# Q-HARNESS: FEW-SHOT NO-REFERENCE IMAGE QUALITY ASSESSMENT AS A SCALABLE SKILL

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

No-reference image quality assessment (NR-IQA) is often needed in settings where a new camera pipeline, restoration method, compression codec, or generative model creates a new target domain faster than a large mean-opinion-score (MOS) database can be collected. In such deployments, NR-IQA should be *few-shot* by design: a practical assessor should adapt from only a few labeled examples and be ready to use. Existing per-dataset fine-tuning does not match this need, because it spends scarce labels re-estimating a large perceptual model, overfits target-domain statistics, and must be repeated for every new domain. We instead frame NR-IQA as a scalable skill of foundation models: modern vision-language models (VLMs) already contain strong quality priors, and the right interface is to elicit and calibrate this skill rather than retrain it. We present **Q-Harness**, a harness system that turns a *frozen* general-purpose VLM into a few-shot quality assessor without backbone training. Given a small set of expert labels, it *elicits* interpretable per-attribute quality subscores (sharpness, noise, exposure, composition, etc.) and maps them to a final MOS through a lightweight ridge calibration layer fitted on those labels. A controlled ablation shows that the harness's gain is *semantic*: querying the same construct in multiple ways yields little benefit, irrelevant attributes receive near-zero weight, and performance improves as quality attributes become less redundant.

## 1 INTRODUCTION

No-reference image quality assessment (NR-IQA) asks a model to predict the mean opinion score (MOS) that a crowd of human raters would assign to an image, given only the image itself. It underpins image acquisition, compression, restoration, and the ranking of generative models, and is measured by the Spearman rank correlation (SRCC) between predicted and ground-truth scores. In practice, however, NR-IQA is rarely just a full-data benchmark problem. A new camera pipeline, restoration model, compression codec, or text-to-image generator can shift the visual domain overnight. For such deployments, the realistic requirement is *few-shot*: collect a small panel of ratings, quickly fit a quality ranker, and use it immediately.

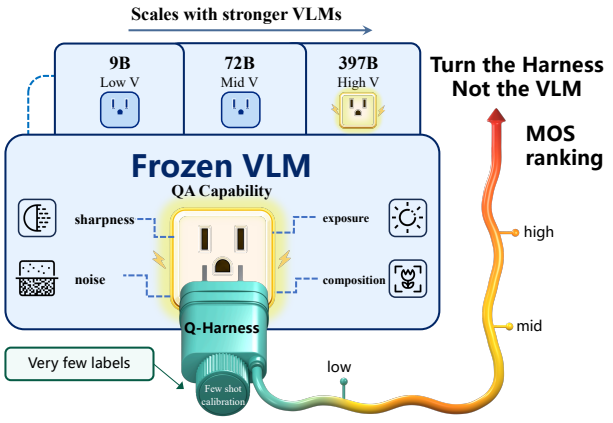


Figure 1: NR-IQA as a scalable skill. A frozen VLM already supplies a quality-assessment ability. Few-shot labels tune the Q-Harness calibration, not the VLM itself. As stronger VLMs replace the backbone, the same harness can read out a stronger quality skill through few-shot ridge calibration.

The dominant supervised recipe is poorly matched to this requirement. Fine-tuning a deep network or multimodal model on a target quality database (Zhang et al., 2020, Su et al., 2020, Ke et al., 2021, Yang et al., 2022, Wu et al., 2024c) uses scarce labels to re-estimate the entire mapping from pixels to quality, entangling domain-general perception with domain-specific scale (Hosu et al., 2020,



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.

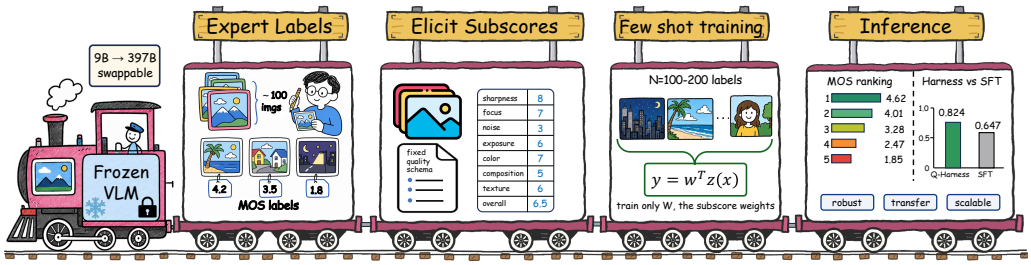


Figure 2: Overview of Q-Harness. A frozen VLM produces per-attribute subscores via a fixed quality schema. A ridge calibration layer trained on 100–200 labels maps them to MOS, achieving PLCC 0.8244 vs. 0.6473 for per-dataset fine-tuning.

Ghadiyaram & Bovik, 2016). This creates two concrete problems: first, with few labels the specialist overfits the target distribution and degrades sharply under label skew. Second, across domains the resulting checkpoint transfers poorly, because it has encoded the statistical characteristics of the source domain. A practical few-shot system should separate visual quality perception, a capability the model already possesses, from the mapping to the target MOS scale. Only this mapping requires adaptation using the few labels available in a specific deployment domain.

We therefore frame NR-IQA as a *scalable skill* of foundation models, as illustrated in Figure 1. Modern vision-language models (VLMs) already contain strong perceptual priors from web-scale pre-training (Radford et al., 2021, Wang et al., 2024, Wu et al., 2024a) and can judge sharpness, noise, exposure, color, and composition. The few labels available in a new domain should not be used to re-learn these perceptions. They should be used to calibrate the mapping between the frozen model's judgments and the target MOS scale. This also makes the system scalable: when a stronger VLM becomes available, the front end can be replaced while retaining the same procedure for eliciting attribute scores and calibrating them to the target MOS scale.

To address these challenges, we propose **Q-Harness**, a harness system that converts a frozen VLM into a few-shot quality assessor, as shown in Figure 2. Given a small set of expert labels, the frozen VLM is prompted to report interpretable per-attribute quality subscores (sharpness, noise, exposure, composition, *etc.*), each on a fixed numeric scale, without updating any backbone weights. A lightweight ridge calibration layer, fitted on the labeled set, then maps the elicited subscores to a final MOS. A controlled ablation further confirms that the subscores carry distinct, decomposed perceptual signal rather than redundant restatements of a single judgment. Because the backbone is never trained, the harness is inexpensive to deploy, straightforward to retarget to a new VLM, and learnable from a handful of labels. Our main contributions are:

- We propose a **few-shot skill framing for NR-IQA** and argue that frozen VLMs provide a scalable perceptual front end whose quality skill can be elicited and calibrated rather than retrained separately for each new quality-assessment domain.
- We propose **Q-Harness**, a harness system that elicits per-attribute subscores from a frozen VLM and maps them to MOS via a few-shot ridge calibration layer, with no backbone training. A controlled ablation confirms the gain is semantic rather than statistical.
- Our Q-Harness achieves **superior label efficiency, robustness, and scalability**, outperforming per-dataset fine-tuning on all five benchmarks and all 25 transfer cells (mean SRCC 0.807 vs. 0.629), while remaining nearly invariant to label-distribution skew.

## 2 Related Work

**Learning-based NR-IQA.** Classical no-reference metrics use handcrafted natural-scene statistics, with supervised BRISQUE and opinion-unaware NIQE (Mittal et al., 2012, 2013). The field then moved to supervised deep models: bilinear CNNs that handle both synthetic and authentic distortions (Zhang et al., 2020), self-adaptive hyper-networks (Su et al., 2020), multi-scale (Ke et al., 2021) and attention-based (Yang et al., 2022) transformers, and aesthetic predictors that regress a score distribution (Talebi & Milanfar, 2018). These achieve high in-domain SRCC but require thousands

of labeled images per target database and transfer unevenly across domains. In contrast, Q-Harness requires only 100–200 labels and transfers without retraining.

**Vision-language models for quality.** Because perceptual quality is entangled with language ("a sharp, well-exposed photo"), CLIP-style models can rank quality zero-shot via antonym prompts (Radford et al., 2021, Wang et al., 2023), and instruction-tuned multimodal models can describe and reason about low-level attributes (Liu et al., 2023, Wu et al., 2024a,b). CLIP-IQA also operates on a frozen backbone, using fixed antonym prompts in its zero-shot version and learned prompts in CLIP-IQA+. At the other extreme, several systems *fine-tune* a multimodal model specifically for scoring (Wu et al., 2024c, Zhu et al., 2024), spending scarce labels on backbone re-estimation. Q-Harness occupies the middle ground: it keeps the VLM frozen like CLIP-IQA but adds a validated, learnable calibration layer over quality-relevant subscores, and we show it outperforms the fine-tuned specialist introduced by Wu et al. (2024c) even at a few-shot budget.

**Attribute decomposition.** That quality decomposes into perceptual attributes is well established: SPAQ collects per-attribute human annotations (brightness, colorfulness, contrast, noisiness, sharpness) alongside overall MOS (Fang et al., 2020), and CLIP-IQA shows that antonym prompts can probe fine-grained image attributes (Wang et al., 2023). Prior work uses such attributes as auxiliary supervision for a trained network. Q-Harness instead makes the elicited quality attributes the *sole representation*: the top-$N$ subscores serve directly as the feature vector for a lightweight calibration layer, with no backbone features involved.

## 3 The Q-Harness Framework

Q-Harness extracts a quality score from a frozen vision-language model in three steps, as illustrated in Figure 2: a small set of expert labels is collected, the frozen VLM *produces* per-attribute quality subscores for each image, and a lightweight ridge calibration layer is *fitted* on those labels to map the subscores to a final MOS prediction. At inference time, the same elicitation procedure and calibration layer produce scores for new images. We additionally present a controlled ablation that *validates* that the subscores carry decomposed perceptual signal rather than redundant copies of a single judgment. We first state the perceptual view that motivates the design, then describe each component.

### 3.1 Quality as Attribute Aggregation

We model the overall quality $y$ a human assigns to an image $x$ as an aggregation of $p$ distinct perceptual attributes, as depicted in Figure 3:

$$y \approx g\big(a_1(x), a_2(x), \ldots, a_p(x)\big), \tag{1}$$

where each attribute function $a_j(x)$ (*e.g.*, sharpness, noise, exposure, color, composition) is a hard non-linear mapping from pixels to a perceptual judgment, but the aggregation function $g$ is simple and stable across images. This view is consistent with multi-attribute subjective IQA protocols such as SPAQ (Fang et al., 2020), which collect separate judgments for brightness, colorfulness, contrast, noisiness, and sharpness alongside an overall quality score. Each $a_j$ is exactly the kind of judgment a large VLM has learned to make from web-scale supervision, whereas $g$ has few effective degrees of freedom. Q-Harness assigns the hard part to the frozen model and learns only $g$. This is why a small labeled set suffices and why the framework transfers: the

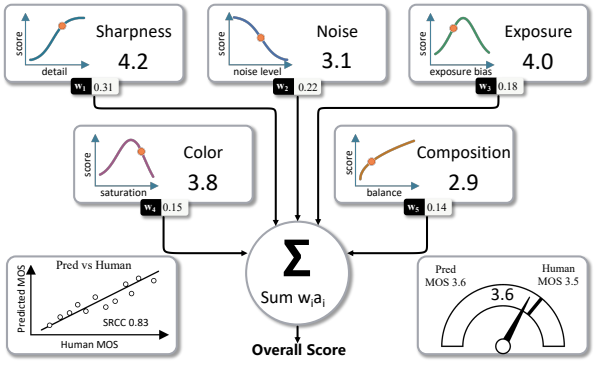


Figure 3: Perceptual view motivating Q-Harness: distinct non-linear attribute perceptions $a_j(x)$ are aggregated by simple weights $w_j$ into MOS. The VLM supplies the hard perceptions. Only the weights are learned (inset, SRCC 0.83 between predicted scores and human MOS).

attribute perceptions are domain-general,
and only their aggregation is fitted per domain.

### 3.2 Eliciting Quality Subscores from a Frozen VLM

Let $M$ denote a frozen instruction-tuned VLM. For an image $x$, we issue a single structured prompt that asks $M$ to rate the image on each of $p$ named quality attributes on a common numeric scale (1–5), and to also provide an overall rating. The prompt specifies the attribute names, the rating scale, and the output format (*e.g.*, a JSON dictionary mapping attribute names to scores), so that parsing is deterministic. We use a low decoding temperature to minimize stochastic variation and issue one elicitation per image with no resampling. Images are evaluated at high resolution ($1024 \times 768$), which we found necessary for stable attribute perception. Lower resolutions degrade the assessment of fine-grained attributes such as noise and texture.

Parsing the response yields a feature vector:

$$\phi(x) = [\hat{a}_1(x), \ldots, \hat{a}_p(x), \hat{m}(x)] \in \mathbb{R}^{p+1}, \tag{2}$$

where $\hat{a}_j$ is the model's reported subscore for attribute $j$ and $\hat{m}$ is its overall rating. We concatenate the outputs of two overlapping prompt variants of the same Q-Harness pipeline. One covers 13 attributes: sharpness, focus, noise, motion blur, flicker, exposure, compression, fluency, content, color, lighting, aesthetic quality, and technical quality. The other covers sharpness, noise, lighting, composition, and texture. Each also provides an overall rating, giving 20 scalar features whose perceptual meanings overlap. These attributes were selected from the quality vocabulary of existing IQA databases (Fang et al., 2020) and VLM-quality benchmarks (Wu et al., 2024a), covering both technical distortions and aesthetic properties. The backbone is never updated, and elicitation requires a constant number of forward passes per image regardless of the number of attributes.

### 3.3 Schema Validation via Controlled Ablation

A subscore schema is only worth deploying if the subscores carry genuinely informative, decomposed signal. A natural concern is that simply aggregating repeated VLM queries might perform equally well, in which case the "decomposition" would amount to variance reduction rather than semantic information. Before fitting any calibration layer, we verify that this is not the case with a controlled ablation in Section 4.2 that asks three questions: Does a semantic decomposition outperform simply querying overall quality several times (*i.e.*, a self-ensemble)? Are irrelevant, non-quality attributes correctly ignored by the calibration layer? Does the benefit scale with attribute diversity, as Equation 1 predicts? If the gain came merely from statistical variance reduction across multiple queries, then arms B and C would perform comparably to a genuine quality decomposition. Passing all three checks certifies that $\phi(x)$ carries decomposed quality signal rather than redundant copies of a single judgment. This validation is performed once when selecting a schema. At deployment time only the chosen top-$N$ attributes are elicited.

### 3.4 Few-Shot Deployment via Calibration Layer

Given a small labeled set $\{(x_i, y_i)\}_{i=1}^{n}$, the harness fits the aggregation $g$ as a regularized linear calibration layer. We standardize each feature using the fit-set statistics and solve a closed-form ridge regression to map the standardized subscores to centered MOS targets:

$$w^\star = \arg\min_{w} \sum_{i=1}^{n} \left(w^\top z_i - (y_i - \bar{y})\right)^2 + \lambda \|w\|_2^2, \qquad z_i = (\phi(x_i) - \mu)/\sigma, \tag{3}$$

where $\mu$ and $\sigma$ are the per-feature mean and standard deviation computed on the fit set, $\bar{y}$ is the mean MOS, and the regularization coefficient $\lambda$ is selected from a logarithmic grid by $k$-fold cross-validated SRCC *inside the fit set* ($k$=5 when the fit set is large enough). Missing features are mean-imputed (*i.e.*, assigned a $z$-score of 0). At test time, the predicted score for an image is $\hat{y} = w^{\star\top} z(x)$.

Because SRCC is invariant to affine transforms, no intercept or per-dataset rescaling is needed, so a calibration layer fitted on one domain can be applied unchanged to another. This property is what makes the $5 \times 5$ cross-dataset transfer grid in Section 4.3 possible without any adaptation. The

calibration layer has only $p+1$ parameters, which makes it learnable from 100–200 labels as shown in Section 4.3 and stable under label-distribution skew.

When using fewer than the full 20 features, we select the top-$N$ by standalone importance: each feature is scored by its individual SRCC with MOS on a held-out validation set, and the top-$N$ are retained. This greedy selection is fast and interpretable. The backbone scaling study uses the top-10 features under a fixed protocol (Section 4.6). Swapping the backbone $M$ for a stronger VLM requires re-eliciting $\phi$ but no other change to the pipeline, and performance improves monotonically.

# 4 Experiments

Our experiments show that Q-Harness outperforms per-dataset fine-tuning across five IQA benchmarks with only 100–200 labels. Controlled ablations support the contribution of semantic attribute decomposition, and further experiments demonstrate cross-dataset transfer and improved performance with stronger backbones. We also examine label-budget saturation, robustness to label-distribution skew, and performance against full-data NR-IQA baselines.

## 4.1 Experiment Setup

Unless noted otherwise, the frozen backbone is a locally served, general-purpose 9B VLM, the larger 397B variant of the same Qwen 3.5 family is used for the scaling study in Section 4.6. The default calibration uses 20 scalar features from the two overlapping prompt variants, including their separate overall ratings. The calibration layer is the ridge regression of Equation 3, with $\lambda$ selected by cross-validated SRCC inside the fit set, separately from the held-out images used for evaluation.

We evaluate on five standard IQA databases spanning distinct domains: KonIQ-10k (Hosu et al., 2020) for in-the-wild web photos, SPAQ (Fang et al., 2020) for smartphone photos, CLIVE (Ghadiyaram & Bovik, 2016) for the LIVE In-the-Wild Challenge, AGIQA-3K (Li et al., 2024) for AI-generated images, and KADID-10k (Lin et al., 2019) for synthetic distortions. These datasets differ in content, distortion type, and MOS scale, with native ranges including 1–5, 0–100, and 0–4, providing a broad test of generalization. The metrics are SRCC and PLCC, with PLCC computed after the standard four-parameter logistic remapping. Training images are removed before computing either correlation, so every reported number is held out.

The principal baseline is Q-Align (Wu et al., 2024c), a dedicated IQA specialist built on mPLUG-Owl2-7B. We fine-tune it per dataset for quality scoring using the same label budget as our calibration layer. This gives a like-for-like comparison: identical labels and images, with a fine-tuned specialist on one side and a frozen backbone with the harness on the other.

## 4.2 Validation of the Elicited Signal

The central claim of the elicitation stage is that its gain comes from *semantic decomposition* rather than from issuing several VLM queries and averaging them. If the benefit were merely statistical, that is, variance reduction from querying the model multiple times, then *any* set of repeated queries should produce a comparable gain regardless of what is asked. To test this, we hold the entire pipeline fixed and vary only the meaning of the five elicited attributes, as shown in Figure 4. **A.** A genuine quality decomposition (sharpness, noise, lighting, composition, texture). **B.** The same overall-quality question asked five times, *i.e.*, a self-ensemble with identical prompts. **C.** Five paraphrased or persona-conditioned versions of the overall-quality question, maximizing linguistic diversity without changing the underlying construct.

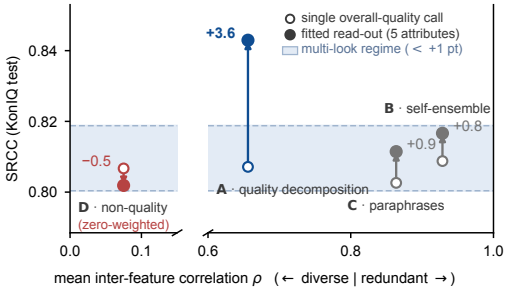


Figure 4: Validation ablation (KonIQ test, frozen 9B). $\Delta$SRCC$\times$100 of the calibration layer over a single overall-quality call, plotted against mean inter-feature correlation $\rho$. Only the genuine quality decomposition (A) escapes the shaded no-semantics band of repeated-query controls.

Table 1: In-domain held-out SRCC/PLCC at a few-shot label budget (training images removed before scoring). PLCC uses the standard four-parameter logistic remapping. "Q-Harness (no fit)" uses the frozen VLM's overall rating directly with no calibration. Q-Harness wins every dataset at both budgets, beating per-dataset fine-tuning of the specialist by $+0.18$ mean SRCC. The calibration helps most on the harder non-natural sets (AGIQA-3K, KADID).

| | $n = 100$ | | | $n = 200$ | | |
|---|---|---|---|---|---|---|
| Dataset | Q-Align | No fit | Q-Harness | Q-Align | No fit | Q-Harness |
| KonIQ | 0.561/0.595 | 0.806/0.824 | **0.828/0.868** | 0.606/0.641 | 0.796/0.812 | **0.834/0.865** |
| AGIQA-3K | 0.710/0.778 | 0.633/0.701 | **0.771/0.828** | 0.751/0.775 | 0.628/0.696 | **0.761/0.833** |
| KADID | 0.634/0.628 | 0.684/0.689 | **0.747/0.723** | 0.714/0.714 | 0.683/0.689 | **0.752/0.741** |
| CLIVE | 0.593/0.622 | 0.759/0.778 | **0.795/0.840** | 0.689/0.710 | 0.744/0.767 | **0.791/0.833** |
| SPAQ | 0.647/0.623 | 0.879/0.872 | **0.894/0.899** | 0.700/0.688 | 0.879/0.872 | **0.904/0.908** |
| Mean | 0.629/0.649 | 0.752/0.773 | **0.807/0.832** | 0.692/0.706 | 0.746/0.767 | **0.808/0.836** |

**D.** Five quality-irrelevant attributes (*e.g.*, presence of people, outdoorness, clutter, vegetation), which provide diverse views but no quality information.

Only the semantic decomposition adds meaningful signal. Re-asking the same construct gains little over a single call, whether identically (B, $+0.8$) or with maximal linguistic diversity (C, $+0.9$), whereas the quality decomposition (A) gains $+3.6$. This threefold difference cannot be explained by variance reduction alone, since arms B and C issue the same number of queries. Irrelevant attributes (D) reduce accuracy slightly ($-0.5$) and receive near-zero weights from the calibration layer, confirming that the ridge regression does not blindly exploit any additional features but selectively weights quality-relevant ones. Moreover, as Equation 1 predicts, the gain tracks attribute diversity: arm A has the lowest inter-feature correlation ($\rho$=0.66) and the highest accuracy, while the redundant ensembles sit at $\rho \geq 0.86$ with no meaningful gain. This pattern is precisely what we expect if the calibration layer is leveraging independent perceptual dimensions rather than averaging correlated copies of the same overall-quality judgment elicited through repeated VLM queries.

### 4.3 Label Efficiency at Few-Shot Budgets

We next compare Q-Harness against per-dataset fine-tuning of the specialist at a fixed, small label budget. Each dataset supplies $n$ uniformly sampled labeled images. Both methods see the same images, which are removed before scoring. Table 1 reports the in-domain held-out SRCC at $n$=100 and $n$=200 labeled images per dataset, with the label budget matched between the two methods.

At $n$=100, Q-Harness reaches a mean held-out SRCC of $0.807$, compared with $0.629$ for the fine-tuned specialist, and is more accurate on each of the five datasets (Figure 6, bottom panel). The gap ($+0.178$) is large by IQA standards, where methods often differ by $0.01$–$0.03$ SRCC. Even the uncalibrated overall rating of the frozen VLM exceeds the fine-tuned specialist on four of the five datasets, indicating that with only a hundred labels, fine-tuning a 7B specialist is less accurate than calibrating the frozen model directly. This confirms our perceptual view: the VLM already has strong quality perception, and the few available labels are better spent on calibration than on re-learning perception. Fitting the calibration layer adds a further $+0.06$ mean SRCC, concentrated on the non-natural domains (AGIQA $+0.14$, KADID $+0.06$), where the overall rating alone is insufficient and a learned aggregation of per-attribute signals helps most.

We also evaluate cross-dataset transfer. Figure 5(b) shows the full $5 \times 5$ transfer grid: the calibration layer is fitted on one dataset and applied to all five, and each cell reports the difference from the fine-tuned specialist trained and evaluated analogously. All $25$ cells favor Q-Harness, both in-domain and under transfer. The margins are largest where the specialist's domain specialization is most damaging: a calibration layer fitted on SPAQ (smartphone photos) transfers to KADID (synthetic distortions) with a $+0.40$ advantage, whereas the fine-tuned specialist collapses when applied out of domain. This behavior is expected: because the calibration layer fits only the aggregation weights over domain-general perceptions, it does not overfit source-domain statistics the way full fine-tuning does when updating the backbone on the same small labeled set.

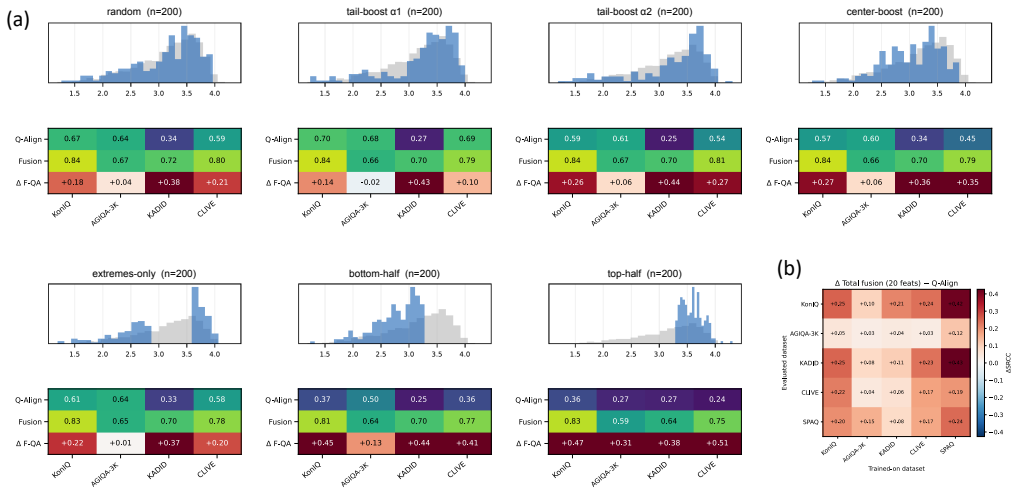


Figure 5: (a) Robustness to label-MOS skew ($n$=200, KonIQ): seven sampling policies with in-domain and cross-dataset SRCC. Q-Harness stays high across all policies while the fine-tuned specialist degrades sharply. (b) Cross-dataset ΔSRCC of 20-feature fusion over Q-Align. All 25 cells favor Q-Harness across the five source datasets and five evaluation datasets.

### 4.4 Robustness to Label-Distribution Skew

In practice, a few-shot label set is rarely balanced: an engineer may opportunistically label whatever images are at hand, which may skew toward easy or hard examples. We therefore measure sensitivity to the label distribution explicitly. We fix the dataset (KonIQ) and the label budget (200 images), and vary the MOS distribution across seven sampling policies. These range from uniform and tail-boosted sampling to severe skew, using only the top or bottom half of the quality range. Both methods are refitted under each policy, and Figure 5(a) reports their in-domain and cross-dataset SRCC on held-out images.

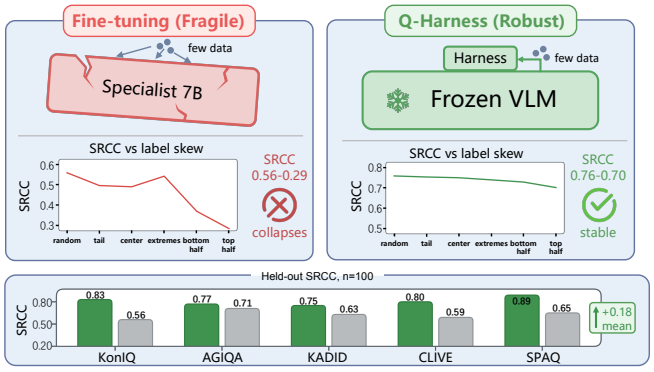


Figure 6: Top (200 KonIQ training labels): mean SRCC over KonIQ, AGIQA-3K, KADID, and CLIVE under label skew. Fine-tuning: 0.56–0.29. Q-Harness: 0.76–0.70. Bottom (100 training labels per dataset): held-out in-domain SRCC.

The fine-tuned specialist is highly sensitive to this skew: its four-dataset mean SRCC drops from 0.56 (uniform) to 0.29 (top-half only) as the label set becomes one-sided. This is expected: when the fine-tuned model sees only low-quality examples, it cannot learn to distinguish high-quality images, and vice versa. Q-Harness is more stable, remaining within 0.70–0.76 across six displayed policies, with a +0.41 SRCC advantage under top-half-only sampling. The explanation is structural (Figure 6, top panels): the frozen VLM's attribute perceptions are not affected by the label distribution because they are pre-trained, and the low-dimensional ridge calibration layer has too few parameters to overfit a skewed sample. This robustness is particularly important for real deployments where the label budget is small and the sampling process is uncontrolled.

Across sampling policies, the calibration domain, evaluation datasets, and label budget are held fixed. Only the quality distribution of the labeled set changes. The comparison therefore tests sensitivity to which quality levels receive supervision. These results support robustness to the seven tested policies, but do not establish invariance to arbitrary label distributions or distortions not captured by the frozen backbone's elicited quality attributes.

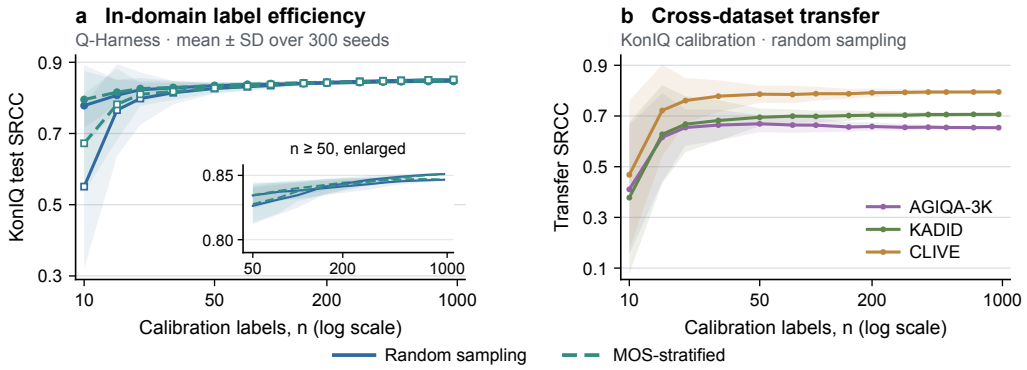


Figure 7: Calibration label efficiency of Q-Harness. (a) KonIQ SRCC under random and MOS-stratified sampling (inset: $n \geq 50$). Circles/squares denote overlapping 5-/13-attribute prompt variants, each with an overall score. (b) Cross-dataset SRCC using the 13-attribute variant, random KonIQ calibration, and fixed weights. Curves and bands show means $\pm 1$ sample standard deviation over 300 calibration seeds (maximum: 954 labels). Predictions and valid-image pools are fixed. The latter differ between the two overlapping prompt variants of the same Q-Harness pipeline.

### 4.5 Calibration Label Efficiency

Across the overlapping prompt variants, in-domain SRCC shows diminishing returns around $n \approx 200$–$300$ labels, as shown in Figure 7(a). Accuracy rises rapidly under both sampling policies, while variability across calibration seeds decreases as the budget grows. With KonIQ-fitted weights held fixed, most transfer gains occur by $n \approx 50$, as shown in Figure 7(b). Later changes are dataset-dependent. AGIQA-3K peaks near $n = 50$ and declines slightly, whereas KADID and CLIVE improve modestly. These trends support calibration with a few hundred labels.

### 4.6 Scaling with Stronger Backbones

Because the harness never trains the backbone, a stronger VLM can be substituted with no other change. This is a key advantage of the frozen-backbone design: as foundation models improve, the quality-assessment system automatically improves without re-engineering the pipeline. To isolate the effect of backbone scale, we fix the protocol: the fit set (a 200-image KonIQ validation split), test splits, prompts, schemas, top-10 feature-selection rule, and ridge calibration layer are held constant across all evaluated model families and backbone sizes.

Table 2: Fixed-protocol scaling study. Values are mean SRCC over the same five test splits. "Zero-shot" is the VLM's overall quality score. "Fusion" is the top-10 subscore calibration fit on the same KonIQ split.

| Family | Backbone | Zero-shot | Fusion | Δ |
|---|---|---|---|---|
| Qwen3.5 | 0.8B | 0.450 | 0.581 | +0.131 |
| | 4B | 0.774 | 0.795 | +0.021 |
| | 9B | 0.783 | 0.814 | +0.031 |
| | 397B | **0.816** | **0.844** | +0.028 |
| InternVL3.5 | 4B | 0.669 | 0.706 | +0.036 |
| | 8B | 0.728 | 0.751 | +0.023 |
| | 241B | 0.762 | 0.776 | +0.014 |
| Llama-3.2 Vision | 11B | 0.727 | 0.746 | +0.019 |
| | 90B | 0.790 | 0.807 | +0.016 |

Table 2 shows two consistent effects across three model families (Qwen3.5 (Qwen Team, 2026), InternVL3.5 (Wang et al., 2025), and Llama-3.2 Vision (Meta AI, 2024)) ranging from 0.8B to 397B parameters. First, stronger backbones provide stronger raw quality judgments: in the Qwen line, the largest gain occurs from 0.8B to 4B ($+0.324$ SRCC), reflecting the transition from a model that struggles with fine-grained perception to one with reliable attribute judgments. Later gains are smaller but still positive ($+0.033$ from 9B to 397B for zero-shot, $+0.030$ for fusion). Second, the harness adds value at every scale and across model families: fusion raises the Qwen3.5 mean from $0.450$ to $0.581$ at 0.8B (a $+0.131$ gain where the base model is weakest) and from $0.816$ to $0.844$ at 397B, and improves all InternVL3.5 and Llama-3.2 Vision configurations. The diminishing $\Delta$ at

Table 3: Comparison with standard IQA baselines. Results are reported as SRCC/PLCC and averaged over seven datasets. Best results are bold. Second-best results are underlined.

| Method | KonIQ | | CLIVE | | SPAQ | | AGIQA-3K | | BID | | KADID | | CSIQ | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC |
| *Handcrafted methods* | | | | | | | | | | | | | | | | |
| NIQE | 0.5271 | 0.5389 | 0.4351 | 0.4811 | 0.5037 | 0.5062 | 0.5346 | 0.5640 | 0.4582 | 0.4606 | 0.3794 | 0.4386 | 0.5788 | 0.6564 | 0.4881 | 0.5208 |
| BRISQUE | 0.1684 | 0.1662 | 0.3213 | 0.3704 | 0.0782 | 0.2378 | 0.5007 | 0.5493 | 0.0909 | 0.2647 | 0.3364 | 0.4015 | 0.4839 | 0.6888 | 0.2828 | 0.3827 |
| *Discriminative deep-learning methods* | | | | | | | | | | | | | | | | |
| MANIQA | 0.7133 | 0.7272 | 0.5514 | 0.5999 | 0.2104 | 0.2789 | 0.4729 | 0.5496 | 0.4424 | 0.4539 | 0.4100 | 0.4685 | 0.5619 | 0.5836 | 0.4803 | 0.5231 |
| UNIQUE | 0.7105 | 0.7798 | 0.5172 | 0.5810 | 0.5188 | 0.5295 | 0.5314 | 0.6260 | 0.6293 | 0.6439 | 0.2809 | 0.4000 | 0.6028 | 0.6571 | 0.5416 | 0.6025 |
| DEIQT | 0.6308 | 0.6656 | 0.4706 | 0.5322 | 0.6242 | 0.6273 | 0.3951 | 0.4569 | 0.4869 | 0.5195 | 0.2766 | 0.3270 | 0.2399 | 0.3814 | 0.4463 | 0.5014 |
| *VLM-based methods* | | | | | | | | | | | | | | | | |
| Q-Align | 0.6650 | 0.7138 | 0.5898 | 0.6362 | 0.7680 | 0.7747 | 0.6398 | 0.7472 | 0.5755 | 0.6029 | 0.3484 | 0.4670 | 0.4913 | 0.5893 | 0.5825 | 0.6473 |
| Qwen3-VL-8B | 0.7498 | 0.8092 | 0.7412 | 0.8086 | 0.8660 | 0.8760 | 0.6545 | **0.7937** | 0.7737 | 0.7761 | 0.3549 | 0.4933 | 0.4564 | 0.6269 | 0.6566 | 0.7405 |
| Qwen3.5-9B | 0.7501 | 0.7418 | 0.5314 | 0.6278 | 0.8180 | 0.8251 | 0.6162 | 0.6939 | 0.7238 | 0.6986 | 0.3481 | 0.3831 | 0.5307 | 0.5806 | 0.6169 | 0.6501 |
| **Q-Harness** | | | | | | | | | | | | | | | | |
| Qwen3.5-9B | 0.8416 | 0.8840 | 0.8114 | 0.8542 | 0.8971 | 0.8979 | 0.6644 | 0.7635 | 0.8457 | 0.8608 | 0.7019 | 0.6878 | 0.7669 | 0.8229 | 0.7899 | 0.8244 |
| Qwen3.5-397B | **0.8563** | **0.8920** | **0.8495** | **0.8715** | **0.8985** | **0.9026** | **0.7103** | 0.7855 | **0.8677** | **0.8760** | **0.7319** | **0.7381** | **0.7794** | **0.8311** | **0.8134** | **0.8424** |

larger scales is expected: a stronger base model already integrates attribute information into its overall rating, leaving less room for the calibration layer to add value. The same elicitation-and-calibration interface thus benefits from better perceptual backbones without any change to the training recipe.

### 4.7 Comparison with Full-Data NR-IQA Baselines

Finally, we compare the strongest fixed-protocol Q-Harness variants against standard NR-IQA baselines in Table 3. The baselines span three categories: handcrafted natural-scene-statistics methods such as opinion-unaware NIQE (Mittal et al., 2013) and supervised BRISQUE (Mittal et al., 2012), discriminative deep-learning models including MANIQA (Yang et al., 2022), UNIQUE (Zhang et al., 2021), and DEIQT (Qin et al., 2023) trained on IQA databases, and VLM-based approaches including Q-Align (Wu et al., 2024c) and general-purpose Qwen backbones that leverage vision-language pre-training. Each cell reports SRCC/PLCC. For Q-Harness, PLCC is computed after the standard IQA logistic remapping used by the baseline evaluator, and averages cover seven datasets: the five main datasets plus BID (Ciancio et al., 2011) and CSIQ (Larson & Chandler, 2010). Unlike the matched-label experiment in Section 4.3, this comparison includes methods with different training regimes. It therefore assesses predictive accuracy against established NR-IQA systems, while the earlier comparison isolates performance under the same limited supervision.

The Qwen3.5-9B harness improves on the strongest non-harness entry, Qwen3-VL-8B, raising average SRCC/PLCC from $.657/.741$ to $.790/.824$. The 397B harness reaches $.813/.842$, achieving the best average SRCC and PLCC and the best score on six of the seven datasets. On AGIQA-3K, Qwen3-VL-8B retains a narrow PLCC edge ($.794$ vs. $.786$), although the 397B harness leads in SRCC ($.710$ vs. $.655$), indicating stronger agreement with the human quality ranking on this dataset.

## 5 Conclusion

We presented **Q-Harness**, a harness system that turns a frozen vision-language model into a few-shot no-reference image-quality assessor by eliciting per-attribute subscores and mapping them to MOS through a lightweight ridge calibration layer. The broader insight is that for perceptual quality, the limiting factors are no longer label availability or the need to train a backbone. The remaining challenge is to calibrate the perception that a general-purpose model already possesses. Q-Harness instantiates this insight: it outperforms per-dataset fine-tuning across all tested conditions while requiring only 100–200 labels, no backbone training, and no domain-specific engineering. The controlled ablation confirms that the gain is genuinely semantic rather than a statistical artifact of repeated querying, and the scaling study demonstrates that the same harness interface benefits monotonically from stronger VLM backbones without any change to the training recipe. Several limitations point to future work: the calibration layer is fitted on a single domain at a time, though cross-dataset transfer already performs well. The elicited schema is fixed rather than discovered, and we study only images. Extending the framework to video would require temporal attributes.

### AI USE STATEMENT

In this work, we used generative AI tools for aiding and polishing the writing. We have not used generative AI tools for research ideation, experiment design, or result interpretation, and generating synthetic datasets and proving mathematical claims are not applicable to this work. Additionally, we used generative AI tools for editing the paper to improve readability. We have reviewed all AI-assisted work. We checked the AI-assisted edits for accuracy and consistency with the intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES

Alexandre G. Ciancio, André Luiz N. Targino da Costa, Eduardo A. B. da Silva, Amir Said, Ramin Samadani, and Pere Obrador. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on Image Processing*, 20(1):64–75, 2011. doi: 10.1109/TIP.2010.2053549.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3677–3686, 2020. doi: 10.1109/CVPR42600.2020.00373.

Deepti Ghadiyaram and Alan C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2016. doi: 10.1109/TIP.2015.2500021.

Vlad Hosu, Hanhe Lin, Tamás Szirányi, 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. doi: 10.1109/TIP.2020.2967829.

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 (ICCV)*, pp. 5148–5157, 2021. doi: 10.1109/ICCV48922.2021.00510.

Eric C. Larson and Damon M. Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010. doi: 10.1117/1.3267105.

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*, 34(8):6833–6846, 2024. doi: 10.1109/TCSVT.2023.3319020.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10k: A large-scale artificially distorted IQA database. In *Proceedings of the 11th International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3. IEEE, 2019. doi: 10.1109/QoMEX.2019.8743252.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2304.08485.

Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`, September 2024.

Anish Mittal, Anush Krishna Moorthy, and Alan C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

Guanyi Qin, Runze Hu, Yutao Liu, Xiawu Zheng, Haotian Liu, Xiu Li, and Yan Zhang. Data-efficient image quality assessment with attention-panel decoder. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2091–2100, 2023. doi: 10.1609/aaai.v37i2.25302.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3667–3676, 2020. doi: 10.1109/CVPR42600.2020.00372.

Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018. doi: 10.1109/TIP.2018.2831899.

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, 2023. doi: 10.1609/aaai.v37i2.25353. arXiv:2207.12396.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-Bench: A benchmark for general-purpose foundation models on low-level vision. In *International Conference on Learning Representations (ICLR)*, 2024a. arXiv:2309.14181.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, Qiong Yan, and Weisi Lin. Q-Instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25490–25500, 2024b. doi: 10.1109/CVPR52733.2024.02408.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, 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 (ICML)*, 2024c. arXiv:2312.17090.

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 (CVPRW)*, pp. 1191–1200, 2022. doi: 10.1109/CVPRW56347.2022.00126.

Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2020. doi: 10.1109/TCSVT.2018.2886771.

Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021. doi: 10.1109/TIP.2021.3061932.

Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang. Adaptive image quality assessment via teaching large multimodal model to compare. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.19298.