Modeling Beyond MOS: Quality Assessment Models Must Integrate Context, Reasoning, and Multimodality

Anonymous Author(s)
Affiliation
Address
email

Abstract

This position paper argues that Mean Opinion Score (MOS), while historically foundational, is no longer sufficient as the sole supervisory signal for multimedia quality assessment models. MOS reduces rich, context-sensitive human judgments to a single scalar, obscuring semantic failures, user intent, and the rationale behind quality decisions. We contend that modern quality assessment models must integrate three interdependent capabilities: (1) context-awareness, to adapt evaluations to task-specific goals and viewing conditions; (2) reasoning, to produce interpretable, evidence-grounded justifications for quality judgments; and (3) multimodality, to align perceptual and semantic cues using vision–language models. We critique the limitations of current MOS-centric benchmarks and propose a roadmap for reform: richer datasets with contextual metadata and expert rationales, and new evaluation metrics that assess semantic alignment, reasoning fidelity, and contextual sensitivity. By reframing quality assessment as a contextual, explainable, and multimodal modeling task, we aim to catalyze a shift toward more robust, human-aligned, and trustworthy evaluation systems.

1 Introduction

Multimedia Quality Assessment (MQA) [1] spans a wide range of tasks, including Image Quality Assessment (IQA) [2], Video Quality Assessment (VQA) [3], Audio Quality Assessment (AQA) [1], and Aesthetic Quality Assessment [4]. These models are foundational to applications such as video streaming [5, 6], teleconferencing [7, 8], and social media delivery [9], where perceptual quality must be optimized under bandwidth and hardware constraints. But they are equally critical in high-stakes domains such as autonomous driving [10], medical imaging [11], and immersive AR/VR environments [12], where even subtle degradations can compromise safety, diagnostic accuracy, or user immersion.

Despite the diversity and complexity of these applications, most MQA systems remain anchored to a single scalar: the Mean Opinion Score (MOS) [13]. Introduced for its simplicity and ease of deployment, MOS has become the default supervisory signal for training and evaluating quality assessment models [2, 14]. It is deeply embedded in dataset design, benchmarking protocols, and learning objectives across IQA, VQA, and related fields.

This paper takes the position that MOS, while historically foundational, is no longer sufficient to guide the development of modern quality assessment models. MOS flattens rich, context-sensitive human judgments into a single number. It cannot explain why a piece of content is judged as high or low quality, nor can it adapt to the specific goals, environments, or user intents that define real-world use cases. This limitation is especially acute in no-reference (NR) settings, where models must assess quality without access to a pristine reference and where semantic failures, contextual mismatches, or task-specific degradations often go undetected.

Consider a few examples: a novice Twitch streamer misconfigures their camera, resulting in poor framing and lighting; a radiological scan contains subtle artifacts that could mislead diagnosis; a generative image appears sharp but contains hallucinated text or broken structure. In each case, a scalar MOS score is not only uninformative it is

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

misleading. What is needed is a modeling framework that reflects how humans actually perceive, reason about, and act on quality.

We argue that quality assessment models must move beyond scalar prediction and integrate three interdependent capabilities:

- **Context-awareness**: to adapt evaluations to specific tasks, user goals, and environmental conditions.
- **Reasoning**: to produce interpretable, evidence-grounded justifications for quality judgments.
- **Multimodality**: to align perceptual and semantic cues using vision–language models and other cross-modal signals.

These pillars are not optional enhancements they are necessary foundations for building quality assessment systems that are robust, trustworthy, and aligned with real-world needs. By reframing quality assessment as a contextual, explainable, and multimodal modeling task, we aim to catalyze a shift in how the field defines, supervises, and evaluates perceptual quality.

The remainder of this paper is structured as follows: Section 2 provides historical context and motivation. Section 3 critiques the limitations of MOS and correlation-based evaluation. Section 4 introduces our proposed modeling paradigm. Section 5 outlines a roadmap for benchmark and metric reform. Section 6 addresses alternative viewpoints, and Section 7 concludes with a call to action.

2 Background and Motivation

The foundations of multimedia quality assessment (MQA) lie in psychophysical research from the 19th century, where early work by Fechner and others introduced methods to quantify human perception of sensory stimuli such as brightness, contrast, and sharpness [15, 16]. Concepts like just-noticeable differences (JNDs) [17], rating scales, and paired comparisons [18] laid the groundwork for structured perceptual evaluation. Thurstone's law of comparative judgment [19] formalized the derivation of interval scales from subjective comparisons, providing a theoretical basis for perceptual modeling.

As multimedia systems evolved, the need for standardized evaluation protocols became critical. Organizations such as the EBU, ITU, and VQEG introduced formal subjective testing methodologies. Standards like ITU-R BT.500 [20] and ITU-T P.910 [13] institutionalized the Mean Opinion Score (MOS), a scalar average of human ratings, as the dominant proxy for perceived quality [21]. MOS has since become the default supervisory signal for computational models and the foundation of most benchmark datasets.

While MOS has enabled decades of progress, it imposes a reductive view of human judgment. It collapses rich, context-sensitive evaluations into a single scalar, discarding semantic nuance, decision rationale, and task relevance. Moreover, subjective testing is resource-intensive, requiring controlled environments, large participant pools, and significant time and cost. These limitations have driven the development of automated quality assessment models, typically categorized by their reliance on reference content:

- **Full-Reference (FR)**: Models that compare a distorted signal to a pristine reference (e.g., PSNR [22], SSIM [23], VMAF [24]).
- **Reduced-Reference (RR)**: Models that use partial information from the reference [25, 26].
- **No-Reference (NR)**: Models that operate without any reference, relying solely on the distorted input (e.g., BRISQE [27], NIQE [28], ARNIQA [29]).

Deep learning has significantly advanced MQA. CNN-based models such as DB-CNN [30], KonCept512 [31], LPIPS [32], and TPOIQ [33], and transformer-based architectures like MUSIQ [34], Chen et al. [35], and Re-IQA [36] have achieved strong performance in both FR and NR settings. However, these models are almost universally trained to regress MOS values. This scalar supervision constrains their expressiveness, pushing them to optimize for perceptual fidelity while suppressing their capacity to model semantic coherence, contextual relevance, or user intent, even when their architectures are capable of doing so.

In these models, context is not explicitly modeled, it is inherited from the subjective testing protocol of the training dataset. A model trained on a single dataset becomes implicitly conditioned on that dataset's assumptions (e.g., display type, viewing distance, task). When trained on multiple datasets, the model effectively averages across contexts, diluting its ability to adapt to any specific use case. This leads to brittle generalization and a lack of task-awareness in deployment.

Recent work has explored vision–language models (VLMs) such as CLIP [37] and BLIP [38], which jointly embed visual and textual information and offer a path toward semantically grounded quality assessment [39]. Extensions like CLIP-IQA [40], QualiCLIP [41], and Q-Align [42] adapt these models to perceptual quality tasks. However, they are still trained on the same MOS-annotated datasets as traditional models, and thus inherit the same contextual limitations. Moreover, they lack structured reasoning capabilities, which are essential for producing interpretable and task-aware assessments.

Multimodal foundation models such as Flamingo [43], LLaVA [44], and Qwen-VL [45] further expand the potential of quality assessment by integrating vision-language alignment with instruction-following capabilities. These systems can generate natural language rationales, interpret visual content in context, and perform multi-step reasoning. However, recent evaluations [46] show that even state-of-the-art models like GPT-4V struggle to detect fine-grained perceptual artifacts and align with human quality judgments, despite their massive scale.

A particularly promising direction is the emergence of Chain-of-Thought (CoT) prompting and reinforcement learning (RL) fine-tuning [47, 48], which enable models to articulate step-by-step reasoning processes. These techniques have improved performance on tasks requiring multi-step inference and justification, and are increasingly being explored in vision-language contexts [49]. RL-based strategies such as GRPO [50] and PPO [51] further enhance alignment with human-like reasoning and preferences.

Finally, we note that most current models underutilize other perceptually and cognitively relevant modalities such as visual attention maps, saliency cues, and artifact localization masks, which could serve as valuable signals for grounding quality judgments in both perceptual and semantic evidence. These modalities remain largely disconnected from current modeling pipelines and are rarely integrated with language-based reasoning or contextual conditioning.

Together, these developments underscore both the feasibility and the urgency of a paradigm shift. **We argue that multimedia quality assessment must move beyond scalar MOS prediction and embrace a modeling framework that is multimodal, explainable, and context-aware one that reflects the richness of human judgment and the complexity of modern multimedia content.**

3 The Limits of MOS and Correlation-Based Evaluation

While MOS remains the de facto standard for image/video quality benchmarks, its single-scalar label **actively hinders** modern systems. By collapsing human judgments into one number, MOS masks semantic failures, erases context, and trains models to chase averages instead of edge-case critical errors. Below, we explore four fundamental flaws of MOS-based training and evaluation.

3.1 MOS as a Bottleneck for Generalization and Semantics

MOS supervision inherently restricts model outputs to a scalar regression target, eliminating the capacity to encode which attributes drive quality assessments. Liu and Bovik [52] demonstrated that IQA models trained exclusively on MOS labels routinely misclassify semantically broken artifacts (e.g., hallucinated text) as high quality, despite appropriately down-weighting benign degradations (e.g., mild blur). Recent transformer-based and vision–language scorers achieve competitive MOS prediction accuracy yet fail to preserve interpretability or object-level reasoning [42].

Also, all contextual metadata like display type, viewing distance, ambient lighting, intended downstream task is baked into the original subjective test protocol (e.g. ITU-T P.910) rather than exposed to the model as features [13]. Consequently, a model trained on a single MOS dataset implicitly adopts its lab's assumptions. When you train across multiple datasets, the model effectively averages over conflicting contexts: in [53], cross-dataset SROCC drops by 25% on no-reference benchmarks.

In no-reference (NR) settings, the absence of any pristine reference further amplifies this limitation: models must infer quality solely from semantic cues, which hinders their ability to distinguish task-critical distortions from benign variations. Zhu et al. [54] report that conventional NR metrics such as VMAF [24] systematically underrate visually sharp but semantically empty gameplay frames, whereas a CLIP-guided NR model that dynamically reweights feature contributions based on scene context yields a 0.15 increase in Spearman's correlation on an out-of-domain "in-the-wild" games dataset. These findings underscore that MOS-based supervision does not encode content relevance, unless implicitly specified during data collection.

3.2 Correlation Does Not Imply Understanding

Contemporary quality-assessment benchmarks judge models almost exclusively by their Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC), or Root Mean Squared Error (RMSE) with respect to averaged MOS labels. These metrics quantify statistical alignment with mean human scores but provide no guarantee that a model captures the mechanisms of human perception or the semantics underlying quality judgments [55], [56]. High correlation thus reflects only that a model has learned to match dataset-wide averages, not that it truly "understands" why one distortion is more objectionable than another [57]. Empirical cross-dataset evaluations reveal this gap starkly. Saha et al. (Re-IQA) demonstrate that leading no-reference IQA methods, achieving SROCC > 0.95 on synthetic distortions suffer a 30% drop in SROCC when tested on authentic "in-the-wild" images (e.g., KonIQ-10k, CLIVE) [36]. Yang et al. [53]further report clear performance degradation of CNN-based NR-IQA models in cross-dataset settings, despite strong in-domain correlations. Hosu et al. confirm that models with $PLCC \simeq 0.92$ on curated datasets fall below $SROCC = 0.70$ on KonIQ-10k's authentic distortions [58]. In the video domain, Li et al. [59] show that VQA algorithms tuned for lab-style degradations fail to generalize to in-the-wild shooting conditions, incurring substantial correlation losses. Ghadiyaram & Bovik [60] likewise find that models trained on synthetic distortions poorly predict quality on real-world mixtures of distortions.

Beyond cross-domain brittleness, correlation-driven training inherently biases models toward the dataset mean, suppressing uncertainty and down-weighting rare but critical distortions. As a result, models optimized for PLCC / SROCC miss high-impact artifacts, semantic anomalies, edge case degradations, and context-sensitive errors, because they are in the tails of the MOS distribution [61]. Recent works like STNS-IQA demonstrate that incorporating scene-aware statistical features and multimodal cues (e.g., CLIP embeddings) recovers over 0.20 SROCC in GAN-generated images where baseline NR metrics fail [62]. Similarly, IE-IQA integrates intelligibility features to improve robustness across contexts, highlighting the necessity of moving beyond pure correlation objectives [63].

These findings confirm that high correlation with MOS benchmarks does not imply human-aligned understanding or robust perceptual fidelity. We therefore advocate for augmenting existing evaluation protocols with :

1. **Semantic stress tests** that probe object, and scene level failure.
2. **Uncertainty calibration metrics** to reward explicit confidence modeling
3. **Context-aware benchmarks** that expose models to a diversity of viewing conditions and user intents

Only by expanding beyond scalar correlation can we ensure quality-assessment systems that genuinely reflect human judgments.

3.3 Structural Failures and the Collapse of Disagreement

MOS reduces inter-subject and intra-subject variability signals of perceptual ambiguity and cultural diversity into a single scalar, thereby discarding information about rating spread and consensus strength [64]. In the KonIQ-10k dataset [58], per-image rating standard deviations range from 0.6 to 1.5 on a five-point scale, correlating with semantic complexity and demonstrating that images with identical MOS can evoke vastly different agreements among observers. The BIQ2021 database similarly reports that images sharing the same mean score can exhibit variance differences exceeding 1.0, indicating fundamentally distinct perceptual gray zones that a single MOS cannot capture [65].

Large-scale crowdsourced experiments on the LIVE-itW database reveal inter-subject variance up to 20 points on a 100-point scale under uncontrolled viewing conditions, yet MOS aggregation obscures these environmental sensitivities [66]. By averaging over these distributions, MOS-trained models become oblivious to borderline or divisive content and cannot flag stimuli that straddle acceptability thresholds for different user groups [67]. The Konstanz Natural Video Database (KoNViD-1k)[68] shows that videos with similar MOS can have rating variances driven by motion ambiguity, yet standard VQA models treat them as uniform quality. This collapse of disagreement limits personalization and accessibility, as models lack the capacity to adjust to individual or cultural preferences.

Furthermore, MOS-centric training underrepresents rare but critical artifacts: generative-image distortions such as semantic misalignments and text hallucinations are virtually absent in legacy datasets, causing no-reference IQA models to assign high quality to AI-generated images despite glaring semantic errors somerecent works like [69] are trying to adress this gap. Similar blind spots occur for short-form video distortions: e.g. frame freezing and ghosting in KVQ, where novel artifacts provoke high disagreement but remain unnoticed when averaged to MOS [70]. Multi-angle video assessments also reveal that wide-angle distortions yield high inter-subject variance, yet MOS

aggregation masks these differences [71]. Even in laboratory-controlled NITS-IQA experiments, significant rating spreads persist under uniform conditions, underscoring that disagreement is not noise but essential information [72].

Addressing this collapse requires preserving and modeling rating distributions via histograms, confidence intervals, or uncertainty estimates, rather than reducing them to a single mean. Only by capturing disagreement can quality assessment models identify divisive content, support personalization, and remain robust to novel, context-sensitive artifacts.

3.4 MOS Discourages Interpretability and Explainability

Mean Opinion Score supervision provides only a scalar target, offering neither the rationale behind a quality judgment nor any spatial or semantic localization of artifacts, especially in NR settings. As a result, MOS-trained models function as black boxes: there is no mechanism to trace *why* a given image or video frame is assigned a specific score. This opacity is untenable in high-stakes domains such as medical imaging or autonomous driving, where understanding model decisions is essential for trust and safety.

Explainable IQA methods illustrate the alternative: Kazemi Ranjbar and Fatemizadeh's ExIQA framework [73] predicts distortion types and strengths using vision–language models and outputs both a quality score and a distortion attribution map. Similarly, recent work on semantic-attribute reasoning in BIQA not only estimates MOS but also produces interpretable feature attributions that highlight object- or scene-level factors influencing quality [74]. In medical imaging, specialized explainability approaches generate high-resolution saliency maps that reveal which anatomical regions drive the quality assessment, enabling radiologists to verify and correct model outputs [75].

By contrast, MOS-only models cannot support human-in-the-loop workflows, adaptive enhancement, or actionable feedback capabilities demonstrated by explainable IQA systems. Without structured explanations, developers cannot debug failure modes, regulators cannot audit compliance, and end users cannot gain confidence in model reliability. We therefore argue that future supervisory signals must integrate interpretability objectives forcing models to justify their scores with localized, evidence-grounded explanations. Recent emerging reasoning CoT in reasoning models like DeepSeek-R1 [47] could also generate a type of interpretability that is more organic to human understanding instead of the post-hoc mechanistic explainability methods used in other approaches.

3.5 MOS Trains Models to Be Perceptually Myopic

Models trained solely to minimize MOS regression losses disproportionately optimize for statistically frequent, low-level distortions blur, noise, blocking while remaining blind to rare distortions, semantically catastrophic or contextually unacceptable anomalies. For instance, analyses on the PIPAL[76] dataset reveal that state-of-the-art IQA metrics (e.g. SSIM, LPIPS) achieve only moderate Spearman correlations on GAN-generated outputs, failing to penalize semantic misalignments that humans deem unacceptable. Similarly, MANIQA demonstrates that NR-IQA networks must be explicitly tailored to GAN distortions to recover just a fraction of their performance on synthetic benchmarks [77].

Hallucinated or implausible content, such as extra limbs or nonsensical text in generative images, can nonetheless receive high scores from MOS-based predictors. The AGIQA-3K [69] study shows that off-the-shelf NR-IQA models correlate poorly with human judgments of text-to-image alignment and semantic fidelity, whereas dedicated semantic metrics substantially improve assessment on text-guided generative outputs. Large multimodal models further confirm this gap: conventional DNN-based IQA methods underperform on AI-generated images with semantic errors, while CLIP-guided architectures regain over 0.20 SROCC by incorporating semantic embeddings [78].

In the video domain, MOS-centric VQA also neglects temporal artifacts and rare but perceptually salient events. The PEA265 database documents six common compression artifacts, including flickering and floating—that standard VQA algorithms fail to detect without bespoke temporal modeling [79]. Saliency-Aware Spatio-Temporal Artifact Measurement (SSTAM) further shows that no-reference VQA methods overlook local flicker unless explicitly trained on artifact annotations, resulting in large drops in correlation with MOS on flicker-rich sequences [80].

These patterns of failure underscore that MOS training instills a form of perceptual myopia: models learn to reward smoothness over structure, sharpness over meaning, and frequency over significance. Emerging approaches, such as IE-IQA's integration of intelligibility features [63], CLIP-AGIQA's exploit of visual–textual alignment [81], and prompt-aware IQA frameworks [82] demonstrate the necessity of multimodal, semantics-driven supervision. To overcome perceptual myopia, future benchmarks must incorporate semantic anomalies, context-aware distortions, and uncertainty modeling rather than rely exclusively on MOS correlation.

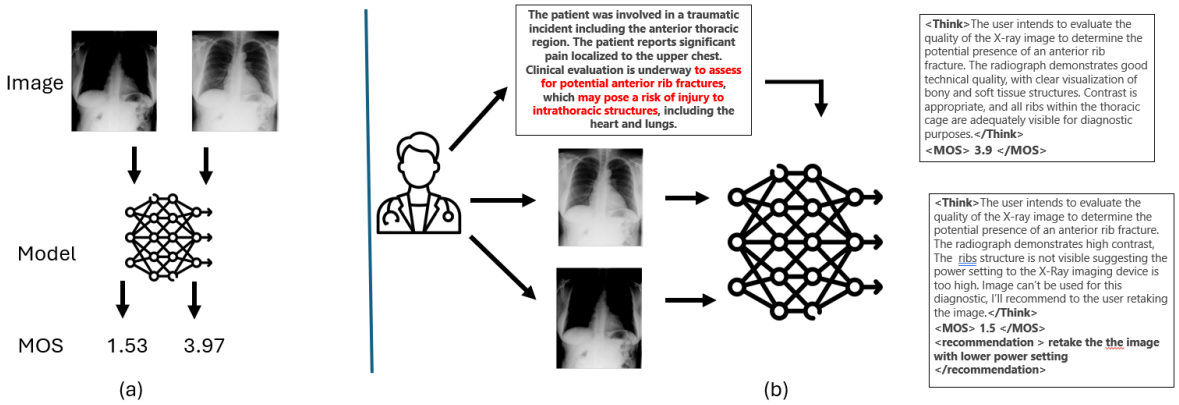

Figure 1: An illustrative example comparing the a reduced pipeline from our paradigm (b) with a traditional MOS quality assessment system (a) in a clinical setting. By providing the context to the model, it can focus on estimating the quality for a specific task, as well as give recommendations, all while the reasoning behind the decision and recommendation stay transparent through the thinking CoT.

4 Modeling Beyond MOS: Toward Contextual, Explainable, and Multimodal Quality Assessment

To overcome the structural limitations of MOS-based supervision, we argue for a modeling paradigm that is explicitly **context-aware**, **reasoning-capable**, and **multimodal**. These are not optional enhancements, they are necessary foundations for building quality assessment systems that are robust, interpretable, and aligned with real-world use cases beyond streaming and compression. In this section, we outline how each of these pillars addresses specific failure modes identified in Section 3, and how they can be operationalized in model design.

4.1 Context-Aware Modeling

MOS-trained models inherit context implicitly from the protocols under which subjective scores were collected, but they lack mechanisms to reason about task, user intent, or viewing conditions [13]. A truly context-aware model must accept explicit metadata inputs such as task type (e.g., medical diagnosis vs. social streaming) or device parameters (e.g., mobile display vs. VR headset) and condition its quality predictions accordingly [83]. MetaIQA employs meta-learning to adapt NR-IQA models to unknown distortions by leveraging task-level priors, demonstrating substantial cross-domain gains when context shifts [83]. Semantic-guided NR-IQA (SGIQA) integrates scene-category prompts into a Swin Transformer backbone, yielding a 0.07 SROCC lift over context-agnostic counterparts on in-the-wild datasets [84].

Context-awareness enables models to distinguish between acceptable and unacceptable degradations based on use case. A compression artifact that is tolerable in a social media clip may be unacceptable in a radiological scan. Without explicit context, models collapse these distinctions, a failure mode that cannot be corrected by MOS regression alone.

4.2 Reasoning-Centric Quality Assessment

MOS supervision discourages interpretability. To reverse this, we advocate for models that generate **explicit chains of reasoning** structured, step-by-step justifications for their quality judgments. This can be achieved through Chain-of-Thought (CoT) prompting, vision-language entailment modules, or reinforcement learning (RL) fine-tuning with reasoning-aligned objectives.

Each reasoning step should be grounded in visual evidence and validated for factual consistency. For example, if a model claims "the license plate is unreadable," it must attend to the relevant region and verify that the claim is supported by the image. This closes the loop between *where* the model looks and *why* it makes a decision. Reasoning also enables models to expose internal conflicts such as high technical fidelity but low semantic integrity, which scalar scores obscure.

Recent advances in multimodal CoT generation [47] and RL-based alignment [50] demonstrate that such reasoning is not only feasible but improves robustness and human trust. In quality assessment, it enables models to surface actionable insights, support human-in-the-loop workflows, and justify decisions in high-stakes domains.

4.3 Multimodal Fusion and Alignment

Human quality judgments are inherently multimodal: they arise from the interplay of visual perception, semantic understanding, and contextual intent. Yet most models reduce this process to a single perceptual embedding. We argue that quality assessment must integrate at least four complementary modalities:

1. **Visual features**: raw pixels or deep perceptual embeddings.
2. **Textual context**: task descriptions, user roles, or usage scenarios.
3. **Visual attention cues**: saliency maps, gaze data, or learned attention heatmaps , [85], [86], [87].
4. **Artifact maps**: localized distortion masks or hallucination flags.

These modalities should be embedded into a joint representation space using a vision-language backbone augmented with attention-fusion layers. This enables a series of alignment checks:

- **Image-attention alignment**: ensures that high-salience regions coincide with perceptually degraded or semantically critical areas.
- **Attention-text alignment**: verifies that textual rationales refer to regions that attracted attention.
- **Artifact-text alignment**: confirms that semantic failure flags correspond to actual distortion masks.

Such alignment is not a luxury, it is a prerequisite for trustworthy, human-aligned quality assessment. Without it, models hallucinate explanations, ignore critical regions, or reward perceptual smoothness over semantic fidelity.

4.4 From Scalar Prediction to Structured Judgment

Together, these modeling principles redefine quality assessment as a structured prediction task instead of a scalar regression problem. A model should output:

- A contextualized quality score (conditioned on task and user intent).
- A natural language rationale (explaining the judgment).
- A visual attention map (showing where the model looked).
- An artifact mask (highlighting localized distortions).

This structured output enables richer supervision, more informative evaluation, and actionable feedback. It also aligns with emerging trends in multimodal AI, where interpretability, grounding, and task adaptation are no longer optional.

5 Reforming Benchmarks for Contextual and Explainable Quality Assessment

To align evaluation with the context- and explainability-driven paradigm outlined in Section 4, benchmarks must shed their dependence on legacy lab-based MOS scores and instead integrate explicit task metadata, persona-conditioned judgments, and rationale annotations. We next detail four targeted reforms, spanning data collection, annotation protocols, simulation, and multi-facet metrics—that realize this shift without expanding benchmark size.

5.1 Contextual and Multi-Perspective Annotation Protocols

To capture the full complexity of real-world quality judgments, benchmarks must go beyond single-condition Opinion Score ratings and incorporate diverse perspectives and explanatory signals:

- **Multi-condition labeling**: Collect separate quality scores for each predefined scenario (e.g., “clinical diagnosis on MRI,” “social media upload on smartphone,” “immersive viewing in VR”), with metadata for task, device, and environment, and explanations to subject before experience.

- **Role-specific ratings**: Assign annotators explicit personas (e.g., radiologist, video editor, end-user) so that perceptual criteria and resulting label distributions, reflect varied expertise and priorities, this can go even further through emotion detection by facial expression[88] and micro-expression [89].
- **Structured rationales**: Require a concise justification for each score, such as a chain-of-thought outline or attribute checklist ("text legibility," "artifact severity")—on a representative subset of items.
- **Semantic-error flags**: Beyond quality scores, annotators must mark semantic failures (e.g., hallucinated objects, missing critical details) that MOS alone cannot expose.
- **Distributional labels**: Preserve the full histogram or confidence interval of ratings per condition and persona rather than a single mean, enabling models to learn from disagreement and quantify uncertainty,as well rare cases.

5.2 Scalable and Targeted Data Collection Infrastructure

Deploying the annotation protocols outlined above at scale necessitates specialized infrastructure for diverse, high-quality labels:

- **Contextual crowdsourcing platforms**: Integrate scenario scripts and persona prompts into task interfaces, ensuring raters understand the specific viewing conditions and user roles before annotation [90].
- **In-the-wild mobile annotation apps**: Embed rating, rationale entry, and declared intent (e.g., "uploading to Instagram," "diagnosing an X-ray") into consumer-facing apps, capturing authentic user judgments in situ.
- **Active learning pipelines**: Use preliminary models to identify semantic edge cases or images with high predicted quality but low MOS, and prioritize these for expert re-annotation, thereby improving label efficiency by up to 40% [91].
- **Gamified annotation interfaces**: Employ game mechanics, such as role-based challenges ("spot the compression artifact as a radiologist") and point systems—to boost annotator engagement and depth of rationale, which has been shown to increase both speed and accuracy in image description tasks [92].
- **Longitudinal expert panels**: Convene panels of domain experts to re-rate a subset of content at regular intervals, revealing perceptual drift and evolving standards; the LEAD methodology achieved inter-rater reliability > 0.85 across repeated conditions [93].

5.3 Simulation and Evaluation with Persona-Conditioned Models

With context- and persona-rich datasets, benchmarks can support simulation-based evaluation:

- **Conditional quality generators**: A reasoning generative model takes an image and a persona-context embedding and outputs an "opinion overlay", an explanation of perceived flaws or strengths.
- **Multi-agent rating simulation**: Multiple persona agents (e.g., radiologist, streamer, casual viewer) each generate a quality score and rationale using their own reasoning policy. This yields a distribution of judgments and explanations [94].
- **Reasoning Metrics**: Introducing recent metrics that evaluate coherence, validity, groundness and factuality [95], and utility of reasoning in LLMs and Multimodal LLMs [96] must be used as training and evaluation objectives alongside perceptual losses.
- **Few-shot persona adaptation**: A lightweight meta-learner adapts a model's scoring head to new personas or tasks with minimal supervision, enabling flexible deployment across domains [94].

5.4 Toward Scientifically Grounded Evaluation Platforms

These reforms transform benchmarks from static MOS collections into scientifically grounded, multi-dimensional evaluation platforms. They support:

- **Context-aware evaluation**: Models are tested under varying task and user conditions.
- **Explainability assessment**: Rationales and attention maps are evaluated for coherence and grounding.
- **Semantic robustness**: Models are challenged with hallucinations, rare artifacts, and ambiguous content.
- **Personalization and fairness**: Rating distributions and persona diversity are preserved and modeled.

These are not speculative enhancements, they are necessary conditions for evaluating the next generation of quality assessment systems, and some preliminary works are being conducted in some aspects. Without them, we cannot measure what matters: whether a model understands quality, adapts to context, and explains its decisions, and by extension accurately model the human visual system behavior.

6 Alternative Viewpoints and Rebuttals

"MOS is sufficient." Mean Opinion Score (MOS) has indeed powered decades of quality assessment, offering a single, interpretable metric collected under standardized protocols (e.g., ITU-T P.910) [13]. Its simplicity enabled rapid dataset creation and straightforward benchmarking[31], [97]. However, this very simplicity glosses over context-dependent semantics, user intent, and the reasoning behind judgments as shown in section 3. We do not discard MOS; rather, we retain it as a baseline while augmenting benchmarks with context-conditioned scores, semantic-failure flags, and reasoning metrics. In constrained scenarios, pure MOS suffices, but in complex or safety-critical domains like medical imaging [11] or autonomous driving[98], additional axes are essential to surface failures that MOS alone conceals.

"Explainability isn't necessary." It is argued that users only need a numerical score, not a chain of thought. Yet extensive work in explainable AI demonstrates that transparency is indispensable when trust and accountability matter [99], [100], [101]. In medical workflows, radiologists require saliency maps and rationales to validate automated assessments [102]. In content-creation pipelines, interpretable feedback accelerates debugging and user learning [103]. Without explicit explanations, models remain opaque black boxes, hindering regulatory approval and end-user confidence.

"Computational complexity concerns." Integrating multimodal reasoning increases system footprint. Yet modern architectures achieve efficient conditional computation, Mixture-of-Experts (MoE) layers route only relevant tokens through heavy subnets [104] leading to substantial savings in computational power exceeding 50%. Adapter modules permit on-demand reasoning at only a fraction of full-model cost [105]. We advocate tiered pipelines: a MOS-only fast path for low-risk use, and a full reasoning path for high-stakes tasks.

"Subjectivity makes standardization impossible." Human judgments vary by culture, expertise, and intent, yet fields such as usability and human–computer interaction routinely formalize subjective assessments into repeatable protocols like ISO9241 [1]. By explicitly modeling personas and contexts, we embrace rather than erase diversity. Persona-conditioned benchmarks reveal demographic sensitivities and enable fairness analyses . Multi-agent evaluation frameworks from dialogue systems illustrate how divergent user profiles can be jointly modeled and simulated [94]. Rather than a barrier, subjectivity becomes an asset for creating robust, personalized quality-assessment systems.

7 Conclusion & Call to Action

MOS alone can no longer serve as the sole supervisory signal for multimedia quality assessment; it must be complemented by explicit *context-awareness*, structured *reasoning*, and integrated *multimodality*. By flattening nuanced, task-dependent human judgments into a single number, MOS-centric approaches obscure semantic failures, disregard user intent and viewing conditions, and provide no insight into the reasoning behind quality decisions. In this paper, we have illustrated these shortcomings with concrete examples from previous studies, outlined a blueprint for richer benchmarks including context-conditioned labels, persona-driven ratings, rationale annotations, semantic-failure flags, and rating distributions and proposed novel evaluation metrics that examine "why," "where," and "when" model outputs align with human perception.

The time has come to evolve our benchmarks and modeling paradigms. In the near term, we encourage the release of datasets annotated with task and persona metadata, structured rationales, and full rating distributions; the organization of shared challenges that assess semantic robustness, explainability, and context sensitivity alongside traditional correlation metrics; and the incorporation of differentiable reasoning and multimodal alignment objectives into model training. Over the longer horizon, we envision interactive, real-time quality advisors that adapt to individual user needs, collaborative platforms for crowdsourced rationale collection, and interdisciplinary efforts to define ethical standards for context and persona modeling. Embracing this expanded framework will yield multimedia quality-assessment systems that are not only accurate but genuinely human-aligned, transparent, and trustworthy.

[1] https://www.iso.org/standard/77520.html

References

[1] Zahid Akhtar and Tiago H Falk, "Audio-visual multimedia quality assessment: A comprehensive survey," *IEEE access*, vol. 5, pp. 21090–21117, 2017.

[2] Guangtao Zhai and Xiongkuo Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, pp. 1–52, 2020.

[3] Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai, "Perceptual video quality assessment: A survey," *Science China Information Sciences*, vol. 67, no. 11, pp. 211301, 2024.

[4] Yubin Deng, Chen Change Loy, and Xiaoou Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80–106, 2017.

[5] Wei Zhou, Xiongkuo Min, Hong Li, and Qiuping Jiang, "A brief survey on adaptive video streaming quality assessment," *Journal of Visual Communication and Image Representation*, vol. 86, pp. 103526, 2022.

[6] Mohammad Ghasempour, Hadi Amirpour, and Christian Timmerer, "Real-time quality-and energy-aware bitrate ladder construction for live video streaming," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2025.

[7] Zhenqiang Ying, Deepti Ghadiyaram, and Alan Bovik, "Telepresence video quality assessment," in *European Conference on Computer Vision*. Springer, 2022, pp. 327–347.

[8] Fazliaty Edora Fadzli, Ajune Wanis Ismail, and Shafina Abd Karim Ishigaki, "A systematic literature review: Real-time 3d reconstruction method for telepresence system," *Plos one*, vol. 18, no. 11, pp. e0287155, 2023.

[9] Qin-Yu Cai, Jing Tang, Si-Zhe Meng, Yi Sun, Xia Lan, and Tai-Hang Liu, "Quality assessment of videos on social media platforms related to gestational diabetes mellitus in china: A cross-section study," *Heliyon*, vol. 10, no. 7, pp. e29020, 2024.

[10] Keke Geng, Ge Dong, and Wenhan Huang, "Robust dual-modal image quality assessment aware deep learning network for traffic targets detection of autonomous vehicles," *Multimedia Tools and Applications*, vol. 81, no. 5, pp. 6801–6826, 2022.

[11] Marouane Tliba, Aymen Sekhri, Mohamed Amine Kerkouri, and Aladine Chetouani, "Deep-based quality assessment of medical images through domain adaptation," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 3692–3696.

[12] Marouane Tliba, Aladine Chetouani, Giuseppe Valenzise, and Frederic Dufaux, "Representation learning optimization for 3d point cloud quality assessment without reference," 10 2022, pp. 3702–3706.

[13] International Telecommunication Union Telecommunication Standardization Sector (ITU-T), "Recommendation p.910: Subjective video quality assessment methods for multimedia applications," `https://www.itu.int/rec/T-REC-P.910-202310-I/en`, 2023, ITU-T Recommendation P.910 (10/23).

[14] Robert C Streijl, Stefan Winkler, and David S Hands, "Mean opinion score (mos) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[15] Stephanie L Hawkins, "William james, gustav fechner, and early psychophysics," *Frontiers in physiology*, vol. 2, pp. 68, 2011.

[16] Norman Burningham, "A brief review of the history and application of psychometrics and scaling to image quality assessment," in *PICS*, 1999, pp. 169–172.

[17] Melissa K Stern and James H Johnson, "Just noticeable difference," *The corsini encyclopedia of psychology*, pp. 1–2, 2010.

[18] Maria Perez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafał K Mantiuk, "From pairwise comparisons and rating to a unified quality scale," *IEEE Transactions on Image Processing*, vol. 29, pp. 1139–1151, 2019.

[19] Louis L Thurstone, "A law of comparative judgment," in *Scaling*, pp. 81–92. Routledge, 2017.

[20] International Telecommunication Union Radiocommunication Sector (ITU-R), "Recommendation bt.500: Methodology for the subjective assessment of the quality of television pictures," `https://www.itu.int/rec/R-REC-BT.500`, 2000, ITU-R Recommendation BT.500-10.

[21] Pedram Mohammadi, Abbas Ebrahimi-Moghadam, and Shahram Shirani, "Subjective and objective quality assessment of image: A survey," *arXiv preprint arXiv:1406.7799*, 2014.

[22] Z. Kotevski and P. Mitrevski, "Experimental comparison of psnr and ssim metrics for video quality estimation," in *ICT Innovations 2009*, Danco Davcev and Juan M. Gómez, Eds., pp. 205–213. Springer, Berlin, Heidelberg, 2010.

[23] BovikAC WangZhou, HR Sheikh, et al., "Image qualityassessment: From errorvisibilitytostructural similarity," *IEEE Transon ImageProcessing*, vol. 13, no. 4, pp. 600, 2004.

[24] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara, "Toward a practical perceptual video quality metric, 2016," *Dostupno na: http://techblog. netflix. com/2016/06/toward-practical-perceptual-video. html [16.8. 2022.]*, 2016.

[25] Xudong Lv and Z. Jane Wang, "Reduced-reference image quality assessment based on perceptual image hashing," *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 4361–4364, 2009.

[26] Xuanqin Mou, Wufeng Xue, and Lei Zhang, "Reduced reference image quality assessment via sub-image similarity based redundancy measurement," in *Electronic imaging*, 2012.

[27] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[28] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[29] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo, "Arniqa: Learning distortion manifold for image quality assessment," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 189–198.

[30] 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*, vol. 30, pp. 36–47, 2019.

[31] 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*, vol. 29, pp. 4041–4056, 2020.

[32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[33] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin, "Topiq: A top-down approach from semantics to distortions for image quality assessment," *IEEE Transactions on Image Processing*, 2024.

[34] 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*, 2021, pp. 5148–5157.

[35] Manri Cheon, Sung-Jun Yoon, Byungyeon Kang, and Junwoo Lee, "Perceptual image quality assessment with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 433–442.

[36] Avinab Saha, Sandeep Mishra, and Alan C Bovik, "Re-iqa: Unsupervised learning for image quality assessment in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5846–5855.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[38] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12888–12900.

[39] Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi, "Benchmark evaluations, applications, and challenges of large vision language models: A survey," *arXiv preprint arXiv:2501.02189*, vol. 1, 2025.

[40] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy, "Exploring clip for assessing the look and feel of images," in *AAAI*, 2023.

[41] Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini, "Quality-aware image-text alignment for real-world image quality assessment," *arXiv preprint arXiv:2403.11176*, vol. 5, no. 6, 2024.

[42] 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.

[43] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al., "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23716–23736, 2022.

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34892–34916, 2023.

[45] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.

[46] Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang, "A comprehensive study of multimodal large language models for image quality assessment," in *European Conference on Computer Vision*. Springer, 2024, pp. 143–160.

[47] DeepSeek-AI, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," 2025.

[48] Qwen Team, "Qwq-32b: Embracing the power of reinforcement learning," March 2025.

[49] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al., "Vlm-r1: A stable and generalizable r1-style large vision-language model," *arXiv preprint arXiv:2504.07615*, 2025.

[50] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al., "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[51] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[52] Anush Krishna Moorthy and Alan Conrad Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[53] Dan Yang, Veli-Tapani Peltoketo, and Joni-Kristian Kämäräinen, "Cnn-based cross-dataset no-reference image quality assessment," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3913–3921.

[54] Kai Zhu, Vignesh Edithal, Le Zhang, Ilia Blank, and Imran Junejo, "Semantically-aware game image quality assessment," *arXiv preprint arXiv:2505.11724*, 2025.

[55] Sergey Kastryulin, Jamil Zakirov, Denis Prokopenko, and Dmitry V Dylov, "Pytorch image quality: Metrics for image quality assessment," *arXiv preprint arXiv:2208.14818*, 2022.

[56] Dounia Hammou, Yancheng Cai, Pavan Madhusudanarao, Christos G Bampis, and Rafał K Mantiuk, "Do image and video quality metrics model low-level human vision?," *arXiv preprint arXiv:2503.16264*, 2025.

[57] Chengqian Ma, Zhengyi Shi, Zhiqiang Lu, Shenghao Xie, Fei Chao, and Yao Sui, "A survey on image quality assessment: Insights, analysis, and future outlook," *arXiv preprint arXiv:2502.08540*, 2025.

[58] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[59] Dingquan Li, Tingting Jiang, and Ming Jiang, "Quality assessment of in-the-wild videos," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 2351–2359.

[60] Deepti Ghadiyaram and Alan C Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of vision*, vol. 17, no. 1, pp. 32–32, 2017.

[61] Guanglu Dong, Xiangyu Liao, Mingyang Li, Guihuan Guo, and Chao Ren, "Exploring semantic feature discrimination for perceptual image super-resolution and opinion-unaware no-reference image quality assessment," *arXiv preprint arXiv:2503.19295*, 2025.

[62] Yuxuan Yang, Zhichun Lei, and Changlu Li, "No-reference image quality assessment combining swin-transformer and natural scene statistics," *Sensors*, vol. 24, no. 16, pp. 5221, 2024.

[63] Tianshu Song, Leida Li, Hancheng Zhu, and Jiansheng Qian, "Ie-iqa: Intelligibility enriched generalizable no-reference image quality assessment," *Frontiers in Neuroscience*, vol. 15, pp. 739138, 2021.

[64] Deepti Ghadiyaram and Alan C Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2015.

[65] Nisar Ahmed and Shahzad Asif, "Biq2021: a large-scale blind image quality assessment database," *Journal of Electronic Imaging*, vol. 31, no. 5, pp. 053010–053010, 2022.

[66] Nithin C Babu, Vignesh Kannan, and Rajiv Soundararajan, "No reference opinion unaware quality assessment of authentically distorted images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2459–2468.

[67] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik, "Patch-vq:'patching up'the video quality problem," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14019–14029.

[68] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe, "The konstanz natural video database (konvid-1k)," in *2017 Ninth international conference on quality of multimedia experience (QoMEX)*. IEEE, 2017, pp. 1–6.

[69] 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*, vol. 34, no. 8, pp. 6833–6846, 2023.

[70] Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen, "Kvq: Kwai video quality assessment for short-form videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25963–25973.

[71] Bo Hu, Wei Wang, Chunyi Li, Lihuo He, Leida Li, and Xinbo Gao, "A multi-annotated and multi-modal dataset for wide-angle video quality assessment," *arXiv preprint arXiv:2501.12082*, 2025.

[72] Jayesh Ruikar and Saurabh Chaudhury, "Nits-iqa database: a new image quality assessment database," *Sensors*, vol. 23, no. 4, pp. 2279, 2023.

[73] Sepehr Kazemi Ranjbar and Emad Fatemizadeh, "Exiqa: Explainable image quality assessment using distortion attributes," *arXiv e-prints*, pp. arXiv–2409, 2024.

[74] Yipo Huang, Leida Li, Yuzhe Yang, Yaqian Li, and Yandong Guo, "Explainable and generalizable blind image quality assessment via semantic attribute reasoning," *IEEE Transactions on Multimedia*, vol. 25, pp. 7672–7685, 2022.

[75] Narbota Amanova, Jörg Martin, and Clemens Elster, "Explainability for deep learning in mammography image quality assessment," *Machine Learning: Science and Technology*, vol. 3, no. 2, pp. 025015, 2022.

[76] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao, "Pipal: a large-scale image quality assessment dataset for perceptual image restoration," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 633–651.

[77] 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*, 2022, pp. 1191–1200.

[78] 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*, 2024, pp. 7803–7812.

[79] Liqun Lin, Shiqi Yu, Liping Zhou, Weiling Chen, Tiesong Zhao, and Zhou Wang, "Pea265: Perceptual assessment of video compression artifacts," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3898–3910, 2020.

[80] Liqun Lin, Yang Zheng, Weiling Chen, Chengdong Lan, and Tiesong Zhao, "Saliency-aware spatio-temporal artifact detection for compressed video quality assessment," *IEEE Signal Processing Letters*, vol. 30, pp. 693–697, 2023.

[81] Zhenchen Tang, Zichuan Wang, Bo Peng, and Jing Dong, "Clip-agiqa: Boosting the performance of ai-generated image quality assessment with clip," in *International Conference on Pattern Recognition*. Springer, 2025, pp. 48–61.

[82] Bowen Qu, Haohui Li, and Wei Gao, "Bringing textual prompt to ai-generated image quality assessment," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[83] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi, "Metaiqa: Deep meta-learning for no-reference image quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14143–14152.

[84] Linpeng Pan, Xiaozhe Zhang, Fengying Xie, Haopeng Zhang, and Yushan Zheng, "Sgiqa: semantic-guided no-reference image quality assessment," *IEEE Transactions on Broadcasting*, 2024.

[85] Mohamed Amine Kerkouri, Marouane Tliba, Aladine Chetouani, and Alessandro Bruno, "A domain adaptive deep learning solution for scanpath prediction of paintings," in *Proceedings of the 19th International Conference on Content-Based Multimedia Indexing*, 2022, pp. 57–63.

[86] Marouane Tliba, Mohamed Amine Kerkouri, Aladine Chetouani, and Alessandro Bruno, "Self supervised scanpath prediction framework for painting images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1539–1548.

[87] Aladine Chetouani and Leida Li, "On the use of a scanpath predictor and convolutional neural network for blind image quality assessment," *Signal Processing: Image Communication*, vol. 89, pp. 115963, 2020.

[88] Simranjit Singh, Amrik Singh, and Baljinder Kaur, "Emotion detection through facial expressions: A survey of ai-based methods," *IJSAT-International Journal on Science and Technology*, vol. 16, no. 1.

[89] Bikash Kumar Jha, Bharat Paudel, Adarsh Mishra, Aabik Maharjan, and Pralhad Chapagain, "Human emotion detection and face recognition system," *International Journal on Engineering Technology*, vol. 2, no. 2, pp. 90–97, 2025.

[90] Jun Xu, Weisi Lin, and Leida Zhang, "A subjective study of image quality assessment metrics using crowdsourcing," *IEEE Transactions on Image Processing*, vol. 30, pp. 1788–1800, 2021.

[91] Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura, "A survey on deep active learning: Recent advances and new frontiers," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[92] Tomislav Ivanjko, "Crowdsourcing image descriptions using gamification: a comparison between game-generated labels and professional descriptors," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2019, pp. 537–541.

[93] Veerle C. Eijsbroek, Katarina Kjell, H. Andrew Schwartz, Jan R. Boehnke, Eiko I. Fried, Daniel N. Klein, Peik Gustafsson, Isabelle Augenstein, Patrick M.M. Bossuyt, and Oscar N.E. Kjell, "The leading guideline: Reporting standards for expert panel, best-estimate diagnosis, and longitudinal expert all data (lead) methods," *Comprehensive Psychiatry*, p. 152603, 2025.

[94] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu, "Scaling synthetic data creation with 1,000,000,000 personas," 2025.

[95] Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong, "QAFactEval: Improved QA-based factual consistency evaluation for summarization," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, Eds., Seattle, United States, July 2022, pp. 2587–2601, Association for Computational Linguistics.

[96] Jinu Lee and J. Hockenmaier, "Evaluating step-by-step reasoning traces: A survey," *ArXiv*, vol. abs/2502.12289, 2025.

[97] Nikolay Ponomarenko, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Lina Jin, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo, "Color image database tid2013: Peculiarities and preliminary results," in *European Workshop on Visual Information Processing (EUVIP)*, 2013, pp. 106–111.

[98] Ce Zhang and Azim Eskandarian, "Image-guided outdoor lidar perception quality assessment for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, pp. 1–12, 2024.

[99] Naeem Ullah, Javed Ali Khan, Ivanoe De Falco, and Giovanna Sannino, "Explainable artificial intelligence: Importance, use domains, stages, output shapes, and challenges," *ACM Comput. Surv.*, vol. 57, no. 4, Dec. 2024.

[100] J. M. Brandenburg, B. P. Müller-Stich, M. Wagner, et al., "Can surgeons trust ai? perspectives on machine learning in surgery and the importance of explainable artificial intelligence (xai)," *Langenbeck's Archives of Surgery*, vol. 410, pp. 53, 2025.

[101] N. Balasubramaniam, M. Kauppinen, K. Hiekkanen, and S. Kujala, "Transparency and explainability of ai systems: Ethical guidelines in practice," in *Requirements Engineering: Foundation for Software Quality*, Vincenzo Gervasi and Andreas Vogelsang, Eds., vol. 13216 of *Lecture Notes in Computer Science*. Springer, Cham, 2022.

[102] Aymen Sekhri, Mohamed A Kerkouri, Aladine Chetouani, Marouane Tliba, Yassine Nasser, Rachid Jennane, and Alessandro Bruno, "Automatic diagnosis of knee osteoarthritis severity using swin transformer," in *Proceedings of the 20th International Conference on Content-Based Multimedia Indexing*, New York, NY, USA, 2023, CBMI '23, p. 41–47, Association for Computing Machinery.

[103] P. Engel-Hermann and A. Skulmowski, "Appealing, but misleading: a warning against a naive ai realism," *AI Ethics*, 2024.

[104] Saeed Masoudnia and Reza Ebrahimpour, "Mixture of experts: a literature survey," *Artificial Intelligence Review*, vol. 42, pp. 275–293, 2014.

[105] Edward J. Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shawn Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022, vol. 1, p. 3.