Beyond Cosine Similarity: Magnitude-Aware CLIP for No-Reference Image Quality Assessment

Zhicheng Liao[1], Dongxu Wu[1], Zhenshan Shi[1], Sijie Mai[1], Hanwei Zhu[2], Lingyu Zhu[3], Yuncheng Jiang[1], Baoliang Chen[1]*

[1]School of Computer Science, South China Normal University, China
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]School of Computer Science, City University of Hong Kong, China
zcliao@m.scnu.edu.cn, blchen@m.scnu.edu.cn

Abstract

Recent efforts have repurposed the Contrastive Language-Image Pre-training (CLIP) model for No-Reference Image Quality Assessment (NR-IQA) by measuring the cosine similarity between the image embedding and textual prompts such as "a good photo" or "a bad photo." However, this semantic similarity overlooks a critical yet underexplored cue: ***the magnitude of the CLIP image features, which we empirically find to exhibit a strong correlation with perceptual quality.*** In this work, we introduce a novel adaptive fusion framework that complements cosine similarity with a magnitude-aware quality cue. Specifically, we first extract the absolute CLIP image features and apply a Box-Cox transformation to statistically normalize the feature distribution and mitigate semantic sensitivity. The resulting scalar summary serves as a semantically-normalized auxiliary cue that complements cosine-based prompt matching. To integrate both cues effectively, we further design a confidence-guided fusion scheme that adaptively weighs each term according to its relative strength. Extensive experiments on multiple benchmark IQA datasets demonstrate that our method consistently outperforms standard CLIP-based IQA and state-of-the-art baselines, ***without any task-specific training***.

Code — https://github.com/zhix000/MA-CLIP

Introduction

Image Quality Assessment (IQA) aims to automatically predict the perceptual quality of an image, playing a critical role in a wide range of applications, such as image enhancement, compression, generation, and transmission. In practical scenarios where pristine reference images are unavailable, No-Reference IQA (NR-IQA) becomes particularly essential, as it assesses image quality solely based on the distorted image.

Recently, significant progress has been made in learning-based NR-IQA models. However, most existing approaches rely heavily on supervised training with specific IQA datasets, often overfitting to dataset-specific distortions or content. This generalization bottleneck severely limits their applicability in real-world systems. The emergence of large-scale vision-language models such as CLIP (Radford et al. 2021) offers a promising alternative. Trained on hundreds of millions of image–text pairs, CLIP demonstrates remarkable generalization and semantic understanding in a zero-shot manner. Recent works have adapted CLIP for NR-IQA by leveraging its ability to compute the cosine similarity between image embedding and quality-descriptive textual prompts (*e.g.*, "a good photo" *vs.*"a bad photo") (Wang, Chan, and Loy 2023). This prompt-based CLIP-IQA technique has shown surprising effectiveness on standard IQA benchmarks without any fine-tuning, providing a compelling direction for generic, training-free quality assessment.

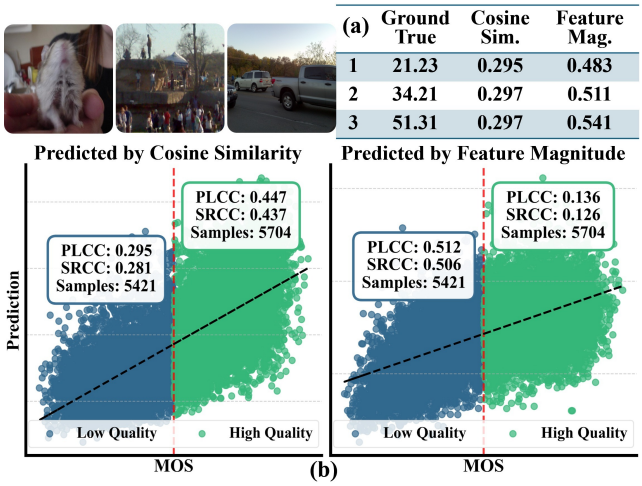


Figure 1: **(a)** Limitations of prompt-based CLIP-IQA: although the images exhibit a wide range of perceptual quality (reflected in their MOS), the cosine similarity between the image embedding and textual prompts remains nearly constant. In contrast, the feature magnitude shows a strong correlation with MOS. **(b)** Complementary behaviors of the two cues across quality levels: As the scatter plots of SPAQ dataset shows that cosine similarity is more reliable in the high-quality region, where semantic features align well with CLIP's pretrained distribution; feature magnitude is more discriminative under low-quality distortions, where semantic alignment breaks down. These observations motivate our dual-cue fusion framework that adaptively integrates both signals for robust quality prediction.

*Corresponding author.

However, CLIP-IQA relies solely on semantic prompt similarity, overlooking another critical yet underexplored signal inherent in the model. Specifically, the cosine similarity computation involves ℓ_2 normalization of the image features, which removes the magnitude information entirely. Through our extensive empirical observations, we find that the magnitude (*i.e.*, norm) of the CLIP image embedding, although ignored in standard usage, is in fact highly indicative of perceptual quality. As illustrated in Fig. 1(a), images with widely varying MOS often yield nearly identical prompt-based similarities, failing to capture true perceptual differences. In contrast, the feature magnitude varies consistently with MOS, increasing for higher-quality images and decreasing for lower-quality ones. Moreover, we observe that cosine-based scores are more reliable in distinguishing high-quality images, where semantic features remain well aligned with CLIP's pretrained distribution, while magnitude cues are more sensitive and consistent in low-quality regimes, where distortions cause semantic misalignment (see Fig. 1(b)). This insight suggests a key conclusion: ***cosine similarity and feature magnitude are complementary***. Motivated by this, we propose to leverage both cues jointly rather than relying on either alone.

To this end, we introduce an adaptive dual-cue fusion framework that integrates semantic and magnitude information for more robust quality prediction. Specifically, we compute two scalar quality indicators: (1) the conventional cosine similarity between image embedding and textual quality prompts, and (2) a magnitude-based cue derived directly from the image features. For the latter, we first take the absolute value of each feature dimension and apply Box-Cox transformation to statistically normalize their distributions across images. This normalization mitigates semantic-content bias and aligns the magnitudes to near-Gaussian distribution. We then average the transformed values to obtain a stable, debiased magnitude score. Finally, to fully exploit their complementary strengths, we design a confidence-guided fusion mechanism that adaptively weights the two cues based on their estimated reliability. This allows the model to trust the cosine score more in high-quality conditions where image semantics are well recognizable, and rely more on the normalized magnitude cue under severe distortions where semantic similarity becomes less reliable.

Extensive experiments on multiple IQA benchmarks validate the effectiveness of our approach. Without any task-specific training, our method substantially outperforms the vanilla CLIP-IQA baseline and recent state-of-the-art NR-IQA models. These results highlight the benefits of combining semantic and magnitude cues for robust and accurate image quality prediction. Our main contributions are summarized as follows:

- We identify the magnitude of CLIP image embedding as a strong and previously overlooked quality cue for NR-IQA that complements traditional cosine similarity.
- We introduce a Box-Cox transformation to normalize per-dimension embedding magnitudes, producing a statistically consistent quality indicator across diverse image contents.
- We design a confidence-guided fusion strategy that adaptively weights cosine similarity and magnitude cues based on their relative reliability.
- Our method is entirely training-free, achieves state-of-the-art zero-shot IQA performance, and generalizes effectively to different image content, demonstrating the versatility of the proposed dual-cue framework.

Related Work

Vision Based NR-IQA Models

Early no-reference IQA approaches relied on handcrafted features that captured natural scene statistics (NSS), with descriptors derived from spatial (Mittal, Soundararajan, and Bovik 2012), wavelet (Moorthy and Bovik 2010), and DCT (Saad, Bovik, and Charrier 2012) domains. Psychovisual models inspired by the free-energy principle of the human visual system further enriched the modeling of perceptual degradation (Gu et al. 2014; Zhai et al. 2011). With the advent of deep learning, CNN-based models became prevalent. Early works such as Kang *et al.* (Kang et al. 2014) learned quality-aware representations directly from image patches, and subsequent extensions introduced multi-task training (Kang et al. 2015) to jointly predict quality and distortion types. Later models such as DIQaM (Bosse et al. 2017), FPR (Chen et al. 2020), and GraphIQA (Sun et al. 2022) further advanced performance by modeling spatial dependencies and incorporating relational reasoning. Transformer-based models have also been explored due to their strong modeling capacity and flexibility. You *et al.* (You and Korhonen 2021) introduced a Transformer backbone for NR-IQA that benefits from features pretrained on large-scale classification tasks. Other approaches leverage perceptual priors from image restoration networks (Lin and Wang 2018; Chen et al. 2022). However, these models struggle to generalize due to limited data and domain-specific bias. To address this, recent work incorporates auxiliary tasks (You and Korhonen 2021; Lin and Wang 2018) and learning strategies such as meta-learning (Zhu et al. 2020), curriculum learning (Wang et al. 2023), and domain generalization (Chen et al. 2021, 2025). Despite progress, generalization in NR-IQA remains an open problem.

Vision-Language Based NR-IQA Models

Vision-language models provide a new paradigm for NR-IQA. Wang *et al.* (Wang, Chan, and Loy 2023) first leveraged CLIP (Radford et al. 2021) to compute the similarity between distorted images and antonym-paired textual prompts. Follow-up work such as IPCE (Peng et al. 2024) mapped cosine similarity scores to discrete quality levels using hand-crafted prompts, while CLIP-AGIQA (Tang et al. 2024) learned prompt tokens for six quality classes and concatenated them with image features for quality regression. In parallel, benchmarks such as Q-Bench (Wu et al. 2023a) and DepictQA (You et al. 2024) evaluated the IQA capability of large multimodal models (LMMs) on low-level visual perception. Zhu *et al.* (Zhu et al. 2024a) examined 2AFC-style prompting for preference judgment, and Q-Align (Wu et al.

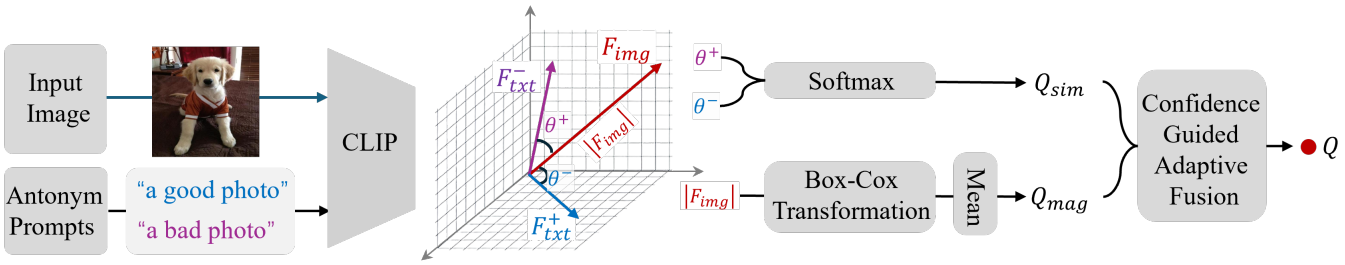


Figure 2: Overview of the Proposed Magnitude-Aware CLIP IQA Framework. Given an input image, we extract its CLIP image embedding and compute two quality signals: (1) Q_{sim}, the image semantic similarity with text prompts, and (2) Q_{mag}, a magnitude-based score obtained via Box-Cox transformation for statistical normalization. To balance these complementary cues, we adopt a confidence discrepancy and generate softmax-based fusion weights, producing the final quality prediction Q.

2023b) proposed rating-level prompts to elicit more consistent predictions. Further improvements have been achieved by leveraging pre-trained LMMs and high-quality instruction datasets (Wu et al. 2024; Zhu et al. 2024b). However, fine-tuning these models on IQA tasks often results in catastrophic forgetting (Luo et al. 2023), weakening performance on other domains. This motivates the need to explore training-free strategies that better preserve the generality of pre-trained vision-language models.

Method

Preliminary: Semantic Similarity Based CLIP-IQA

In the classical CLIP-IQA model, the zero-shot capability of the CLIP model for NR-IQA has been explored. In particular, the image quality is estimated by measuring the alignment between image embedding and handcrafted quality prompts, such as $\{$*"a good photo"*$\}$ *vs.*$\{$*"a bad photo"*$\}$. Formally, let $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ denote an input image, and let $\mathcal{T}_{pos}$ and $\mathcal{T}_{neg}$ denote a pair of antonymic textual descriptions reflecting high and low perceptual quality, respectively. The CLIP model is utilized to encode the image and text into embedding vectors:

$$F_{img} = \phi_{img}(\mathbf{x}) \in \mathbb{R}^D, \tag{1}$$

$$F_{txt}^{+} = \phi_{txt}(\mathcal{T}_{pos}) \in \mathbb{R}^D, \tag{2}$$

$$F_{txt}^{-} = \phi_{txt}(\mathcal{T}_{neg}) \in \mathbb{R}^D, \tag{3}$$

where $\phi_{img}(\cdot)$ and $\phi_{text}(\cdot)$ are the image and text encoders of the pre-trained CLIP model, respectively. The cosine similarity between the normalized embeddings is computed as:

$$s^{+} = \cos(F_{img}, F_{text}^{+}) = \hat{F}_{img} \cdot \hat{F}_{text}^{+}, \tag{4}$$

$$s^{-} = \cos(F_{img}, F_{text}^{-}) = \hat{F}_{img} \cdot \hat{F}_{text}^{-}, \tag{5}$$

where $\hat{F} = F/\|F\|_2$, denotes the ℓ_2-normalized results. The final quality score is obtained using a softmax-based probability:

$$Q_{sim} = \frac{\exp(s^{+}/\tau)}{\exp(s^{+}/\tau) + \exp(s^{-}/\tau)}, \tag{6}$$

where τ is a temperature hyperparameter. The Q_{sim} reflects the relative probability that the image with good quality than the bad one, which can be deemed as the image quality assessment result.

Magnitude-Aware CLIP IQA Model

An overview of the proposed framework is illustrated in Fig. 2. The pipeline are described in detail as follows.

Limitations of Cosine Similarity As depicted in Eqn. (4), the cosine similarity inherently normalizes both input vectors, which discards the magnitude information of the image embedding. However, we empirically observe that the magnitude also presents a high correlation with perceptual quality: High-quality images typically yield rich and discriminative features, reflected by larger magnitudes, while heavily degraded images exhibit reduced embedding norms. Nevertheless, as shown in Fig. 3, the magnitude distributions vary significantly across different image content, even under similar perceptual quality. This content-dependent variation introduces a semantic bias that impairs direct comparison of magnitude scores across samples.

Statistical Normalization via Box-Cox Transformation To mitigate the semantic bias inherent in raw CLIP embeddings, we introduce a statistical normalization approach based on the Box-Cox transformation (Box and Cox 1964), which is a classical power transform that stabilizes variance and reduces distributional skewness. Unlike cosine similarity, which discards magnitude information via ℓ_2 normalization, our goal is to retain and standardize this information for quality prediction. Given the image-level feature embedding $F_{img} \in \mathbb{R}^D$, we first take its element-wise absolute value to remove the polarity and retain only activation strength:

$$\hat{F} = |F_{img}| \in \mathbb{R}^D. \tag{7}$$

Then we normalize it by its standard deviation:

$$\tilde{F} = \frac{\hat{F}}{\sigma + \varepsilon}, \tag{8}$$

where σ is the standard deviation computed over all dimensions, and ε is a small constant for numerical stability. The Box-Cox transformation is finally applied independently to each dimension:

$$\mathbf{T}_d = \begin{cases} \dfrac{(\tilde{F}_d + 1)^{\lambda} - 1}{\lambda}, & \lambda \neq 0, \\ \log(\tilde{F}_d + 1), & \lambda = 0, \end{cases} \tag{9}$$

where $\mathbf{T}_d$ denotes the transformed d-th feature, and λ is the power parameter. Finally, the normalized magnitude-based quality score is obtained by averaging across all dimensions:

$$Q_{mag} = \frac{1}{D}\sum_{d=1}^{D}\mathbf{T}_d. \tag{10}$$

This transformation yields a statistically normalized scalar that effectively captures perceptual quality variations while remaining robust to semantic content.

Confidence-Guided Adaptive Fusion While both Q_{sim} and Q_{mag} offer valuable but distinct quality cues, their reliability is not uniform across different image conditions. Specifically, Q_{sim}, which relies on semantic similarity between image embedding and textual prompts, is more robust when semantic content is well-preserved, such as in high-quality images. In contrast, Q_{mag} captures distortions through statistical deviations in embedding magnitude, making it more responsive under severe degradation where semantic alignment is compromised.

To adaptively leverage their complementary strengths, we design a fusion scheme that dynamically adjusts the contribution of each cue based on their agreement. We begin by computing the discrepancy between the two estimates:

$$\Delta = Q_{sim} - Q_{mag}, \tag{11}$$

which quantifies the direction and degree of disagreement. A large positive Δ suggests that Q_{sim} is relatively confident (*e.g.*, in a clean image), whereas a negative Δ indicates greater reliability in Q_{mag} (*e.g.*, when content is distorted). This discrepancy implicitly reflects the underlying quality level of the image and serves as a signal for confidence reweighting. We convert Δ into two fusion logits through an affine transformation:

$$\gamma_{sim} = 1.0 + \alpha\Delta, \tag{12}$$

$$\gamma_{mag} = 0.6 - \alpha\Delta, \tag{13}$$

where α is a tunable hyperparameter controlling the sensitivity of the fusion to confidence gaps. The base constants (1.0 and 0.6) encode prior trust in the two metrics, while Δ adaptively adjusts these values based on content quality. We then apply softmax normalization to ensure the resulting weights form a valid probability distribution:

$$[w_{sim}, w_{mag}] = \text{softmax}([\gamma_{sim}, \gamma_{mag}]). \tag{14}$$

Finally, the overall perceptual quality score is obtained as a convex combination of both cues:

$$Q = w_{sim}\cdot Q_{sim} + w_{mag}\cdot Q_{mag}. \tag{15}$$

Experiments

Experimental Settings

Implementation Details We implement our method using PyTorch, based on the pretrained Resnet50 model. All evaluations are conducted under a zero-shot setting, where no ground-truth supervision is used during model optimization. For the Box-Cox transformation in Q_{mag}, the power parameter λ is empirically set to 0.5, and we add 1.0 as an offset to ensure positivity. The α in Eqn. (12) is fixed by 1.0. All experiments are conducted using a single NVIDIA 3090 GPU.

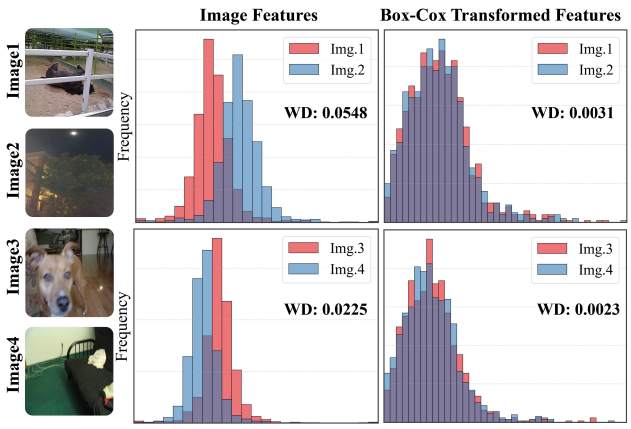


Figure 3: Semantic Bias exists in CLIP feature. This feature magnitudes for visually similar-quality images differ substantially across semantic categories. Statistical normalization is vital to make magnitude cues reliable. WD represents the Wasserstein Distance between two Feature distribution.

Datasets We evaluate our method on a wide range of datasets to verify its generalization and robustness. These include: **(1) Synthetic distortion datasets**: CSIQ (Larson and Chandler 2010), TID2013 (Ponomarenko et al. 2015), and KADID-10k (Lin, Hosu, and Saupe 2019), containing various artificially generated distortions. **(2) Authentic distortion datasets**: CLIVE (Ghadiyaram and Bovik 2015), KonIQ (Hosu et al. 2020), and SPAQ (Fang et al. 2020), reflecting real-world degradations from mobile photography. **(3) Image restoration (IR) datasets**: PIPAL (Jinjin et al. 2020), featuring restored images from SR/denoising/deblurring pipelines. **(4) AIGC quality datasets**: AGIQA-1k (Zhang et al. 2023) and AGIQA-3k (Li et al. 2023), designed for quality assessment of AI-generated content. All evaluations follow the standard protocol, where Spearman's Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC) are reported.

Comparison Methods We compare our method with two categories of existing IQA approaches: **(1) Opinion-Unaware (OU) Methods.** These methods do not require subjective opinion scores for training and are typically used in a zero-shot manner. We include NIQE (Mittal, Soundararajan, and Bovik 2012), QAC (Xue, Zhang, and Mou 2013), PIQE (Venkatanath et al. 2015), LPSI (Wu, Wang, and Li 2015), dipIQ (Ma et al. 2017), SNP-NIQE (Liu et al. 2019), NPQI (Liu et al. 2020), CLIPIQA (Wang, Chan, and Loy 2023), ContentSep (Babu, Kannan, and Soundararajan 2023), and MDFS (Ni et al. 2024) in this category. Since our method is also a zero-shot approach that requires no training on opinion scores, this group serves as the most appropriate baseline for a fair comparison. **(2) Learning-based Methods.** These methods rely on training with human-annotated quality scores and are typically optimized for specific IQA datasets. We include Re-IQA (Saha, Mishra, and Bovik 2023), ARNIQA (Agnolucci et al. 2024), CLIP-IQA$^+$ (Wang, Chan, and Loy 2023), and GRepQ (Sri-

	Dataset	Method									
		NIQE	QAC	PIQE	dipIQ	SNP-NIQE	NPQI	ContentSep	MDFS	CLIP-IQA	MA-CLIP
SRCC	CLIVE	0.4495	0.2258	0.2325	0.2089	0.4654	0.4752	0.5060	0.4821	0.7019	**0.7428** (+5.8%)
	CSIQ	0.6191	0.4804	0.5120	0.5191	0.6090	0.6341	0.5871	**0.7774**	0.6807	0.7374 (+8.3%)
	TID2013	0.3106	0.3719	0.3636	0.4377	0.3329	0.2804	0.2530	0.5363	0.5786	**0.5990** (+3.5%)
	KADID	0.3779	0.2394	0.2372	0.2977	0.3719	0.3909	0.5060	**0.5983**	0.5009	0.5251 (+4.8%)
	KonIQ	0.5300	0.3397	0.2452	0.2375	0.6284	0.6132	0.6401	0.7333	0.6846	**0.7645** (+11.7%)
	SPAQ	0.3105	0.4397	0.2317	0.2189	0.5402	0.5999	0.7084	0.7408	0.7144	**0.7725** (+8.1%)
	AVG	0.4329	0.3495	0.3037	0.3200	0.4913	0.4990	0.5682	0.6487	0.6296	**0.6902** (+9.6%)
PLCC	CLIVE	0.4939	0.2841	0.3144	0.3163	0.5199	0.4920	0.5130	0.5364	0.7217	**0.7680** (+6.4%)
	CSIQ	0.6901	0.5934	0.6279	0.7009	0.6962	0.6479	0.3632	**0.7907**	0.7270	0.7828 (+7.7%)
	TID2013	0.3789	0.4190	0.4615	0.4746	0.4055	0.4000	0.2203	0.6242	0.6552	**0.6756** (+3.1%)
	KADID	0.3883	0.3088	0.2887	0.3832	0.4212	0.3401	0.3568	**0.5939**	0.5204	0.5489 (+5.5%)
	KonIQ	0.4835	0.2906	0.2061	0.3773	0.6222	0.6139	0.6274	0.7123	0.7124	**0.8035** (+12.8%)
	SPAQ	0.2639	0.4497	0.2488	0.2239	0.5469	0.6155	0.6648	0.7177	0.7179	**0.7775** (+8.3%)
	AVG	0.4498	0.3909	0.3579	0.4127	0.5353	0.5182	0.4576	0.6625	0.6981	**0.7261** (+4.0%)

Table 1: Performance comparison of opinion-unaware IQA models on six benchmark datasets. The best results are **bolded**. Relative gains of MA-CLIP over CLIP-IQA are annotated in each cell.

Method	Setting		PIPAL		AGIQA-1k		AGIQA-3k	
	CD	ZS	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Re-IQA	✓		0.568	0.587	0.783	0.840	0.811	0.874
ARNIQA	✓		0.634	0.666	0.768	0.849	0.803	0.881
CLIP-IQA$^+$	✓		0.552	0.558	0.817	0.855	0.844	0.894
GRepQ	✓		0.554	0.568	0.740	0.797	0.807	0.858
NIQE		✓	0.167	0.181	0.623	0.721	0.510	0.526
IL-NIQE		✓	0.231	0.220	**0.645**	**0.757**	0.528	0.544
CL-MI		✓	0.281	0.282	0.474	0.621	0.591	0.665
CLIP-IQA		✓	0.332	0.339	0.511	0.644	0.658	0.716
MA-CLIP		✓	**0.371**	**0.393**	0.528	**0.668**	**0.706**	**0.764**
Gain			**+11.7%**	**+15.9%**	**+3.3%**	**+3.9%**	**+7.3%**	**+6.7%**

Table 2: Quantitative results on image restoration and AIGC datasets. CD: Cross-dataset, ZS: Zero-shot. The last row shows the relative gain of MA-CLIP compared to CLIP-IQA. Best scores are highlighted in bold.

nath et al. 2024). For a comprehensive evaluation of generalization capability, we compare with these learning-based methods trained on the training split of each testing dataset.

Quantitative Comparison

As summarized in Table 1, our Magnitude-Aware CLIP (MA-CLIP) achieves consistent and significant performance gains compared with CLIP-IQA across all benchmark categories, demonstrating its robustness under various distortion types and domains.

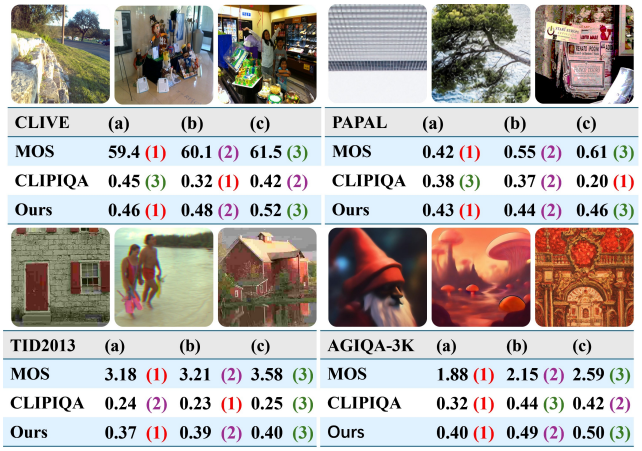
On synthetic distortion datasets of CSIQ, our method outperforms existing opinion-unaware baselines and CLIP-based models by a large margin. For instance, on TID2013, MA-CLIP achieves an SRCC of 0.599, surpassing CLIP-IQA by 3.5%. This improvement reflects the benefit of incorporating distortion-sensitive features via our normalized magnitude modeling, which enhances the model's ability to detect subtle degradation patterns often present in synthetic benchmarks. On real-world datasets like KonIQ-10k and SPAQ, where distortions are more diverse and semantically entangled, MA-CLIP achieves an SRCC of 0.765 on KonIQ-10k, outperforming all OU methods. This shows that the combination of Q_{sim} and statistically normalized Q_{mag} enables a more holistic perception of quality.

To further test generalization in downstream applications, Table 2 reports results on image restoration (PIPAL) and AIGC (AGIQA-1k/3k) datasets. These datasets feature challenging distribution shifts, such as hallucination artifacts, over-smoothing, or texture inconsistency, which are often poorly handled by traditional or purely semantic-based metrics. MA-CLIP achieves the best PLCC of 0.706 and SRCC of 0.764 on AGIQA-3k, surpassing recent multimodal methods like MDFS. It also performs competitively on PIPAL, where many models struggle due to the diverse restoration algorithms and overfitting risks.

The performance gain confirms the advantage of our adaptive fusion design, which dynamically balances semantic and magnitude cues based on image-specific confidence. In addition, we compare against learning-based methods trained on KonIQ-10k (*e.g.*, GRepQ, CLIP-IQA$^+$). While these models leverage large-scale opinion-aware data, our zero-shot MA-CLIP still achieves highly competitive results, especially on datasets it has never seen during training. This highlights the strong generalization capability of our method, without sacrificing interpretability or requiring expensive annotations.

Qualitative Visualization

To further complement the quantitative results, we provide two qualitative visualization that illustrate the efficacy of our proposed method from a human-understandable perspective:



CLIVE	(a)	(b)	(c)
MOS	59.4 (1)	60.1 (2)	61.5 (3)
CLIPIQA	0.45 (3)	0.32 (1)	0.42 (2)
Ours	0.46 (1)	0.48 (2)	0.52 (3)

PAPAL	(a)	(b)	(c)
MOS	0.42 (1)	0.55 (2)	0.61 (3)
CLIPIQA	0.38 (3)	0.37 (2)	0.20 (1)
Ours	0.43 (1)	0.44 (2)	0.46 (3)

TID2013	(a)	(b)	(c)
MOS	3.18 (1)	3.21 (2)	3.58 (3)
CLIPIQA	0.24 (2)	0.23 (1)	0.25 (3)
Ours	0.37 (1)	0.39 (2)	0.40 (3)

AGIQA-3K	(a)	(b)	(c)
MOS	1.88 (1)	2.15 (2)	2.59 (3)
CLIPIQA	0.32 (1)	0.44 (3)	0.42 (2)
Ours	0.40 (1)	0.49 (2)	0.50 (3)

Figure 4: MOS alignment visualization. Representative examples from multiple datasets illustrating the ranking alignment between MOSs and our MA-CLIP predictions. Each triplet shows three images with their MOSs and predicted quality scores of CLIP-IQA and our MA-CLIP.

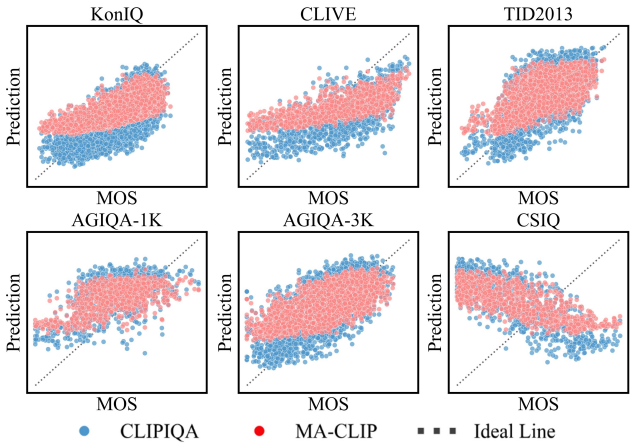


Figure 5: Scatter plot comparison of CLIP-IQA and MA-CLIP. The x-axis represents the MOS, while the y-axis shows the prediction. As the scatter gets closer to the ideal line, it indicates that the model predicts better.

(1) MOS alignment visualization As shown in Fig. 4, we present representative examples from diverse datasets where the original CLIP-IQA model fails to correctly rank the image qualities in accordance with the MOS. In each set, we display three images along with their MOS and the predicted quality scores from both CLIP-IQA and our proposed MA-CLIP. It can be observed that CLIP-IQA often over-relies on semantic content and yields inverted or inconsistent rankings. In contrast, our method incorporates magnitude-aware correction, which adjusts the quality estimates to better reflect perceptual degradation. As a result, the predicted order aligns more closely with the MOS-based ranking.

(2) Scatter plot comparison In Fig. 5, we visualize the scatter plots comparing CLIP-IQA and MA-CLIP over six datasets including four representative groups: (a) synthetic distortions, (b) authentic distortions, (c) IR (mage restoration, and (d) AIGC-generated images. Each plot maps the predicted scores versus the MOSs, where the ideal prediction would lie along the diagonal line. The scatter patterns reveal that CLIP-IQA suffers from more dispersed and biased predictions, especially on complex or underrepresented distortions. In contrast, MA-CLIP exhibits tighter clustering around the diagonal, indicating improved consistency and robustness in ranking.

Ablation Study

To thoroughly validate the effectiveness and design choices of our proposed MA-CLIP framework, we conduct a series of ablation experiments. These experiments are designed to isolate and quantify the contribution of each key component, including individual prediction branches, magnitude feature extraction strategies, fusion mechanisms, statistical parameters, and backbone architectures.

(1) Contribution of Each Branch. We first evaluate the independent performance of the two scoring branches:

Dataset	Q_{sim}		Q_{mag}		**Fusion**	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
CLIVE	0.702	0.722	0.418	0.503	**0.743**	**0.768**
CSIQ	0.681	0.727	0.448	0.460	**0.737**	**0.783**
KonIQ-10k	0.685	0.712	0.557	0.616	**0.765**	**0.804**
SPAQ	0.714	0.718	0.578	0.598	**0.773**	**0.775**
AVG	0.696	0.720	0.500	0.544	**0.755**	**0.783**

Table 3: Ablation study on the contribution of each scoring branch. Q_{sim}: semantic-only; Q_{mag}: magnitude-only; Fusion: confidence-guided combination.

Q_{sim}, which captures semantic similarity between image and text embeddings by CLIP model; and Q_{mag}, which estimates distortion severity via magnitude information. As reported in Table 3, Q_{sim} performs well on content-consistent distortions but fails to capture quality degradation in texture-corrupt or over-smoothed cases, often overemphasizing high-level semantics. In contrast, Q_{mag} exhibits stronger sensitivity to signal-level degradations but is less reliable when semantic preservation is critical. The proposed confidence-guided fusion of both branches results in a substantial SRCC gain over 8.5% , demonstrating their complementary nature.

(2) Variants of Magnitude Feature Extraction. We compare three strategies for computing the magnitude-based feature: (i) L1 norm, (ii) L2 norm, and (iii) our proposed Box-Cox normalized features norm. As shown in Table 4, the Box-Cox-based normalization yields consistently higher correlation with perceptual quality scores across all datasets. In particular, it improves SRCC by 59% over the L2 variant on KonIQ-10k, highlighting the benefits of distributional stabilization. When each variant is used independently (without fusion), Box-Cox normalization again outperforms

Type	CSIQ		CLIVE		KonIQ-10k	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
L1	0.126	0.152	0.345	0.369	0.495	0.530
L2	0.299	0.386	0.193	0.304	0.350	0.349
Ours	**0.448**	**0.461**	**0.418**	**0.503**	**0.557**	**0.616**

Table 4: Comparison on perceptual-optimized feature norm with different feature norm on CSIQ, CLIVE and KonIQ.

Type	CG-Fusion		CSIQ		KADID-10k	
	w_{sim}	w_{mag}	SRCC	PLCC	SRCC	PLCC
L1		✓	0.126	0.152	0.394	0.397
L2		✓	0.280	0.398	0.338	0.345
Ours	0.8	0.2	0.736	0.779	0.523	0.544
	0.2	0.8	0.700	0.742	0.445	0.492
	0.5	0.5	0.734	0.779	0.520	0.545
		✓	**0.737**	**0.783**	**0.525**	**0.549**

Table 5: Ablation study on weight combination with different feature norm on CLIVE and KADID-10k.

the alternatives, indicating its standalone robustness in capturing distortion-aware cues.

(3) Fusion Strategy Comparison. To assess the effectiveness of our confidence-weighted adaptive fusion, we compare it against several baselines: (i) equal-weighted average, and (ii) fixed-weight summations with various ratios. Results in Table 5 show that adaptive fusion achieves superior performance across all tested datasets. This confirms that confidence-aware fusion dynamically adjusts to different image conditions, enabling better utilization of the complementary strengths of Q_{sim} and Q_{mag}.

(4) Sensitivity to Box-Cox Transformation Parameter λ. To investigate the robustness of the Box-Cox normalization, we conduct a sensitivity analysis on the transformation parameter λ, which controls the degree of non-linearity applied to magnitude values. As plotted in Fig. 6, we observe that small positive values (*e.g.*, $\lambda = 0.5$) consistently yield stable and high SRCC values. Very large λ leads to a performance drop due to either over-flattening (loss of signal variance) or numerical instability. These findings suggest that light nonlinear normalization is sufficient to suppress magnitude outliers while preserving distortion-relevant information.

(5) Impact of Backbone Architectures. We further assess the generality of MA-CLIP across diverse CLIP backbones, including ResNet-50, ResNet-101, ViT-B/32, and ViT-L/14. As shown in Table 6, our magnitude-aware design consistently improves performance across all architectures, surpassing the corresponding CLIP-IQA baselines in both SRCC and PLCC. The consistent gains highlight the effectiveness of magnitude-aware correction even when applied to models with high semantic representation capacity.

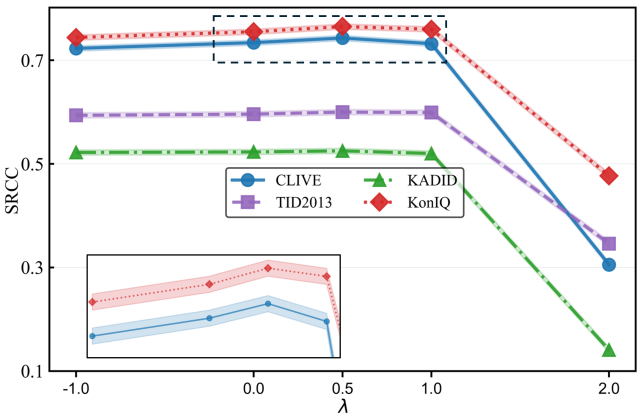


Figure 6: Sensitivity of Box-Cox parameter λ on SRCC across on CLIVE, TID2013, KonIQ-10k and KADID-10k.

	Method	CSIQ		KonIQ-10k		AGIQA-3k	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
R50	CLIPIQA	0.681	0.727	0.685	0.712	0.658	0.716
	Ours	**0.737**	**0.783**	**0.765**	**0.804**	**0.706**	**0.764**
R101	CLIPIQA	0.715	0.705	0.710	0.730	0.643	0.697
	Ours	**0.741**	**0.764**	**0.727**	**0.748**	**0.667**	**0.743**
B/32	CLIPIQA	0.763	0.783	0.715	0.743	0.663	0.710
	Ours	**0.783**	**0.810**	**0.760**	**0.803**	**0.694**	**0.757**
L/14	CLIPIQA	0.622	0.628	0.682	0.709	0.699	0.788
	Ours	**0.666**	**0.680**	**0.717**	**0.752**	**0.717**	**0.814**

Table 6: Impact of different CLIP backbones on MA-CLIP performance (SRCC/PLCC). Best scores for each backbone are highlighted in bold.

Conclusion

In this work, we revisit CLIP-based NR-IQA by identifying a crucial yet previously overlooked quality cue: the magnitude of CLIP image features. While existing CLIP-IQA approaches rely solely on prompt-based cosine similarity, we demonstrate that feature magnitude exhibits strong and complementary correlation with perceptual quality. To harness both cues effectively, we propose a novel, training-free dual-source framework that integrates a statistically normalized magnitude score with semantic similarity via a confidence-guided fusion strategy. Extensive experiments across diverse IQA benchmarks show that our method consistently outperforms both CLIP-IQA and state-of-the-art NR-IQA models, without requiring any task-specific fine-tuning. These findings highlight the value of revisiting internal properties of pretrained models and open new directions for plug-and-play quality assessment leveraging multimodal embeddings.

Acknowledgments

The work was supported by the National Natural Science Foundation of China under Grant No. 62401214, the National Natural Science Foundation of China under Grant No.

62477015, the Key Research and Development Program of Guangdong of China under Grant No. 2023B0303010004, and the Innovation Team Project for Universities in Guangdong Province in China under Grant No. 2023KCXTD011.

References

Agnolucci, L.; Galteri, L.; Bertini, M.; and Del Bimbo, A. 2024. ARNIQA: Learning distortion manifold for image quality assessment. In *IEEE Winter Conference on Applications of Computer Vision*, 189–198.

Babu, N. C.; Kannan, V.; and Soundararajan, R. 2023. No reference opinion unaware quality assessment of authentically distorted images. In *IEEE Winter Conference on Applications of Computer Vision*, 2459–2468.

Bosse, S.; Maniry, D.; Müller, K.-R.; Wiegand, T.; and Samek, W. 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1): 206–219.

Box, G. E.; and Cox, D. R. 1964. An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2): 211–243.

Chen, B.; Li, H.; Fan, H.; and Wang, S. 2020. No-reference screen content image quality assessment with unsupervised domain adaptation. *arXiv preprint arXiv:2008.08561*.

Chen, B.; Xiao, K.; Shen, X.; and Wang, S. 2025. Monotonic and Invertible Network: A General Framework for Learning IQA Model from Mixed Datasets. *International Journal of Computer Vision*, 1–22.

Chen, B.; Zhu, L.; Kong, C.; Zhu, H.; Wang, S.; and Li, Z. 2022. No-reference image quality assessment by hallucinating pristine features. *IEEE Transactions on Image Processing*, 31: 6139–6151.

Chen, B.; Zhu, L.; Li, G.; Lu, F.; Fan, H.; and Wang, S. 2021. Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4): 1903–1916.

Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; and Wang, Z. 2020. Perceptual quality assessment of smartphone photography. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3677–3686.

Ghadiyaram, D.; and Bovik, A. C. 2015. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1): 372–387.

Gu, K.; Zhai, G.; Yang, X.; and Zhang, W. 2014. Using free energy principle for blind image quality assessment. *IEEE Transactions on Multimedia*, 17(1): 50–63.

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

Jinjin, G.; Haoming, C.; Haoyu, C.; Xiaoxing, Y.; Ren, J. S.; and Chao, D. 2020. PIPAL: a Large-Scale Image Quality Assessment Dataset for Perceptual Image Restoration. In *European Conference on Computer Vision*, 633–651. Springer.

Kang, L.; Ye, P.; Li, Y.; and Doermann, D. 2014. Convolutional neural networks for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1733–1740.

Kang, L.; Ye, P.; Li, Y.; and Doermann, D. 2015. Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In *IEEE International Conference on Image Processing*, 2791–2795.

Larson, E. C.; and Chandler, D. M. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1): 011006.

Li, C.; Zhang, Z.; Wu, H.; Sun, W.; Min, X.; Liu, X.; Zhai, G.; and Lin, W. 2023. AGIQA-3K: an Open Database for AI-Generated Image Quality Assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8): 6833–6846.

Lin, H.; Hosu, V.; and Saupe, D. 2019. KADID-10k: A Large-Scale Artificially Distorted IQA Database. In *Proceedings of the IEEE Eleventh International Conference on Quality of Multimedia Experience*, 1–3.

Lin, K.-Y.; and Wang, G. 2018. Hallucinated-IQA: No-reference image quality assessment via adversarial learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 732–741.

Liu, Y.; Gu, K.; Li, X.; and Zhang, Y. 2020. Blind image quality assessment by natural scene statistics and perceptual characteristics. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(3): 1–91.

Liu, Y.; Gu, K.; Zhang, Y.; Li, X.; Zhai, G.; Zhao, D.; and Gao, W. 2019. Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4): 929–943.

Luo, Y.; Yang, Z.; Meng, F.; Li, Y.; Zhou, J.; and Zhang, Y. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Ma, K.; Liu, W.; Liu, T.; Wang, Z.; and Tao, D. 2017. dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, 26(8): 3951–3964.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3): 209–212.

Moorthy, A. K.; and Bovik, A. C. 2010. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, 17(5): 513–516.

Ni, Z.; Liu, Y.; Ding, K.; Yang, W.; Wang, H.; and Wang, S. 2024. Opinion-Unaware Blind Image Quality Assessment using Multi-Scale Deep Feature Statistics. *IEEE Transactions on Multimedia*.

Peng, F.; Fu, H.; Ming, A.; Wang, C.; Ma, H.; He, S.; Dou, Z.; and Chen, S. 2024. AIGC image quality assessment

via image-prompt correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6432–6441.

Ponomarenko, N.; Jin, L.; Ieremeiev, O.; Lukin, V.; Egiazarian, K.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F.; et al. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30: 57–77.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Saad, M. A.; Bovik, A. C.; and Charrier, C. 2012. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, 21(8): 3339–3352.

Saha, A.; Mishra, S.; and Bovik, A. C. 2023. Re-IQA: Unsupervised Learning for Image Quality Assessment in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5846–5855.

Srinath, S.; Mitra, S.; Rao, S.; and Soundararajan, R. 2024. Learning Generalizable Perceptual Representations for Data-Efficient No-Reference Image Quality Assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 22–31.

Sun, S.; Yu, T.; Xu, J.; Zhou, W.; and Chen, Z. 2022. GraphIQA: Learning distortion graph representations for blind image quality assessment. *IEEE Transactions on Multimedia*.

Tang, Z.; Wang, Z.; Peng, B.; and Dong, J. 2024. CLIP-AGIQA: boosting the performance of ai-generated image quality assessment with clip. In *International Conference on Pattern Recognition*, 48–61. Springer.

Venkatanath, N.; Praneeth, D.; Bh, M. C.; Channappayya, S. S.; and Medasani, S. S. 2015. Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications (NCC)*, 1–6. IEEE.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 2555–2563.

Wang, J.; Chen, Z.; Yuan, C.; Li, B.; Ma, W.; and Hu, W. 2023. Hierarchical curriculum learning for no-reference image quality assessment. *International Journal of Computer Vision*, 131(11): 3074–3093.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Li, C.; Sun, W.; Yan, Q.; Zhai, G.; et al. 2023a. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2023b. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*.

Wu, H.; Zhu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Li, C.; Wang, A.; Sun, W.; Yan, Q.; et al. 2024. Towards open-ended visual quality comparison. *arXiv preprint arXiv:2402.16641*.

Wu, Q.; Wang, Z.; and Li, H. 2015. A Highly Efficient Method for Blind Image Quality Assessment. In *Proceedings of the IEEE International Conference on Image Processing*, 339–343.

Xue, W.; Zhang, L.; and Mou, X. 2013. Learning without human scores for blind image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 995–1002.

You, J.; and Korhonen, J. 2021. Transformer for image quality assessment. In *IEEE International Conference on Image Processing (ICIP)*, 1389–1393. IEEE.

You, Z.; Gu, J.; Li, Z.; Cai, X.; Zhu, K.; Xue, T.; and Dong, C. 2024. Descriptive Image Quality Assessment in the Wild. *arXiv preprint arXiv:2405.18842*.

Zhai, G.; Wu, X.; Yang, X.; Lin, W.; and Zhang, W. 2011. A psychovisual quality metric in free-energy principle. *IEEE Transactions on Image Processing*, 21(1): 41–52.

Zhang, Z.; Li, C.; Sun, W.; Liu, X.; Min, X.; and Zhai, G. 2023. A Perceptual Quality Assessment Exploration for AIGC Images. In *IEEE International Conference on Multimedia and Expo Workshops*, 440–445. IEEE.

Zhu, H.; Li, L.; Wu, J.; Dong, W.; and Shi, G. 2020. MetaIQA: Deep meta-learning for no-reference image quality assessment. In *IEEEF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14143–14152.

Zhu, H.; Sui, X.; Chen, B.; Liu, X.; Chen, P.; Fang, Y.; and Wang, S. 2024a. 2AFC prompting of large multimodal models for image quality assessment. *arXiv preprint arXiv:2402.01162*.

Zhu, H.; Wu, H.; Li, Y.; Zhang, Z.; Chen, B.; Zhu, L.; Fang, Y.; Zhai, G.; Lin, W.; and Wang, S. 2024b. Adaptive Image Quality Assessment via Teaching Large Multimodal Model to Compare. *arXiv preprint arXiv:2405.19298*.