tau$.

3.2.3 DISTILLATION OBJECTIVE

We distill quality prior from the teacher using dense supervision on $\mathcal{D}$. Specifically, we leverage both (i) pointwise soft scores $\hat{y}^T(x)$ and (ii) pairwise preferences $\mathcal{P}$ to train a lightweight student regressor.

Student predictor. We denote the student as an IQA regressor $s_\theta : \mathcal{X} \to \mathbb{R}$ that outputs a scalar quality score. Our formulation does not rely on a specific architecture, requiring only real-valued outputs suitable for regression and ranking.

Point-wise distillation. Given teacher-provided soft scores $\hat{y}^T(x)$ for $x \sim \mathcal{D}$, we minimize

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{x\sim\mathcal{D}} \left[\text{SmoothL1} \left(s_\theta(x), \hat{y}^T(x) \right) \right]. \tag{6}$$

Pair-wise distillation. Given teacher-provided pairs $\mathcal{P} = \{(x_a, x_b, t_{ab}, \omega_{ab})\}$, we define

$$P_\theta(a \succ b) = \sigma\big(s_\theta(x_a) - s_\theta(x_b)\big), \qquad \sigma(z) = \frac{1}{1+e^{-z}}. \tag{7}$$

We optimize a weighted binary cross-entropy loss:

$$\mathcal{L}_{\text{rank}} = \mathbb{E}_{(x_a,x_b,t_{ab},\omega_{ab})\sim\mathcal{P}} \left[\omega_{ab} \cdot \text{BCE} \left(P_\theta(a \succ b),\, t_{ab} \right) \right]. \tag{8}$$

Joint objective. The objective of this stage is

$$\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{reg}} + \lambda_{dis} \mathcal{L}_{\text{rank}}, \tag{9}$$

where λ_{dis} is used to balance point-wise and pair-wise distillation. We select the checkpoint without using any MOS annotations in the Mos free setting. In the few-shot setting, checkpoint selection is performed using a held-out split of the same MOS subset $\mathcal{D}_{\text{MOS}}$ that is visible for calibration fine-tuning stage, without introducing any additional labeled data or test-set leakage.

3.3 CALIBRATION FINE-TUNING (CFT)

To align the student predictions with the human MOS scale, we fine-tune the student on the small labeled subset $\mathcal{D}_{\text{MOS}}$ with a regression and a correlation objective.

Let $\mathbf{s} = [s_\theta(x_i)]_{(x_i,y_i)\in\mathcal{D}_{\text{MOS}}}$ and $\mathbf{y} = [y_i]$ denote vectors of predictions and MOS. We first perform scale calibration using mean squared error:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_{\text{MOS}}} \left[\big(s_\theta(x_i) - y_i\big)^2 \right]. \tag{10}$$

Meanwhile, we optimize the PLCC calculated within each mini-batch for differentiability by minimizing

$$\mathcal{L}_{\text{PLCC}} = 1 - \rho(\mathbf{s}, \mathbf{y}), \qquad \rho(\mathbf{s}, \mathbf{y}) = \frac{\text{cov}(\mathbf{s}, \mathbf{y})}{\sigma(\mathbf{s})\,\sigma(\mathbf{y})}. \tag{11}$$

The overall objective of the calibration fine-tuning stage is

$$\mathcal{L}_{\text{cal}} = \mathcal{L}_{\text{MSE}} + \lambda_{cal} \mathcal{L}_{\text{PLCC}}, \tag{12}$$

where λ_{cal} balances absolute MOS calibration and correlation optimization.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

The teacher model is based on the InternVL-3.5-8B and remains frozen during all stages. We use a ConvNeXt-Base as the student model pretrained on ImageNet. All images are resized to 256

Table 1: Comparison with state-of-the-art IQA methods on AIGC and UGC benchmarks.

Methods	AIGC Datasets				UGC Datasets			
	AGIQA-3K		AIGIQA-20K		KonIQ-10k		SPAQ	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Supervised								
BRISQUE (Mittal et al., 2012a) (TIP, 2012)	0.472	0.561	0.466	0.558	0.705	0.707	0.802	0.805
HyperIQA (Su et al., 2020) (CVPR, 2020)	0.850	0.904	0.816	0.832	0.904	0.915	0.915	0.918
MANIQA (Yang et al., 2022) (CVPR, 2022)	0.861	0.911	0.850	0.887	0.930	0.946	0.922	0.927
DBCNN (Zhang et al., 2020) (TCSVT, 2020)	0.826	0.890	0.805	0.848	0.844	0.862	0.909	0.927
MUSIQ (Ke et al., 2021) (ICCV, 2021)	0.820	0.865	0.832	0.864	0.824	0.937	0.873	0.868
StairIQA (Sun et al., 2023) (JSTSP, 2023)	0.834	0.893	0.789	0.842	0.920	0.936	0.923	0.929
Q-Align (Wu et al., 2023) (TOMM, 2023)	0.852	0.881	**0.874**	0.889	0.922	0.911	0.887	0.886
MA-AGIQA (Wang et al., 2024) (ACMMM, 2024)	**0.893**	**0.927**	0.864	**0.905**	**0.933**	**0.948**	**0.927**	**0.932**
Weak-supervised								
CONTRIQUE (Madhusudana et al., 2022) (TIP, 2022)	0.817	0.879	0.788	0.807	0.894	0.906	0.914	0.919
Re-IQA (Saha et al., 2023) (CVPR, 2023)	0.811	0.874	0.787	0.811	**0.914**	**0.923**	0.918	**0.925**
CLIP-IQA+ (Wang et al., 2023) (AAAI, 2023)	0.844	0.894	0.833	0.854	0.895	0.909	0.864	0.866
ARNIQA (Agnolucci et al., 2024) (WACV, 2024)	0.803	0.881	0.778	0.792	0.869	0.883	0.904	0.909
GRepQ-D (Srinath et al., 2024) (WACV, 2024)	0.807	0.858	0.789	0.810	0.855	0.868	0.903	0.917
Ours (10% MOS)	0.841	0.899	0.839	0.878	0.867	0.903	0.896	0.902
Ours (30% MOS)	**0.868**	**0.914**	**0.860**	**0.905**	0.899	0.916	**0.921**	0.922
Label-free								
NIQE (Mittal et al., 2012b) (ISPL, 2012)	0.523	0.566	0.208	0.337	0.551	0.488	0.703	0.670
ILNIQE (Zhang et al., 2015) (TIP, 2015)	0.609	0.655	0.335	0.455	0.453	0.467	0.719	0.654
Re-IQA (Saha et al., 2023) (CVPR, 2023)	0.627	0.633	0.611	0.626	0.580	0.568	0.613	0.616
CLIP-IQA (Wang et al., 2023) (AAAI, 2023)	0.638	0.711	0.388	0.537	0.695	0.727	0.738	0.735
MDFS (Ni et al., 2024) (TMM, 2024)	0.672	0.676	0.691	0.695	0.733	0.737	0.741	0.754
GRepQ-Z (Srinath et al., 2024) (WACV, 2024)	0.613	0.734	0.624	0.634	0.768	0.784	0.823	0.839
DUBMA (Wang et al., 2025b) (IJCAI, 2025)	0.684	0.701	0.695	0.697	0.703	0.740	0.834	0.841
Ours	**0.749**	**0.811**	**0.696**	**0.762**	**0.777**	**0.801**	**0.861**	**0.867**

and cropped to 224×224. Random resized cropping and horizontal flipping are applied during training, while center cropping is used for evaluation. We use AdamW for model optimization and adopt mixed precision training. Stage 1 is trained jointly with the objective by using point-wise judgments and pairwise preferences provided by the teacher, where $\lambda_{dis} = 0.5$. Stage-2 fine-tunes the student on the labeled subset using a correlation-based objective to align the prediction results with human MOS. Hyperparameters and implementation details, including pair construction and filtering strategies, are provided in the appendix.

4.2 Datasets and Evaluation Protocol

We evaluate on the image quality assessment benchmarks for both user-generated content (UGC) and AI-generated content (AIGC). For UGC, we use KonIQ-10k (Hosu et al., 2020) and SPAQ (Fang et al., 2020), which contain natural images with authentic distortions from real-world capture and processing pipelines. For AIGC, we use AGIQA-3K (Li et al., 2023), AIGIQA-20K (Li et al., 2024a), which cover images generated by diverse text-to-image models with varying perceptual qualities.

To quantify label efficiency, we adjusted the MOS-visible ratio on the training split, resulting in three evaluation settings that correspond to Table 1: (i) label-free: the student is trained solely with teacher supervision, without using any human MOS. (ii) 10% MOS: only a randomly sampled 10% subset of training images contain MOS labels, while the remaining training samples are treated as MOS-invisible. (iii) 30% MOS: similarly, 30% of the training images contain MOS labels.

We repeat the sampling process five times with different random seeds and report the average SRCC/PLCC values across the five runs. Across all settings, we construct the pair-wise supervision set by uniformly sampling image pairs at random, with the number of sampled pairs matching the size of the dataset. The ablation experiment on the number of samples is provided in the appendix.

4.3 Comparison with State-of-the-Art Methods

We compare our method with representative weak-supervised and label-free IQA approaches on both AIGC and UGC benchmarks, as shown in Table 1. For AIGC, we report results on AGIQA-

3K and AIGIQA-20K. For UGC, we evaluate the performance on KonIQ-10k and SPAQ. In each case, we consider both label-free and weak-supervised settings to validate the effectiveness of the proposed method.

4.3.1 AIGC BENCHMARKS

Without using any human MOS, our method shows strong correlation with human judgments on AGIQA-3K, with an SRCC of 0.749 and a PLCC of 0.811, outperforming the strongest label-free baseline in the table (SRCC = 0.684). On AIGIQA-20K, it reaches 0.696 and 0.762 in terms of SRCC and PLCC, respectively, achieving the best SRCC among label-free competitors while exceeding the best label-free PLCC 0.697. These results indicate that teacher-driven dense supervision can transfer reliable quality perception even without human labels.

With ten percent MOS, our method achieves an SRCC of 0.841 and a PLCC of 0.899 on AGIQA-3K, and 0.839 and 0.878 on AIGIQA-20K, respectively. On AGIQA-3K, the PLCC is the highest among weak-supervised methods, while the SRCC remains close to the strongest competitor. For reference, most weak-supervised IQA methods in the table are evaluated using a large portion of MOS labels (often around 70% (Saha et al., 2023; Agnolucci et al., 2024; Madhusudana et al., 2022)) to calibrate the regression head. In contrast, we report results with 30% MOS in our framework, with the performance on AGIQA-3K further improving to 0.868 and 0.914, and to 0.860 and 0.905 on AIGIQA-20K in terms of SRCC and PLCC, respectively. Notably, the PLCC on AIGIQA-20K matches the best supervised result reported in the table.

4.3.2 UGC BENCHMARKS

For label-free setting, our method achieves strong correlations on both UGC benchmarks. On KonIQ-10k, it attains an SRCC of 0.777 and a PLCC of 0.801, while on SPAQ, it reaches 0.861 and 0.867 of SRCC and PLCC, respectively. On both datasets, our method achieved the best results among label-free competitors. In particular, on SPAQ, it improved SRCC from 0.834 to 0.861 compared to the strongest competing baseline, and in terms of PLCC, it improved from 0.841 to 0.867.

With 10% MOS, our method achieves 0.867 and 0.903 on KonIQ-10k, and 0.896 and 0.902 on SPAQ in terms of SRCC and PLCC. For reference, the compared weak-supervised methods commonly rely on around 70% of MOS labels for regression-based calibration. Using only thirty percent MOS, our method further improves performance to 0.899 and 0.916 on KonIQ-10k, and to 0.921 and 0.922 on SPAQ, achieving competitive results with fewer MOS among weak-supervised methods reported in Table 1. These results show that limited MOS supervision already yields strong performance, while additional MOS labels consistently improve the correlation.

4.4 EFFICIENCY ANALYSIS

Figure 4 illustrates label efficiency on AGIQA-3K by varying the MOS availability ratio in Calibration Fine-Tuning (CFT) stage while keeping Teacher-Guided Distillation (TGD) stage fixed. Overall, performance improves sharply with a small amount of MOS and then gradually saturates. Without any MOS, the model achieves an SRCC of 0.749 and a PLCC of 0.811. When the MOS ratio reaches 10%, performance increases to 0.841 and 0.899 in terms of SRCC and PLCC, respectively. As the MOS ratio further increases, the gains become marginal, reaching 0.868 and 0.914 at 30% MOS, indicating that CFT primarily serves to calibrate the dataset-specific score scale with limited MOS annotations. More efficiency analysis is shown in the appendix.

4.5 ABLATION STUDIES

Table 2 presents different supervision strategies used in TGD on AGIQA-3K with 10% visible MOS, including point-wise judgments, pair-wise preferences, and confidence-aware weighting on pairs.

The first row reports the CFT-only baseline, where Stage-1 distillation is removed, resulting in 0.759 and 0.812. Using only point-wise supervision to distill the teacher significantly improves the student to 0.834 and 0.892, indicating that point-wise soft judgments provide strong quality guidance. Adding pair-wise supervision on top of point-wise supervision boosts performance to 0.836 and

Table 2: Ablation study of different supervision strategies in Teacher-Guided Distillation.

Point	Pair	Pair Conf.	SRCC	PLCC
✗	✗	✗	0.759	0.812
✓	✗	✗	0.834	0.892
✓	✓	✗	0.836	0.895
✗	✓	✓	0.829	0.889
✓	✓	✓	**0.841**	**0.899**

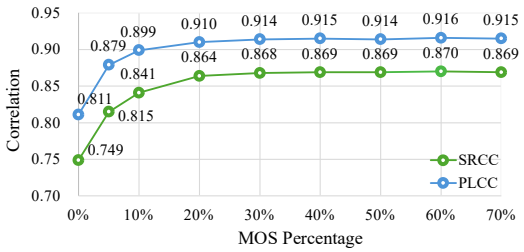

Figure 4: Impact of MOS availability during MOS calibration on AGIQA-3K.

Table 3: Robustness to teacher models and student backbone architectures.

Setting	Params	SRCC	PLCC
MLLMs			
Qwen3-VL	8B	0.835	0.873
Qwen3-VL	32B	0.848	0.907
InternVL-3.5	8B	0.841	0.899
InternVL-3.5	38B	0.850	0.912
Student Backbones			
ResNet-18	11.7M	0.807	0.868
ResNet-50	25.6M	0.815	0.886
ConvNeXt-Tiny	28.6M	0.833	0.896
ConvNeXt-Base	88.6M	0.841	0.899
ConvNeXt-Large	197M	0.836	0.882
ConvNeXt-XL	350M	0.843	0.900
ViT-B	86M	0.840	0.900
Swin-B	88M	0.842	0.901

0.895, showing that relative comparisons provide complementary ranking information. When relying solely on pair-wise supervision with confidence weighting, the student reaches 0.829 and 0.889. This setting still brings a noticeable performance gain, showing that pair-wise preferences can transfer useful relative quality information.

4.6 Robustness to Different Backbones

Table 3 studies the robustness of our framework to different teacher and student choices on AGIQA-3K with 10% visible MOS. Overall, stronger teachers can provide higher-quality supervision signals, and the student achieves consistent performance across various backbone network capacities.

For Qwen3-VL, SRCC and PLCC increased from 0.835 and 0.873 with the 8B model to 0.848 and 0.907 with the 32B model. Meanwhile, InternVL-3.5 provided a competitive and generally stronger supervision, reaching 0.841 and 0.899 with the 8B model and achieving the best performance of 0.850 and 0.912 with the 38B model. For student backbones, performance improves from ResNet-18 to ConvNeXt variants. Among ConvNeXt variants, ConvNeXt-XL achieves the highest SRCC/PLCC, while ConvNeXt-Base provides a favorable trade-off between performance and model size.

5 Conclusion

In this work, we revisit the role of multimodal large language models in image quality assessment and argue that the main challenge lies not in perceptual understanding, but in aligning model outputs with dataset-specific MOS scales. Based on this insight, we propose LEAF, a label-efficient IQA framework that decouples perceptual knowledge transfer from MOS calibration. The MLLM teacher provides point-wise judgments and pair-wise preferences, while the student model learns robust quality perception without relying on human annotations in the distillation stage. Finally, a lightweight calibration stage with a small MOS subset is sufficient to align predictions with human annotations.

Extensive experiments on both user-generated and AI-generated IQA benchmarks demonstrate that LEAF significantly reduces the dependence on annotations while achieving competitive or superior correlation with human judgments. This work provides a practical alternative to direct regression with heavy annotation requirements. We believe the proposed framework offers a promising direction for future research on label-efficient quality assessment tasks.

AI USE STATEMENT

In this work, we used generative AI tools for writing auxiliary scripts for data preprocessing, grammar checking, and language polishing. We have not used generative AI tools for generating core research ideas, synthesizing experimental data, or generating main figures, and mathematical proofs and theoretical derivations are not applicable to this work. We have reviewed all AI-assisted work. Specifically, all LLM-generated code was manually reviewed, verified for correctness, and tested by two authors before integration. All text edits and literature summaries suggested by AI were carefully checked by the authors to ensure scientific accuracy and to prevent any unintended plagiarism. No AI tool was used to generate, manipulate, or interpolate empirical results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate full reproducibility, we provide comprehensive implementation details in the appendix and supplementary material. The complete configuration and training protocol of the student model, including the backbone architecture, preprocessing pipeline, and the two-stage training procedure (Stage 1 pairwise distillation and Stage 2 MOS calibration), are documented in Appendix D. The design of teacher supervision signals, including prompt templates, point-wise soft score construction, pair-wise preference extraction, and confidence weighting, is detailed in Appendix C. The dataset descriptions and evaluation protocols for both UGC and AIGC benchmarks are provided in Appendix F. Additional ablation studies, including the impact of the number of sampled pairs in Stage 1 and the computational efficiency of LEAF, are reported in Appendix E. The full code will be released upon publication. We confirm that the main results reported in the paper can be reproduced by following the configurations and protocols described in the above sections.

REFERENCES

Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. Arniqa: Learning distortion manifold for image quality assessment. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 188–197, Jan 2024. doi: 10.1109/WACV57701.2024.00026.

Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini. Quality-aware image-text alignment for opinion-unaware image quality assessment, 2025. URL https://arxiv.org/abs/2403.11176.

Nithin C Babu, Vignesh Kannan, and Rajiv Soundararajan. No reference opinion unaware quality assessment of authentically distorted images. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2458–2467, Jan 2023. doi: 10.1109/WACV56688.2023.00249.

Sebastian Bosse, Dominique Maniry, Thomas Wiegand, and Wojciech Samek. A deep neural network for image quality assessment. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3773–3777, Sep. 2016. doi: 10.1109/ICIP.2016.7533065.

Diqi Chen, Yizhou Wang, and Wen Gao. No-reference image quality assessment: An attention driven approach. *IEEE Transactions on Image Processing*, 29:6496–6506, 2020. ISSN 1941-0042. doi: 10.1109/TIP.2020.2990342.

Zewen Chen, Juan Wang, Wen Wang, Sunhan Xu, Hang Xiong, Yun Zeng, Jian Guo, Shuxun Wang, Chunfeng Yuan, Bing Li, and Weiming Hu. Seagull: No-reference image quality assessment for regions of interest via vision-language instruction tuning. *CoRR*, abs/2411.10161, 2024. URL https://doi.org/10.48550/arXiv.2411.10161.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, May 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2020.3045810.

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)*, pp. 3674–3683. IEEE, 2020.

Jun Fu, Wei Zhou, Qiuping Jiang, Hantao Liu, and Guangtao Zhai. Vision-language consistency guided multi-modal prompt learning for blind ai generated image quality assessment. *IEEE Signal Processing Letters*, 31:1820–1824, 2024. ISSN 1558-2361. doi: 10.1109/LSP.2024.3420083.

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. ISSN 1941-0042. doi: 10.1109/tip.2020.2967829. URL http://dx.doi.org/10.1109/TIP.2020.2967829.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137, 2021. doi: 10.1109/ICCV48922.2021.00510.

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*, pp. 1–1, 2023. doi: 10.1109/TCSVT.2023.3319020.

Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Aigiqa-20k: A large database for ai-generated image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 6327–6336, June 2024a.

Yixuan Li, Peilin Chen, Hanwei Zhu, Keyan Ding, Leida Li, and Shiqi Wang. Deep shape-texture statistics for completely blind image quality evaluation. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(12), November 2024b. ISSN 1551-6857. doi: 10.1145/3694977. URL https://doi.org/10.1145/3694977.

Yunhao Li, Sijing Wu, Wei Sun, Zhichao Zhang, Yucheng Zhu, Zicheng Zhang, Huiyu Duan, Xiongkuo Min, and Guangtao Zhai. Aghi-qa: A subjective-aligned dataset and metric for ai-generated human images, 2025. URL https://arxiv.org/abs/2504.21308.

Kwan-Yee Lin and Guanxiang Wang. Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 732–741, June 2018. doi: 10.1109/CVPR.2018.00083.

Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022. ISSN 1941-0042. doi: 10.1109/TIP.2022.3181496.

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, 2012a.

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

Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, Dec 2011. ISSN 1941-0042. doi: 10.1109/TIP.2011.2147325.

Zhangkai Ni, Yue Liu, Keyan Ding, Wenhan Yang, Hanli Wang, and Shiqi Wang. Opinion-unaware blind image quality assessment using multi-scale deep feature statistics. *IEEE Transactions on Multimedia*, 26:10211–10224, 2024. ISSN 1941-0077. doi: 10.1109/TMM.2024.3405729.

Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1808–1817, June 2018. doi: 10.1109/CVPR.2018.00194.

Avinab Saha, Sandeep Mishra, and Alan C. Bovik. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5846–5855, June 2023. doi: 10.1109/CVPR52729.2023.00566.

Michael Seufert, Sebastian Egger, Martin Slanina, Thomas Zinner, Tobias Hoßfeld, and Phuoc Tran-Gia. A survey on quality of experience of http adaptive streaming. *IEEE Communications Surveys & Tutorials*, 17(1):469–492, 2014.

Jinsong Shi, Pan Gao, and Jie Qin. Transformer-based no-reference image quality assessment via supervised contrastive learning. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i5.28285. URL https://doi.org/10.1609/aaai.v38i5.28285.

Ankit Shukla, Avinash Upadhyay, Swati Bhugra, and Manoj Sharma. Opinion unaware image quality assessment via adversarial convolutional variational autoencoder. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2142–2152, Jan 2024. doi: 10.1109/WACV57701.2024.00215.

Suhas Srinath, Shankhanil Mitra, Shika Rao, and Rajiv Soundararajan. Learning generalizable perceptual representations for data-efficient no-reference image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 22–31, January 2024.

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)*, pp. 3664–3673, June 2020. doi: 10.1109/CVPR42600.2020.00372.

Wei Sun, Xiongkuo Min, Danyang Tu, Siwei Ma, and Guangtao Zhai. Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. *IEEE Journal of Selected Topics in Signal Processing*, 2023.

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. doi: 10.1609/aaai.v37i2.25353.

Jiarui Wang, Huiyu Duan, Yu Zhao, Juntong Wang, Guangtao Zhai, and Xiongkuo Min. Lmm4lmm: Benchmarking and evaluating large-multimodal image generation with lmms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17312–17323, October 2025a.

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, pp. 7803–7812, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681471. URL https://doi.org/10.1145/3664647.3681471.

Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402 Vol.2, 2003. doi: 10.1109/ACSSC.2003.1292216.

Zhihua Wang, Xuelin Liu, Jiebin Yan, Jie Wen, Wei Wang, and Chao Huang. Deep opinion-unaware blind image quality assessment by learning and adapting from multiple annotators. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 2036–2044, 2025b.

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. *arXiv preprint arXiv:2509.25787*, 2025.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang. A comprehensive study of multimodal large language models for image quality assessment, 2024. URL https://arxiv.org/abs/2403.10854.

Zitong Xu, Huiyu Duan, Bingnan Liu, Guangji Ma, Jiarui Wang, Liu Yang, Shiqi Gao, Xiaoyu Wang, Jia Wang, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Lmm4edit: Benchmarking and evaluating multimodal image editing with lmms. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 6908–6917, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755294. URL https://doi.org/10.1145/3746027.3755294.

Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 995–1002, June 2013. doi: 10.1109/CVPR.2013.133.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1190–1199, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/CVPRW56347.2022.00126. URL https://doi.ieeecomputersociety.org/10.1109/CVPRW56347.2022.00126.

Shuaiyu Yuan, Tristan Whitmarsh, Dimitri A Kessler, Otso Arponen, Mary A McLean, Gabrielle Baxter, Frank Riemer, Aneurin J Kennerley, William J Brackenbury, Fiona J Gilbert, and Joshua D Kaggie. A deep-learning-based label-free no-reference image quality assessment metric: Application in sodium mri denoising, 2024. URL https://arxiv.org/abs/2408.16481.

Lin Zhang, Lei Zhang, Xuanqin Mou, and D. Zhang. Fsim: A feature similarity index for image quality assessment. *Trans. Img. Proc.*, 20(8):2378–2386, August 2011. ISSN 1057-7149. doi: 10.1109/TIP.2011.2109730. URL https://doi.org/10.1109/TIP.2011.2109730.

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, Aug 2015. ISSN 1941-0042. doi: 10.1109/TIP.2015.2426416.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric . In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, Los Alamitos, CA, USA, June 2018. IEEE Computer Society. doi: 10.1109/CVPR.2018.00068. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00068.

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, Jan 2020. ISSN 1558-2205. doi: 10.1109/TCSVT.2018.2886771.

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 (CVPR)*, pp. 14071–14081, June 2023.

Zhichao Zhang, Xinyue Li, Wei Sun, Zicheng Zhang, Yunhao Li, Xiaohong Liu, and Guangtao Zhai. Leveraging multimodal large language models for joint discrete and continuous evaluation in text-to-image alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 977–986, June 2025.

Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. Quality-aware pretrained models for blind image quality assessment. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22302–22313, June 2023. doi: 10.1109/CVPR52729.2023.02136.

A OVERVIEW

This appendix provides more details to support the methods, analyses, and experimental protocols in the main paper. The specific content includes:

a) Details for the motivating analysis in Figure 2, including the MLLM point-wise scoring process, the 10% MOS calibration setup with a frozen MLLM, the residual computation, and the analysis for MLLM's struggle with scale mapping (Appendix B).
b) More details of the teacher supervision signals, including prompt design, point-wise soft score construction, pair-wise preference extraction, and confidence weighting (Appendix C).
c) Configuration and training protocol of the student model, including the backbone, preprocessing, and the two-stage training procedure (Stage 1 distillation and Stage 2 MOS calibration) (Appendix D).
d) Additional experimental details, including the impact of the number of sampled pairs used in Stage 1 pairwise distillation and the efficiency of LEAF(Appendix E).
e) Dataset descriptions and evaluation protocol for both UGC and AIGC benchmarks (Appendix F).
f) Limitation (Appendix G).

B DETAILS FOR FIGURE 2

This section elaborates the motivating analysis in Figure 2. Importantly, the MLLM scoring process in Figure 2(a) is identical to our teacher point-wise supervision described in Section C, i.e., it uses the same prompt, the same five quality tokens, and the same soft score construction from token log-likelihoods.

Dataset, normalization, and evaluation. All results are reported on AGIQA-3K according to the official split. For visualization in Figure 2, MOS values are min-max normalized to $[0, 1]$, and model prediction values are also min-max normalized to $[0, 1]$ using the corresponding split statistics. SRCC/PLCC are computed between normalized predictions and normalized MOS on the evaluation split.

(a) Direct MLLM scoring without MOS. For Figure 2(a), we use the pretrained MLLM with greedy decoding and record token-level log-probabilities at the first generated position. Using the fixed token set $\mathcal{C} = \{\texttt{Excellent}, \texttt{Good}, \texttt{Fair}, \texttt{Poor}, \texttt{Bad}\}$, we compute a soft distribution p_c over $\mathcal{C}$ and form a continuous score $s_{\text{teacher}} \in [1, 5]$ by the expectation rule in Section C. Human MOS are not used in this process. This results in the scatter plot in Figure 2(a), where the ranking correlation can be strong, but the absolute scale shows noticeable miscalibration.

(b) Calibrating with 10% MOS using a lightweight head. For Figure 2(b), we keep the pretrained MLLM fully frozen and perform MOS calibration using only a randomly sampled 10% MOS subset from the training split (fixed seed). Specifically, we attach a lightweight regression head (a single fully-connected layer) after the last Transformer encoder and train only this head to predict MOS. The parameters of the MLLM backbone network will not be updated. Then, the calibrated predictor will be evaluated on the same evaluation set as (a). This experiment shows that a small amount of MOS supervision can significantly improve MOS alignment while leaving the pretrained MLLM intact.

(c) Residual distribution (gray vs. blue). Figure 2(c) shows the distribution of residuals $\hat{y} - y$, where $\hat{y}$ represents the normalized prediction and y denotes the normalized MOS (both in $[0, 1]$). The gray histogram corresponds to direct MLLM point-wise scoring in (a), while the blue histogram corresponds to the calibrated (frozen-MLLM + lightweight head) predictor in (b). We report the mean residual to quantify systematic bias and the MAE to quantify overall calibration error. As shown, the 10% MOS calibration significantly reduces residual bias and improves MOS-aligned accuracy.

The Analysis for MLLM's struggle with scale mapping. MOS prediction requires dataset-specific scale calibration, which is typically not learned during MLLM pretraining, as reflected in Figure 2. Moreover, our experiments show that autoregressive MLLMs tend to favor a few discrete values (e.g., 2.0, 2.5, 3.0, 3.5) rather than truly continuous scores.

C Teacher Signals

C.1 Teacher Model and Inference Setup

This subsection describes the teacher model and inference configuration used to provide supervision for student training. We use InternVL3.5-8B as the multimodal teacher. To keep the teacher outputs consistent across repeated runs, we disable sampling during inference process and use greedy decoding (with temperature set to zero). We also record token-level log-probabilities, keeping the top 50 candidate tokens at each generated position. These log-probabilities are used to construct (i) soft point-wise quality scores and (ii) probabilistic pair-wise preferences. For point-wise scoring, the teacher is constrained to generate a single token (one-word category). For pair-wise comparison, we allow up to eight tokens to ensure the model can produce a clean A/B decision while still keeping the output concise.

C.2 Point-wise Prompt and Soft Score Construction

This subsection describes how the point-wise prompt generate a five-level judgment and how we convert it into a continuous soft score.

Prompt.

> `<image>`
> *You are an expert image quality evaluator. The degree of image quality should be categorized into five levels, from high to low: "Excellent", "Good", "Fair", "Poor", and "Bad". Answer with ONLY ONE of the following words: "Excellent", "Good", "Fair", "Poor", or "Bad".*

Intuitively, one may consider a simpler design that directly prompts the teacher to output a continuous quality score (e.g., a real number in $[1, 5]$). However, in practical applications, we find that this direct numeric scoring tends to collapse to a small set of preferred values rather than producing truly continuous outputs. For example, the teacher frequently emits scores on a coarse grid with fixed spacing (often around 0.25), behaving like a slightly finer discrete classification rather than genuine regression. This quantization effect makes it difficult to fully utilize the teacher's perceptual sensitivity, as many visually distinct samples are forced into the same narrow score range.

A reasonable explanation is that an autoregressive LLM represents a numeric rating as a discrete string and assigns probability mass according to both task evidence and strong format priors learned from text corpora. Without an explicit calibration objective, the model tends to reuse a small set of high-priority "rating anchors" (e.g., half- or quarter-point increments) that are prevalent in natural language reviews, yielding grid-like and highly concentrated outputs. Moreover, the likelihood landscape over numeric strings is typically not smooth with respect to perceptual quality, so visually distinct samples may map to the same stable numeric outputs. Therefore, instead of relying on free-form numeric generation, we restrict the teacher to a fixed set of quality tokens and recover a continuous score from token-level log-likelihoods, which provides a stable and uncertainty-aware supervision signal.

Probability extraction. We read the teacher's top-50 token log-probabilities at the first generated position and form a distribution over the five candidate labels

$$\mathcal{C} = \{\texttt{Excellent}, \texttt{Good}, \texttt{Fair}, \texttt{Poor}, \texttt{Bad}\}.$$

For each candidate $c \in \mathcal{C}$, we take its log-probability ℓ_c if it appears in the top-50 list; otherwise we assign a very small value so that it receives negligible probability mass.

Soft score. We compute a stable softmax over $\{\ell_c\}$:

$$p_c = \frac{\exp(\ell_c - m)}{\sum_{c' \in \mathcal{C}} \exp(\ell_{c'} - m)}, \quad m = \max_{c \in \mathcal{C}} \ell_c,$$

and form a continuous score by a weighted average:

$$s_{\text{teacher}} = \sum_{c \in \mathcal{C}} p_c \cdot r(c),$$

where $r(\texttt{Excellent}) = 5$, $r(\texttt{Good}) = 4$, $r(\texttt{Fair}) = 3$, $r(\texttt{Poor}) = 2$, $r(\texttt{Bad}) = 1$. This yields $s_{\text{teacher}} \in [1, 5]$ and preserves the teacher's uncertainty when multiple categories are plausible.

Outputs. For each image, we store the teacher soft score and the image identifier. If MOS is available in the dataset, it is recorded only for analysis and is not used during teacher inference.

C.3 Pair-wise Prompt and Preference Extraction

This section describes how to obtain pair-wise preferences and how to extract individual A/B decisions.

Prompt.

> `<image><image>`
> *You are an expert image quality evaluator. You will be given TWO images: Image A and Image B. Please compare their OVERALL subjective visual quality (e.g., sharpness, noise, artifacts, aesthetics). Choose which one is better in overall quality:*
> *- Output ONLY ONE character: A or B*
> *- No extra words, no punctuation, no explanation*
> *- If they are close, you MUST still choose A or B.*
> *Answer:*

Pair construction. We construct the pair-wise supervision set by uniformly sampling image pairs at random from the *Train* split, with a fixed random seed and duplicate removal. Following the main paper, the number of sampled pairs is set to match the number of images, i.e., $|\mathcal{P}| = |\mathcal{D}_{\text{Train}}|$ for each dataset. Each sampled pair is ordered as (x_a, x_b) and presented to the teacher as `Image A` and `Image B` in the prompt.

Decision extraction and decision position. To ensure a well-defined decision position for probability extraction, we define the decision token as the *first* generated token after the `Answer:` prefix. We normalize token strings by stripping leading/trailing whitespace and removing common tokenizer boundary markers, and then map the normalized token to `A` or `B` if matched. If the first generated token cannot be mapped to A/B, the pair is discarded. This design aligns the hard decision and the probability extraction to the same decision position.

C.4 Preference Probability and Reliability Weight

At the decision position, we read the teacher log-probabilities and compute a two-class distribution over $\{\texttt{A}, \texttt{B}\}$. Let (ℓ_A, ℓ_B) denote the token log-probabilities for `A` and `B` at the decision position, respectively. We then compute:

$$p_A = \frac{e^{\ell_A}}{e^{\ell_A} + e^{\ell_B}}, \quad p_B = \frac{e^{\ell_B}}{e^{\ell_A} + e^{\ell_B}}.$$

The hard preference label is

$$t_{ab} = \begin{cases} 1, & \text{if } p_A \geq p_B \text{ (teacher prefers A)}, \\ 0, & \text{otherwise.} \end{cases}$$

Confidence weighting. Not all teachers' decisions are equally reliable. We quantify confidence using the normalized entropy of $p = (p_A, p_B)$:

$$H(p) = -\sum_{c\in\{A,B\}} p_c \log(p_c + \epsilon), \qquad \omega_{ab} = 1 - \frac{H(p)}{\log 2} \in [0, 1],$$

where $\epsilon = 10^{-12}$ ensures numerical stability. Larger ω_{ab} indicates a more confident decision. In practice, we discard low-confidence pairs with $\omega_{ab} < \tau$; unless otherwise stated, we set $\tau = 0.05$.

Outputs. Each valid pair is stored with its image identifiers (x_a, x_b), the preference label t_{ab}, probabilities (p_A, p_B), and the confidence weight ω_{ab}.

D Student Setup

D.1 Student Architecture

This subsection describes the student model used for quality regression. The student model is a ConvNeXt-based regressor that outputs a single scalar quality score for each image. Unless otherwise specified, all experiments use the ConvNeXt-Base version with ImageNet pre-trained weights. The regressor head maps pooled backbone features to a 1D prediction.

D.2 Input Resolution and Preprocessing

This subsection describes the image preprocessing methods used for training and evaluation. All images are resized to 256 on the shorter side and then cropped to 224×224. During the training process, we apply random resized cropping (scale in $[0.8, 1.0]$) and random horizontal flipping (probability 0.5). During evaluation, we use center cropping. Images are normalized using the mean and standard deviation provided by the corresponding pretrained ConvNeXt weight metadata.

D.3 Two-stage Training Protocol

This section introduces the two-stage optimization process used to train the students.

Stage 1: teacher-guided distillation. The student is trained using the following methods: (i) point-wise regression to the teacher's soft scores and (ii) pair-wise ranking supervision from teacher-generated comparisons. The point-wise objective uses SmoothL1 loss (Huber loss, $\beta = 1.0$) between the student prediction and the teacher score. The overall objective is a weighted sum of point-wise and pair-wise losses, with the pair-wise term weighted by $\lambda_{\text{dis}} = 0.5$. We train for 15 epochs using AdamW with learning rate 1×10^{-5} and weight decay 1×10^{-4}, and apply a linear learning-rate decay to zero over the 15 epochs.

Stage 2: MOS calibration. We fine-tune the student on a MOS-visible subset sampled from the *Train* split to align the prediction results with the dataset-specific MOS scale.

The MOS-visible ratio r follows the main paper and takes values in $\{0, 0.1, 0.3\}$. When $r = 0$, Stage 2 is skipped, and the student is evaluated directly after Stage 1. When $r > 0$, we uniformly sample an $r\%$ subset from the *Train* split as $\mathcal{D}_{\text{MOS}}$ for calibration fine-tuning.

Let $\mathcal{D}_{\text{MOS}} = \{(x_i, y_i)\}$ denote the MOS-visible subset, where y_i is the MOS label. We optimize a regression objective together with a differentiable Pearson correlation objective computed within each mini-batch:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_{\text{MOS}}}\left[\big(s_\theta(x_i) - y_i\big)^2\right],$$

$$\mathcal{L}_{\text{PLCC}} = 1 - \rho(\mathbf{s}, \mathbf{y}), \qquad \rho(\mathbf{s}, \mathbf{y}) = \frac{\operatorname{cov}(\mathbf{s}, \mathbf{y})}{\sigma(\mathbf{s})\,\sigma(\mathbf{y}) + \epsilon},$$

Table 4: Effect of the number of sampled pairs used in Stage 1 pair-wise distillation on AGIQA-3K. We vary the pair budget as $|\mathcal{P}| = \alpha\,|\mathcal{D}_{\text{Train}}|$ while keeping all other settings fixed.

Pair budget α	SRCC	PLCC
0	0.834	0.892
0.25	0.837	0.894
0.5	0.839	0.894
1.0	0.841	0.899
2.0	0.844	0.900

where $\mathbf{s} = [s_\theta(x_i)]$ and $\mathbf{y} = [y_i]$ are the prediction and MOS vectors within a mini-batch, and $\epsilon = 10^{-12}$. The overall objective is

$$\mathcal{L}_{\text{cal}} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{cal}}\mathcal{L}_{\text{PLCC}}.$$

Unless otherwise stated, we set $\lambda_{\text{cal}} = 0.5$.

Optimization details. Both stages adopt mixed precision training. For both single-point and double-point samples, the batch size is 32, and the random seed is fixed at 42.

D.4 Pair-wise Supervision in Student Training

This subsection describes how the teacher-generated paired data is applied during the training process. Paired supervision is loaded from the paired dataset generated by the teacher and filtered through a minimum Confidence Level threshold of 0.05. For each pair, the loss contribution can be weighted; by default, we use the teacher-provided reliability weight ω_{ab}.

E Additional Experiment Details

E.1 Effect of the Number of Sampled Pairs

Table 4 shows that introducing pair-wise supervision consistently improves performance over using point-wise supervision alone ($\alpha = 0$). Increasing the pair budget yields monotonic but gradually saturating gains, with most of the improvement achieved around $\alpha = 1$. Therefore, we adopt $|\mathcal{P}| = |\mathcal{D}_{\text{Train}}|$ as the default setting in the main paper, balancing performance and supervision cost.

E.2 Efficiency of LEAF

The overall cost of LEAF is shown below. All results are measured on an NVIDIA GeForce RTX 4090.

(1) Label efficiency: LEAF (10% MOS) achieves $0.841/0.899$, while existing fully supervised methods (Q-Align) achieve $0.852/0.881$.

(2) At training, Teacher: 8B (used for offline inference, only once, Inference latency: 127.87 ms/image, GPU memory: 16.95 GB) Student: 88.6 M (1.1% parameters compared to the teacher, GPU memory: 14.47 GB)

(3) At deployment, only student (Parameters: 88.6 M, Inference latency: 11.10 ms/image, GPU memory: 0.51 GB)

The proposed method significantly reduces both annotation and inference-time costs.

F Datasets and Evaluation Protocol

F.1 UGC vs. AIGC: Conceptual Differences

This section aims to clarify the differences between UGC (User-Generated Content) and AIGC (AI-Generated Content) benchmarks. UGC images are captured in real-world environments and typi-

cally contain real distortion elements, such as compression artifacts, sensor noise, exposure issues, motion blur, and compression distortion. In contrast, AIGC images are generated by generative models and often exhibit generation-specific degradations, including synthetic artifacts, structural inconsistencies, texture drifting, and over-sharpening or over-smoothing effects. Since these two data sources originate from different distortion mechanisms, their perceptual score distributions and associated MOS metrics may vary, which prompts explicit score calibration in benchmark testing.

F.2 Dataset Descriptions

This subsection summarizes the benchmarks used in our experiments.

KonIQ-10k (UGC). KonIQ-10k is an in-the-wild IQA benchmark containing 10,073 images with diverse content and authentic distortions. Subjective quality is collected via crowdsourcing using a 5-point absolute category rating (ACR) protocol, and MOS is provided for each image.

SPAQ (UGC). SPAQ is a smartphone photography IQA benchmark containing 11,125 images captured by 66 smartphones. It provides MOS obtained from a well-controlled subjective study, with MOS values reported on a continuous scale (0–100). This dataset focuses on real mobile shooting conditions and the resulting real degradation scenarios.

AGIQA-3K (AIGC). AGIQA-3K contains 2,982 text-to-image samples generated by six representative generative models spanning different paradigms. The dataset provides human MOS annotations for AI-generated images. In this work, we only used the MOS of perception (visual quality) for training and evaluation, and did not involve text-image alignment evaluation.

AIGIQA-20K (AIGC). AIGIQA-20K is a large-scale AI-generated IQA benchmark that includes 20,000 images generated by 15 mainstream text-to-image models. It provides MOS values obtained through subjective evaluation, and its scale and model coverage can adapt to evaluations under a wide range of generation settings.

F.3 Evaluation Metrics and Reporting

This subsection describes the evaluation metrics and the reporting protocol. Given predictions $\{\hat{y}_i\}_{i=1}^N$ and MOS labels $\{y_i\}_{i=1}^N$, we report SRCC and PLCC.

SRCC. Let $R(\hat{y}_i)$ and $R(y_i)$ denote the ranks of $\hat{y}_i$ and y_i, and $d_i = R(\hat{y}_i) - R(y_i)$. Then

$$\mathrm{SRCC} \;=\; 1 \;-\; \frac{6\sum_{i=1}^{N} d_i^2}{N\,(N^2-1)}\,. \tag{13}$$

PLCC. Let $\bar{\hat{y}} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. Then

$$\mathrm{PLCC} \;=\; \frac{\sum_{i=1}^{N}(\hat{y}_i-\bar{\hat{y}})\,(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(\hat{y}_i-\bar{\hat{y}})^2}\,\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}}\,. \tag{14}$$

F.4 Split Strategy

This subsection describes how we constructed the training subset and how the MOS visible subset used for calibration was sampled. We follow the protocol of the official dataset and do not create separate training/validation/test partitions. All training samples are drawn *only* from the official *Train* split when it is provided. For datasets without official split files, we create a fixed random *seen/unseen* partition at the image level with a fixed seed, where the *seen* part is used for training and the *unseen* part remains hidden throughout model development. All unseen images are complementary to the training subset and are never accessed for training, model selection, or hyperparameter tuning.

Within the training-visible set (the official *Train* split or the constructed *seen* partition), we randomly sample a MOS-visible subset $\mathcal{D}_{\mathrm{MOS}}$ whose size corresponds to an $r\%$ portion of the *entire dataset*,

where $r \in \{0, 10, 30\}$. When $r > 0$, the subset is drawn uniformly at random with fixed seeds and is used only in Stage 2 for MOS calibration; when $r = 0$, Stage 2 is skipped.

F.5 Allowed Supervision Signals by Stage

This subsection specifies which supervision signals are used in each stage.

Stage 1: teacher-guided distillation. Stage 1 uses only teacher-generated supervision: (i) point-wise soft scores and (ii) pair-wise preferences with confidence weights. No human MOS labels are used for optimization in Stage 1.

Stage 2: MOS calibration. Stage 2 uses MOS labels only from the MOS-visible subset $\mathcal{D}_{\text{MOS}}$ sampled from the training-visible set (with ratio $r \in \{0, 10, 30\}$, measured w.r.t. the entire dataset) to align the student predictions to the dataset-specific MOS scale. When $r = 0$, Stage 2 is not performed. No images or MOS labels from the unseen split are used for optimization, model selection, or hyperparameter tuning.

F.6 Reproducibility

All partitions and MOS subsets are generated using a fixed random seed and reused across multiple runs, thereby ensuring that all reported results are obtained under the same data partitioning conditions. For the setting of $r > 0$, we sample the MOS subset five times using different random seeds and report the average SRCC/PLCC values obtained from these five runs.

G Limitations

Although LEAF reduces the need for MOS annotations and avoids deploying large MLLMs at inference time, it still relies on an MLLM teacher to generate offline supervision signals. Therefore, the quality and efficiency of the framework are partly bound by the teacher model. In particular, if the teacher exhibits systematic errors on certain fine-grained distortions, artistic styles, or generation-specific artifacts, the student may inherit these biases, and a small MOS calibration subset may not fully correct them. In addition, this work mainly focuses on image quality assessment. Extending the perception-calibration decoupling principle to video quality assessment requires further consideration of temporal artifacts, motion consistency, and frame-level quality aggregation. We leave these directions for future work.