LLM-TTA: Leveraging Large Language Models for Test-Time Adaptation in Image Quality Assessment

Ambreen Bashir
Indian Institute of Technology Jammu
Jammu and Kashmir, India
2018rcs0049@iitjammu.ac.in

Rahul Kumar
Indian Institute of Technology Jammu
Jammu and Kashmir, India
2021ucs0110@iitjammu.ac.in

Vinit Jakhetiya
Indian Institute of Technology Jammu
Jammu and Kashmir, India
vinit.jakhetiya@iitjammu.ac.in

Badri N. Subudhi
Indian Institute of Technology Jammu
Jammu and Kashmir, India
subudhi.badri@iitjammu.ac.in

Sunil Jaiswal
K|Lens GmbH
Saarbrücken, Germany
sunil.jaiswal@k-lens.de

Weisi Lin
Nanyang Technological University
Singapore, Singapore
wslin@ntu.edu.sg

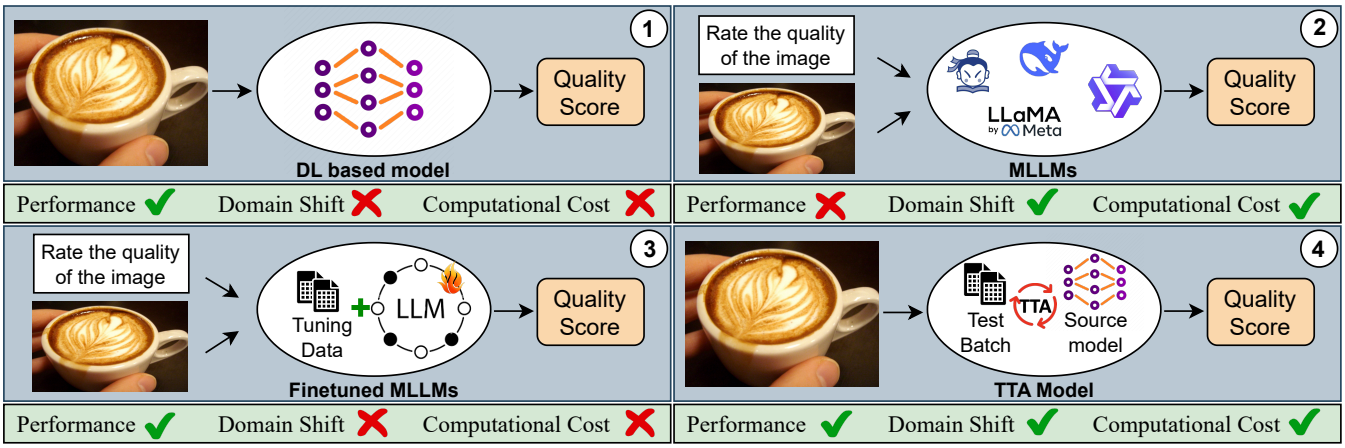

Figure 1: Comparison of four paradigms for No-Reference Image Quality Assessment (NRIQA). (1) Traditional deep learning-based models offer good performance but struggle under domain shift. (2) Zero-shot multi-modal Large Language Models (LLMs) are domain-robust but lack fine-grained quality perception. (3) Finetuned LLMs improve performance but incur high tuning costs. (4) Test-Time Adaptation (TTA) methods adapt pre-trained models on-the-fly using test batches, balancing performance, robustness, and efficiency.

Abstract

Recent test-time adaptation (TTA) methods for image quality assessment (IQA) aim to mitigate the distribution shift between training and testing data by adapting the batch normalization layers of the base IQA model. These methods typically leverage auxiliary tasks—such as group contrastive loss and rank loss—based on feature clustering of low- and high-quality images. However, existing approaches using base models or pre-trained VGG-16 networks struggle to generate meaningful quality-based clusters, limiting their effectiveness. In this work, we introduce LLM-TTA, a novel TTA framework that exploits the perceptual ranking capabilities of pre-trained multi-modal Large Language Models (LLMs).

The LLM-TTA algorithm is founded on the observation that, although pre-trained LLMs exhibit limited capability in predicting accurate absolute quality scores, they are highly effective in estimating relative quality rankings and generating more discriminative clusters—both of which are essential for the efficient computation of group contrastive loss. This simple yet impactful insight enables LLM-TTA to substantially improve the adaptability and performance of base IQA models. We validate our method on leading architectures (TReS, MUSIQ, MetaIQA, UIQA) across diverse natural (Koniq-10K, LIVE, PIPAL) and AI-generated (AGIQA-3K, PKU-AGIQA-4K, AIGCIQA2023) image quality assessment datasets. Experimental results demonstrate that LLM-TTA outperforms both existing TTA strategies and standalone LLMs, establishing a new state of the art in test-time IQA adaptation.

McGE '25, Dublin, Ireland

ACM ISBN 979-8-4007-2060-4/2025/10
https://doi.org/10.1145/3746278.3759394

CCS Concepts

• **Computing methodologies** → *Computer vision problems.*

Keywords

Image Quality Assessment, Large Language Models, Test Time Adaptation, Group Contrastive Loss.

ACM Reference Format:

Ambreen Bashir, Rahul Kumar, Vinit Jakhetiya, Badri N. Subudhi, Sunil Jaiswal, and Weisi Lin. 2025. LLM-TTA: Leveraging Large Language Models for Test-Time Adaptation in Image Quality Assessment. In *Proceedings of the 3rd International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice (McGE '25), October 27–28, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746278.3759394

1 Introduction

Image Quality Assessment (IQA) is a key task in image processing and computer vision, aiming to objectively predict the perceptual quality of images [18]. IQA methods are broadly classified into full reference (FR-IQA) and non-reference (NR-IQA) approaches. Although FR-IQA methods [29] rely on pristine reference images, NR-IQA methods estimate perceptual quality without any reference, making them more applicable in real-world scenarios such as social media, surveillance, and generative content evaluation [5].

Traditional NR-IQA approaches used hand-crafted features and classical regression models to predict image quality [8, 21]. In contrast, modern deep learning based methods have significantly improved NR-IQA performance by learning mappings from image content to quality scores using convolutional neural network [6, 19, 26, 48] or transformer-based [1, 31, 41] backbones. However, these models often suffer from performance drops under domain shifts, due to their reliance on specific training distributions. Unsupervised methods using self-supervised learning objectives like contrastive learning [20] or ranking-based loss [47] have been proposed to improve generalization but still lag behind supervised methods in complex scenarios.

To bridge this gap, Test-Time Adaptation (TTA) has emerged as a promising approach [23, 38], enabling models to adapt during inference using only unlabeled test data. This is particularly valuable for NR-IQA, where distributional changes caused by variations in capture devices, scene content, lighting conditions, context, or compression artifacts can significantly degrade performance. However, existing TTA strategies developed for classification tasks—such as self-training [30], correlative sampling [44], augmentation-based techniques [46], entropy minimization [33], and pseudo-label refinement [4] do not transfer directly to perceptual quality estimation. While contrastive learning-based methods like TTT++ [16] offer some adaptability, they are not designed to capture the fine-grained distinctions critical for IQA.

Despite recent advancements in TTA, its application to NR-IQA remains relatively unexplored. Roy et al. [27] proposed TTA-IQA, using group contrastive (GC) and rank losses to adapt NR-IQA models during inference. FA-TTA-IQA [11] extended this by using perceptual features from VGG-16 and selecting the highest-quality image in a batch as an anchor for contrastive learning. While promising, these methods rely on the base model's initial prediction, which may be incorrect.

Figure 1 illustrates the comparative limitations of existing IQA paradigms. Traditional deep learning based NR-IQA models perform well on seen distributions but fail under domain shifts. Recently, Large Language Models (LLMs), especially multimodal ones (MLLMs), have shown promise in visual tasks [34, 35], due to their zero-shot capabilities. However, while MLLMs handle domain shift well, their scoring accuracy is often limited. In this paper, the terms LLM and MLLM are used interchangeably, as we focus on LLMs with vision-language capabilities. Fine-tuning LLMs for IQA can improve performance [3, 32] but incurs high computational costs and does not fully resolve generalization issues. Meanwhile, TTA-based strategies [11, 27] offer a more efficient alternative, aiming to balance accuracy and adaptability.

Motivated by these insights, we propose LLM-TTA, a novel TTA framework that leverages the perceptual ranking capabilities of LLMs to improve the adaptability of NR-IQA models under distribution shifts. LLM-TTA uses LLMs to rank images in a test batch, enabling more reliable high-quality and low-quality group formation, which are later used for contrastive loss calculation. Importantly, LLM-TTA does not require any ground-truth labels or retraining, and is compatible with a variety of existing NR-IQA backbones. The major contributions of the proposed work are summarized as follows:

- We present the first framework that integrates LLMs with efficient TTA for NR-IQA.
- Our proposed method, LLM-TTA, is motivated by the observation that while MLLMs may struggle to produce accurate perceptual quality scores in a zero-shot setting, they excel at providing reliable relative quality rankings. These rankings are used for effective quality cluster formation in a group contrastive loss framework.
- We demonstrate the effectiveness of LLM-TTA across several backbones and datasets, consistently outperforming state-of-the-art TTA-based IQA methods.

2 Motivation

The TTA methods available in literature such as TTA-IQA [27] and FA-TTA-IQA [11] adapt NR-IQA models using group contrastive loss as an auxiliary objective to align batch normalization parameters with the test distribution. Specifically, TTA-IQA computes contrastive loss based on the base model's direct predictions, while

Table 1: A comparison of TTA-IQA, FA-TTA-IQA, and the proposed LLM-TTA across multiple datasets for the creation of high-quality and low-quality clusters, as well as the accuracy in identifying the image with the best perceptual quality.

Dataset	Model	TTA-IQA C/W	LLM-TTA C/W	FA-TTA Top-1 (%)	LLM-TTA Top-1 (%)
Koniq-10K	Meta	2296/2740	2906/2130	27.96	35.27
	UIQA	1841/3195	2906/2130	18.19	35.27
AIGC	Meta	439/761	654/546	19.67	32.33
	UIQA	555/645	654/546	25.67	32.33
SPAQ	Meta	2352/3216	3486/2078	33.91	43.68
	UIQA	1797/3771	3486/2078	29.09	43.68

Figure 2: An example batch of 8 test images from the Koniq-10K dataset [9] used for adapting MetaIQA [48]. 'MOS' denotes the ground-truth mean opinion scores, 'Baseline' shows predictions from the base MetaIQA model, and 'Qwen' displays the quality scores obtained using the pretrained Qwen [40]. Rankings that match the ground-truth MOS are highlighted in blue, while mismatches are shown in red.

FA-TTA-IQA selects the image with the highest perceptual quality, identified using the base model, and computes the contrastive loss using feature similarity between this image and the rest of the batch, with features extracted via a pre-trained VGG-16 network.

However, both methods rely heavily on the base model's ability to produce accurate perceptual rankings. Under significant distribution shifts, such as those introduced by AI-generated images, the base model often fails to distinguish between high-quality and low-quality images, leading to inaccurate contrastive loss estimation and poor adaptation. This is reflected in Table 5, where performance drops significantly on out-of-distribution data. Although FA-TTA-IQA generally outperforms TTA-IQA on natural images, its performance relies on the base model to select the image with the best perceptual quality in a batch, while the remaining samples in the cluster are chosen using features extracted from the pretrained VGG-16 network. At the same time, the performance of the FA-TTA-IQA algorithm is limited for AI-generated datasets because the features of VGG-16 are less perceptually relevant compared to those of natural images.

To quantitatively validate these points, Table 1 presents the frequency with which the base model correctly selects the top-2 and bottom-2 images (from a batch of 8) based on the ground-truth MOS, across various IQA datasets. The results reveal that the base model often misidentifies both high- and low-quality images, which diminishes the effectiveness of the group contrastive loss in TTA-IQA and results in suboptimal batch normalization updates. Similarly, FA-TTA-IQA depends on the base model's ability to correctly select the highest-quality image, a task it fails to perform consistently, thereby limiting adaptation performance. The fifth column of Table 1 shows the percentage of times the base model successfully identifies the highest-quality image.

This quantitative and qualitative analysis, illustrated in Figure 2, emphasizes the necessity for a novel algorithm that can efficiently classify images into high- and low-quality clusters. Such a classification is crucial for the precise computation of group contrastive loss, ensuring more accurate and reliable adaptation.

3 Proposed Algorithm

With the advancement of multimodal large language models (LLMs), these models have been increasingly applied to various image processing and computer vision tasks [45]. In the IQA domain, studies such as [36] and [39] indicate that pretrained LLMs can predict the perceptual quality of images to some extent. In this work, we observe that while zero-shot LLMs struggle to match ground-truth MOS scores, they excel in producing reliable relative quality rankings, which is crucial for TTA tasks. To substantiate this argument, the fourth and sixth columns of Table 1 report the performance of the pretrained Qwen model in (i) identifying the top 2 highest- and bottom 2 lowest-quality images, and (ii) accurately selecting the highest-quality image within a batch for constructing high- and low-quality clusters. Figure 2 also illustrates a batch of 8 images along with their perceptual rankings as determined by the base model and the Qwen LLM. These experiments reveal that the Qwen-based selection demonstrates significantly higher accuracy compared to the clustering-based approach used in TTA-IQA and

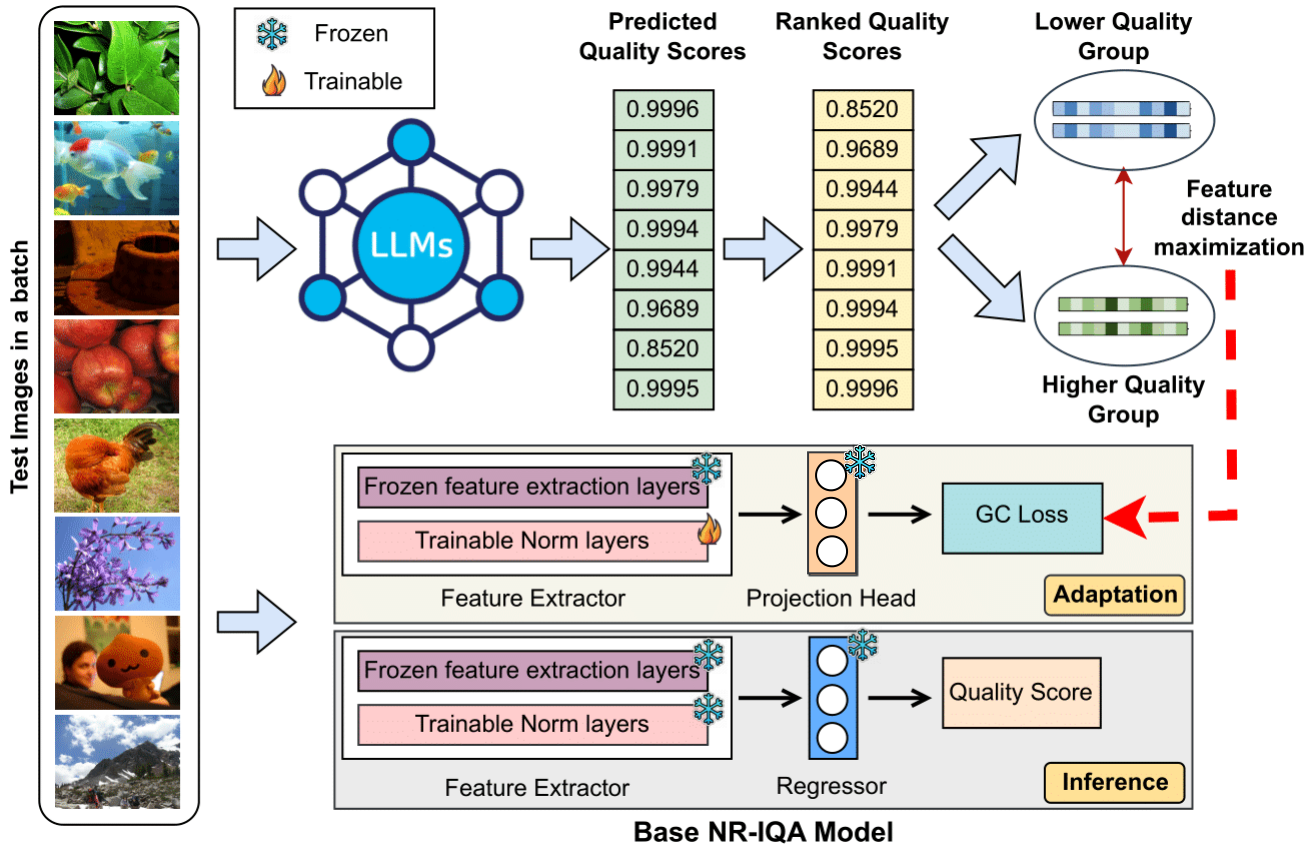

Figure 3: Overview of the proposed LLM-TTA framework: Test images are processed by pretrained LLMs to obtain perceptual rankings, which are used to compute group contrastive loss and adapt the batch/layer normalization parameters of the base IQA model.

FA-TTA-IQA with the base model, enabling more reliable group formation for contrastive learning.

In general, a deep learning-based IQA pipeline consists of a feature extractor to capture quality-related features, followed by a quality regressor that fuses these features to predict the final quality score. In TTA-IQA [27], only batch normalization layers are updated using auxiliary tasks such as group contrastive loss and rank loss. The group contrastive loss aims to maximize the feature distance between high-quality and low-quality image groups, thereby enhancing the model's discriminative capability. In parallel, the rank loss enforces the model to distinguish between high-quality reference images and their corresponding degraded (noisy) counterparts. In the proposed LLM-TTA algorithm, only group contrastive loss is employed. This design choice is motivated by advancements in sensor technology, which have led to the acquisition of generally high-quality images. Consequently, the base IQA models are typically capable of distinguishing between clean and noisy versions of the same image, reducing the necessity for an additional rank loss.

The overall pipeline of the proposed LLM-TTA framework is illustrated in Figure 3. During adaptation, the parameters of the layer normalization and batch normalization layers within the feature extractor (Θ_e) are updated. The projection head (Θ_p) maps the extracted features to a lower-dimensional space. Based on LLM-guided grouping, the following loss function is used to adapt the parameters of the feature extractor (Θ_e) through group contrastive (GC) learning.

$$\Theta_e^* = \arg\min_{\Theta_e} \mathcal{L}_{\text{LLM-GC}}(B) \tag{1}$$

where B is the batch size of images. Each image in a batch of B is individually fed into the LLM with a fixed prompt (i.e., "The quality of the image is"). The effect of using different fixed prompts on the performance of the proposed LLM-TTA algorithm is analyzed in Subsection 4.3. For each image, the output logits corresponding to the score tokens good ($\text{I}_{\text{Logit}(good)}$) and poor ($\text{I}_{\text{Logit}(poor)}$) are extracted from the full set of logits. The final quality score of I^{th} image is then obtained by applying the softmax function to these logits, as described in [36]:

$$\text{S}_\text{I} = \frac{\exp(\text{I}_{\text{Logit}(good)})}{\exp(\text{I}_{\text{Logit}(good)}) + \exp(\text{I}_{\text{Logit}(poor)})} \tag{2}$$

After computing the quality scores $S_{I_1}, S_{I_2}, \ldots, S_{I_B}$ for each image $I_1, I_2, \ldots, I_B$ in a batch of size B, the images are sorted based on their scores. The two images with the lowest scores and the two

with the highest scores are selected to form the low-quality and high-quality clusters, respectively. Table 1 and Figure 2 illustrate the effectiveness of the pretrained Qwen LLM in guiding the clustering process. It is evident that the proposed LLM-based clustering strategy for group contrastive loss outperforms the conventional clustering mechanisms employed in TTA-IQA and FA-TTA-IQA. Once the high-quality and low-quality clusters are formed, the group contrastive loss is applied to minimize the feature distances among images within the same cluster (i.e., intra-cluster similarity) and maximize the distances between images from different clusters (i.e., inter-cluster dissimilarity). The proposed group contrastive loss ($\mathcal{L}_{\text{LLM-GC}}$) for i^{th} sample is calculated as:

$$\mathcal{L}_{\text{GCL}} = -\log \frac{\exp\left(\text{sim}(z_i, z_j)/\tau\right)}{\sum_{(k)\in A(i)} \exp\left(\text{sim}(z_i, z_k)/\tau\right)} \tag{3}$$

Here, $z_i = \Theta_p(\Theta_e(I_i))$ and $z_j = \Theta_p(\Theta_e(I_j))$ represent the projected feature vectors of image samples I_i and I_j, which belong to the same quality cluster. The function $\text{sim}(\cdot,\cdot)$ denotes the similarity between these features, computed using cosine similarity. $z_k = \Theta_p(\Theta_e(I_k))$ corresponds to the projected features of samples I_k from the opposite (negative) cluster. $A(i)$ denotes the set of all such contrasting samples for anchor i, and τ is the temperature parameter that controls the sharpness of the softmax distribution. A similar $\mathcal{L}_{\text{GCL}}$ is computed for each sample within the high-quality and low-quality groups, and the final group contrastive loss is obtained by averaging these individual losses.

It is also noteworthy that, during the ranking process within a batch, images are not simultaneously fed to the LLM for direct relative ranking (i.e., multi-stimulus setting). Instead, each image is evaluated individually in a single-stimulus manner, and the relative ranking is derived based on the predicted quality scores. As highlighted in [39], the single-stimulus approach has been shown to yield more reliable and consistent quality assessments compared to the multi-stimulus setting.

4 Experimental Results

The proposed LLM-TTA algorithm is applied to adapt several NR-IQA models, including TReS [6], MUSIQ [12], MetaIQA [48], and UIQA [31], across a diverse set of IQA datasets. These datasets encompass both natural image quality assessment datasets, such as LIVE [28], PIPAL [10], and Koniq-10K [9], as well as AI-generated image quality assessment datasets like AGIQA-3K [13], PKU-AGIQA-4K [43], and AIGCIQA [24]. Detailed descriptions of the IQA datasets and models are provided in Tables 2 and 3, respectively. The proposed LLM-TTA algorithm is exclusively compared with existing TTA-based methods, namely TTA-IQA [27] and FA-TTA-IQA [11], as, to the best of the authors' knowledge, no other TTA techniques have been introduced in the literature for the IQA task.

4.1 Implementation Details

All experiments are carried out on a Tesla V100-PCIE-32GB GPU using PyTorch 2.4.0 and CUDA 11.8.2 in a Linux machine. To ensure reproducibility, the random seed is set to 2021, and a consistent batch size of 8 is maintained. During the TTA phase, all parameters of the base model are kept frozen, except for the layer normalization and batch normalization layers, which are adapted using the proposed group contrastive loss ($\mathcal{L}_{\text{GCL}}$). The adaptation is performed using the Adam optimizer with a learning rate of 0.001. Each batch undergoes 10 adaptation iterations; more iterations are avoided to prevent the base model's learned representations from being overwritten. For every batch of test images, the original source model is reloaded and adapted independently.

Table 2: Details of the datasets used for the evaluation of the proposed algorithm in comparison with existing TTA algorithms.

Dataset	Description
LIVE [28]	779 distorted images affected by compression artifacts, noise contamination, and other common visual degradations.
PIPAL [10]	1,130 images containing restoration-related distortions generated by deep learning-based models.
Koniq-10K [9]	10,073 authentically distorted images collected from the web, exhibiting a wide range of real-world distortions such as blur, noise, compression artifacts, and exposure issues.
AGIQA-3K [13]	3,000 AI-generated images with diverse prompts and distortion types.
PKU-AGIQA-4K [43]	4,000 AI-generated images sourced from various generative models such as Stable Diffusion, DALL·E, and Midjourney.
AIGCIQA [24]	1,200 AI-generated images from various models and prompt styles.

4.2 Performance Comparison

The evaluation is conducted using two standard metrics: Spearman's Rank Correlation Coefficient (SRCC) for ranking consistency and Pearson's Linear Correlation Coefficient (PLCC) for assessing prediction linearity.

Table 4 reports the performance of the proposed LLM-TTA framework, where different LLMs, including InternLM-XComposer2-VL (ILM) [2], LLaVA-v1.5-13B (LLaVa) [15], Deepseek-VL-7B-base (DS) [17], Llama-3.2-11B-Vision (Llama) [7], and Qwen2.5-VL-7B-Instruct (Qwen) [40], are employed to compute group contrastive loss and adapt the normalization layers for natural image quality assessment. The table also presents the performance of the pre-trained LLMs, along with a comparative analysis against state-of-the-art TTA methods, namely TTA-IQA [27] and FA-TTA-IQA [11]. It is

Table 3: Details of NR-IQA models adapted using test time adaptation methods.

Model	Training Dataset
MetaIQA [48]	Trained on TID2013 [25] and KADID-10K dataset [14].
TReS [6]	Trained on LIVEFB database [42].
MUSIQ [12]	Trained on LIVEFB database [42].
UIQA [31]	Trained on AVA database [22].

Table 4: Performance comparison of the proposed LLM-TTA-IQA with pretrained LLMs and existing TTA methods on multiple NR-IQA backbones across natural image datasets. Higher values denote better performance; the best results are highlighted in blue, and the second-best in red. Here, Proposed$_{\text{ILM}}$ represents the proposed LLM-TTA algorithm leveraging the internLM.

Backbone	Datasets	Koniq-10k		LIVE		PIPAL		Average	
	Method	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Pretrained LLMs	ILM [2]	0.4432	0.4362	0.4556	0.4534	0.1957	0.1972	0.3648	0.3623
	LLaVa [15]	0.1117	0.0992	0.2625	0.2733	0.0220	0.0167	0.1321	0.1297
	Deepseek [17]	0.2864	0.3028	0.0026	0.0155	0.2084	0.2160	0.1658	0.1781
	LLaMa [7]	0.1554	0.1996	0.4628	0.4659	0.1430	0.1452	0.2537	0.2702
	Qwen [40]	0.6387	0.5931	0.6667	0.5356	0.4195	0.4425	0.5750	0.5237
TReS	Baseline	0.6690	0.6920	0.5434	0.4450	0.3624	0.3592	0.5249	0.4987
	TTA-IQA [27]	0.6578	0.7074	0.6722	0.5963	0.4278	0.4204	0.5859	0.5747
	FA-TTA-IQA [11]	0.6917	**0.7381**	0.6922	0.6309	**0.4729**	**0.4765**	0.6189	0.6152
	Proposed$_{\text{ILM}}$	0.6432	0.6584	0.4544	0.4328	0.4448	0.4490	0.5141	0.5134
	Proposed$_{\text{LLaVa}}$	0.2640	0.2791	0.3951	0.4218	0.4258	0.4419	0.3616	0.3809
	Proposed$_{\text{DS}}$	0.5049	0.5283	0.4396	0.4832	0.4222	0.4067	0.4556	0.4727
	Proposed$_{\text{Llama}}$	0.4209	0.4624	**0.8005**	0.7164	0.4253	0.4325	0.5489	0.5371
	Proposed$_{\text{Qwen}}$	**0.7005**	0.7309	0.7843	**0.7202**	0.4506	0.4502	**0.6451**	**0.6338**
MUSIQ	Baseline	0.6452	0.6816	0.2685	0.3393	0.3407	0.3307	0.4181	0.4505
	TTA-IQA [27]	0.6693	0.7230	0.3649	0.4031	0.3743	0.3731	0.4695	0.4997
	FA-TTA-IQA [11]	0.6711	0.7263	0.3561	0.4063	0.3738	0.3744	0.4670	0.5023
	Proposed$_{\text{ILM}}$	**0.6907**	**0.7373**	0.4201	0.4257	0.3767	0.3737	0.4958	0.5122
	Proposed$_{\text{LLaVa}}$	0.6652	0.7191	0.3966	0.4174	0.3756	0.3725	0.4791	0.5030
	Proposed$_{\text{DS}}$	0.6751	0.7250	0.3818	0.4104	0.3765	0.3752	0.4778	0.5035
	Proposed$_{\text{Llama}}$	0.6652	0.7189	0.3891	0.4156	0.3785	0.3778	0.4776	0.5041
	Proposed$_{\text{Qwen}}$	0.6816	0.7316	**0.4552**	**0.4409**	**0.3815**	**0.3805**	**0.5061**	**0.5177**
MetaIQA	Baseline	0.5209	0.4760	0.7323	0.6732	0.3441	0.3168	0.5324	0.4887
	TTA-IQA [27]	0.5549	0.5329	0.7899	0.7646	0.3524	0.3319	0.5657	0.5431
	FA-TTA-IQA [11]	0.5735	0.5529	0.8521	0.8326	0.3637	0.3424	0.5964	0.5760
	Proposed$_{\text{ILM}}$	0.6294	0.5907	0.8626	0.8499	**0.4476**	0.3907	**0.6465**	**0.6104**
	Proposed$_{\text{LLaVa}}$	0.4994	0.4806	0.7628	0.7316	0.3888	0.3469	0.5503	0.5197
	Proposed$_{\text{DS}}$	0.4941	0.4662	0.7731	0.7464	0.2441	0.2301	0.5038	0.4809
	Proposed$_{\text{Llama}}$	0.4711	0.4436	**0.8731**	**0.8510**	0.4077	0.3721	0.5840	0.5556
	Proposed$_{\text{Qwen}}$	**0.6539**	**0.5979**	0.8405	0.8227	0.4420	**0.3925**	0.6455	0.6044
UIQA	Baseline	0.7515	0.7468	0.6956	0.6654	0.1705	0.1668	0.5392	0.5263
	TTA-IQA [27]	0.7648	0.7573	0.7024	0.6717	0.2233	0.2170	0.5635	0.5487
	FA-TTA-IQA [11]	0.7294	0.7274	0.6669	0.6180	0.1943	0.1890	0.5302	0.5115
	Proposed$_{\text{ILM}}$	0.7421	0.7225	0.6936	0.6628	0.3436	0.3413	0.5931	0.5755
	Proposed$_{\text{LLaVa}}$	0.6320	0.6461	0.4682	0.4451	0.2804	0.2773	0.4602	0.4562
	Proposed$_{\text{DS}}$	0.6840	0.6985	0.4185	0.4129	0.1762	0.1658	0.4262	0.4257
	Proposed$_{\text{Llama}}$	0.6959	0.6969	0.6784	0.6495	0.1707	0.1640	0.5150	0.5035
	Proposed$_{\text{Qwen}}$	**0.7937**	**0.7872**	**0.7402**	**0.7073**	**0.3650**	**0.3722**	**0.6330**	**0.6222**

evident from the results that the proposed LLM-TTA algorithm consistently outperforms both TTA-IQA and FA-TTA-IQA across multiple benchmarks. Furthermore, among the evaluated LLMs, Qwen demonstrates the best standalone performance, and its integration with the proposed LLM-TTA framework also yields the best results. These experimental results are in line with previous studies [36, 37, 39]. Interestingly, although some LLMs do not exhibit strong performance individually, their integration within the LLM-TTA framework significantly improves their effectiveness, often surpassing the base IQA algorithm in most scenarios.

Similarly, Table 5 reports the performance of pretrained LLMs, existing TTA methods (TTA-IQA [27] and FA-TTA-IQA [11]) and the proposed LLM-TTA framework across AI-generated IQA datasets: AGIQA-3K [13], PKU-AGIQA-4K [43], and AIGCIQA2023 [24], using various pretrained NRIQA as backbones. The experimental results demonstrate that, despite the base models not being explicitly trained on AI-generated content, the proposed LLM-TTA

Table 5: Performance comparison of the proposed LLM-TTA-IQA with pretrained LLMs and existing TTA methods on multiple NR-IQA backbones across AI-generated datasets. Higher values indicate better performance; best results are marked in blue, and second-best in red.

Backbone	Datasets	AGIQA-3K		PKU-AGIQA-4K		AIGCIQA2023		Average	
	Model	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Pretrained LLMs	ILM [2]	0.4590	0.5075	0.0402	0.0323	0.3118	0.3138	0.2703	0.2845
	LLaVa [15]	0.2431	0.2369	0.1740	0.1675	0.1126	0.1027	0.1766	0.1690
	Deepseek [7]	0.2395	0.2553	0.0250	0.0221	0.1493	0.1346	0.1379	0.1373
	LLaMa [17]	0.4992	0.5122	0.1482	0.1824	0.4808	0.4284	0.3761	0.3743
	Qwen [40]	0.6169	0.6525	0.1444	0.0571	0.7106	0.5304	0.4906	0.4133
TReS [6]	Baseline	0.5052	0.5285	0.0171	0.0132	0.5864	0.5370	0.3696	0.3596
	TTA-IQA [27]	0.5209	0.5780	0.0251	0.0182	0.5958	0.5971	0.3806	0.3978
	FA-TTA-IQA [11]	0.5092	0.5859	0.0692	0.0616	0.5503	0.5339	0.3762	0.3938
	$\text{Proposed}_{\text{ILM}}$	0.4967	0.5318	0.0081	0.0234	0.5132	0.5109	0.3393	0.3554
	$\text{Proposed}_{\text{LLaVa}}$	0.4597	0.5017	0.0999	0.1052	0.3926	0.3944	0.3174	0.3338
	$\text{Proposed}_{\text{DS}}$	0.5934	**0.6748**	0.0713	0.0658	0.6440	0.6494	0.4362	0.4633
	$\text{Proposed}_{\text{Llama}}$	0.4836	0.5526	0.0249	0.0267	0.5009	0.4642	0.3365	0.3478
	$\text{Proposed}_{\text{Qwen}}$	**0.6029**	0.6685	**0.2625**	**0.2496**	**0.6852**	**0.6946**	**0.5169**	**0.5376**
MUSIQ [12]	Baseline	0.4435	0.5686	0.0524	0.0712	0.6024	0.5810	0.3661	0.4069
	TTA-IQA [27]	0.4596	0.5852	0.0782	0.0854	0.6214	0.5882	0.3864	0.4196
	FA-TTA-IQA [11]	0.4470	0.5797	0.0772	0.0882	0.6003	0.5735	0.3748	0.4138
	$\text{Proposed}_{\text{ILM}}$	**0.4643**	0.5861	0.0743	0.0808	0.6176	0.5817	0.3854	0.4162
	$\text{Proposed}_{\text{LLaVa}}$	0.4536	0.5794	0.0702	0.0763	0.6069	0.5749	0.3769	0.4102
	$\text{Proposed}_{\text{DS}}$	0.4601	0.5863	0.0774	0.0828	0.6213	0.5877	0.3863	0.4189
	$\text{Proposed}_{\text{Llama}}$	0.4591	0.5834	0.0789	0.0854	0.6060	0.5717	0.3813	0.4135
	$\text{Proposed}_{\text{Qwen}}$	0.4626	**0.5879**	**0.0953**	**0.0995**	**0.6272**	**0.5916**	**0.3950**	**0.4263**
MetaIQA [48]	Baseline	0.5119	0.4886	0.1397	0.1329	0.4504	0.4139	0.3673	0.3451
	TTA-IQA [27]	0.5342	0.5279	0.1263	0.1269	0.4612	0.4470	0.3739	0.3673
	FA-TTA-IQA [11]	0.5336	0.5168	0.1261	0.1241	0.4768	0.4606	0.3788	0.3672
	$\text{Proposed}_{\text{ILM}}$	0.5333	0.4823	0.0981	0.1054	0.4888	0.4687	0.3734	0.3521
	$\text{Proposed}_{\text{LLaVa}}$	0.5454	0.4793	0.1482	0.1447	0.4243	0.4007	0.3726	0.3416
	$\text{Proposed}_{\text{DS}}$	0.5327	0.5039	0.1309	0.1315	0.5231	0.4662	0.3956	0.3672
	$\text{Proposed}_{\text{Llama}}$	0.5168	0.5345	0.0874	0.0913	0.4684	0.4334	0.3575	0.3531
	$\text{Proposed}_{\text{Qwen}}$	**0.5689**	**0.5452**	**0.2334**	**0.2219**	**0.5897**	**0.5451**	**0.4640**	**0.4374**
UIQA [31]	Baseline	0.6181	0.7009	0.2040	0.2307	0.6566	0.6733	0.4929	0.5350
	TTA-IQA [27]	0.6322	0.6932	**0.2362**	**0.2665**	0.6777	0.7016	0.5154	0.5538
	FA-TTA-IQA [11]	0.6182	0.6945	0.2127	0.2372	0.6399	0.6520	0.4903	0.5279
	$\text{Proposed}_{\text{ILM}}$	0.6352	0.7201	0.1938	0.2202	0.6673	0.6850	0.4988	0.5418
	$\text{Proposed}_{\text{LLaVa}}$	0.5285	0.5786	0.2206	0.2453	0.5509	0.5549	0.4333	0.4596
	$\text{Proposed}_{\text{DS}}$	0.6338	0.7229	0.2236	0.2452	0.7147	0.7454	0.5240	0.5712
	$\text{Proposed}_{\text{Llama}}$	0.6027	0.6499	0.2000	0.2190	0.6217	0.6326	0.4748	0.5005
	$\text{Proposed}_{\text{Qwen}}$	**0.6529**	**0.7479**	0.2178	0.2281	**0.7236**	**0.7588**	**0.5314**	**0.5783**

method consistently enhances their performance across these challenging datasets. Additionally, TTA-IQA outperforms FA-TTA-IQA, as the latter relies on feature similarity using a pretrained VGG-16 architecture, which is originally trained on natural images and therefore struggles to capture the relevant perceptual features required for assessing AI-generated images. On average, the proposed LLM-TTA achieves PLCC improvements of 13.52%, 9.77%, and 7.85% over pretrained Qwen, TTA-IQA, and FA-TTA-IQA, respectively, on natural image quality assessment datasets. For AI-generated image datasets, it achieves even greater gains—19.73%, 13.87%, and 16.26%, respectively.

These experimental results clearly validate that the proposed LLM-TTA algorithm outperforms both the existing IQA-TTA algorithms and the base IQA models. Integrating LLMs with TTA significantly boosts performance compared to both the base models and the existing TTA-IQA approaches. From this perspective, further advances in LLMs tailored for low-level vision tasks are

Table 6: Performance comparison of the proposed LLM-TTA algorithm using different prompts (Prompt1, Prompt2, Prompt3) on the LIVE and AIGC datasets.

	Model	P1		P2		P3	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
LIVE [28]	Score	0.591	0.519	0.667	0.536	0.621	0.643
	MetaIQA [48]	0.876	0.862	0.841	0.823	0.865	0.850
	MuSIQ [12]	0.456	0.440	0.455	0.441	0.463	0.444
	TReS [6]	0.756	0.699	0.784	0.720	0.667	0.611
	UIQA [31]	0.723	0.689	0.740	0.707	0.744	0.710
AIGC [24]	Score	0.673	0.499	0.711	0.530	0.698	0.668
	MetaIQA [48]	0.590	0.546	0.589	0.545	0.417	0.413
	MuSIQ [12]	0.625	0.591	0.627	0.591	0.623	0.589
	TReS [6]	0.674	0.676	0.685	0.695	0.679	0.691
	UIQA [31]	0.722	0.757	0.723	0.759	0.727	0.763

expected to complement and enhance the effectiveness of the proposed LLM-TTA framework.

4.3 Ablation Studies

The input prompts play a crucial role in influencing the performance of LLMs and, consequently, the effectiveness of the proposed LLM-TTA framework. To investigate this, we experimented with three different prompt formulations:

(1) Prompt1: **Rate the quality of the image**, followed by applying softmax over the logits corresponding to 'good' and 'poor' tokens;
(2) Prompt2: **The quality of the image is**, followed by softmax of the logits for 'good' and 'poor' tokens as suggested in [36];
(3) Prompt3: **Rate the quality of the image in terms of sharpness, noise, distortion, and clarity. Provide a quality score from 1 (bad) to 10 (good) only. Format: Score: x**, as proposed in [39].

Table 6 presents the results of an ablation study evaluating the impact of different prompts used with the Qwen LLM for generating clusters in the proposed LLM-TTA algorithm on the LIVE and AIGC datasets. The first row shows the PLCC and SRCC values of the quality predictions made by the pretrained LLM, Qwen, under each prompt. The subsequent rows report the performance of various IQA models after adaptation using the LLM-TTA framework. AAmong the evaluated prompts, Prompt 2 demonstrates the highest performance, aligning with the findings previously reported in [36]. Based on this observation, all the experimental results presented in Tables 4 and 5 are obtained using Prompt 2.

The scatter plots in Fig. 4 illustrate the relationship between MOS and the predicted scores for the base MetaIQA model, as well as its adapted versions using TTA-IQA, FA-TTA-IQA, and the proposed LLM-TTA algorithm. It is evident that the predictions from LLM-TTA align more closely with the regression line than other methods. These results demonstrate the effectiveness of the proposed approach and highlight the advantages of incorporating LLMs into the TTA pipeline.

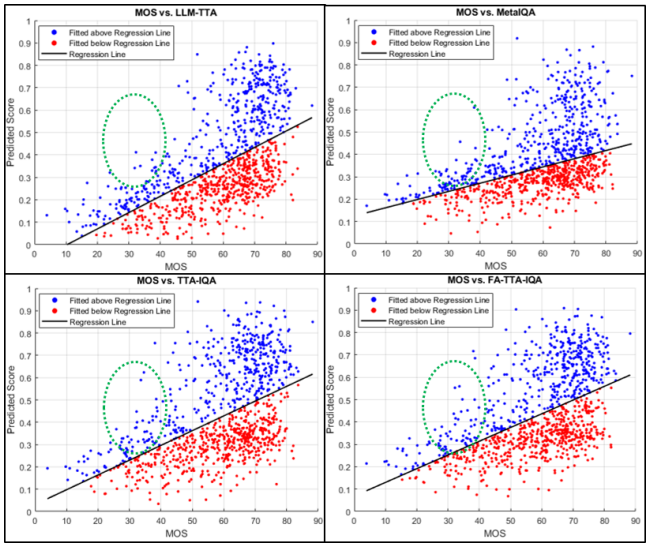

Figure 4: Scatter plots comparing the proposed LLM-TTA with the baseline MetaIQA model, TTA-IQA, and FA-TTA-IQA methods on the KONIQ-10K dataset [9].

5 Conclusions and Future Work

In this paper, we proposed a novel TTA algorithm for NR-IQA that leverages the capabilities of pretrained LLMs. The key insight behind the proposed LLM-TTA algorithm is that while pre-trained LLMs may struggle to provide accurate absolute quality scores, they are effective in generating consistent relative quality rankings of images, a property that is highly beneficial for TTA. Based on this, our approach employs LLM-based clustering to guide the adaptation of NR-IQA models through a group contrastive loss as an auxiliary task, eliminating the need for ground-truth labels at test time. Extensive experiments on both natural and AI-generated datasets show that LLM-TTA significantly boosts the performance of various existing NR-IQA models. These results highlight the promising role of LLMs as auxiliary tools in improving test-time adaptation for perceptual quality assessment. Furthermore, the proposed framework opens up opportunities for extending this concept to related image restoration tasks such as super-resolution and denoising.

One of the key open challenges in the domain of test-time adaptation is the development of a reliable early-stopping criterion. We hypothesize that large language models can contribute to solving this problem, for example, by terminating the adaptation process when the clustering results produced by the base model converge with those generated by the LLM. We plan to investigate this direction in future work. Another area of future work involves enhancing the clustering of low- and high-quality images by ensembling multiple LLMs.

Acknowledgments

This work utilized virtual machine resources provided by the Indian Institute of Technology Jammu.

References

[1] Nasim Jamshidi Avanaki, Abhijay Ghildyal, Nabajeet Barman, and Saman Zadtootaghaj. 2024. LAR-IQA: A Lightweight, Accurate, and Robust No-Reference Image Quality Assessment Model. *arXiv preprint arXiv:2408.17057* (2024).

[2] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297* (2024).

[3] Chaofeng Chen, Sensen Yang, Haoning Wu, Liang Liao, Zicheng Zhang, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2024. Q-ground: Image quality grounding with large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 486–495.

[4] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. 2022. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 295–305.

[5] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. 2020. Assessing image quality issues for real-world problems. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3646–3656.

[6] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. 2022. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1220–1230.

[7] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[8] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. 2014. Hybrid no-reference quality metric for singly and multiply distorted images. *IEEE Transactions on Broadcasting* 60, 3 (2014), 555–567.

[9] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing* 29 (2020), 4041–4056.

[10] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. 2020. 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, 633–651.

[11] Meghna Kapoor, Vinit Jakhetiya, Badri Subudhi, Ankur Bansal, and Weisi Lin. 2025. Feature Affinity based Clustering for Test-Time Adaptation for Image Quality Assessment. In *Proceedings of the IEEE International Conference on Multimedia and Expo*.

[12] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5148–5157.

[13] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. 2023. Agiqa-3k: An open database for ai-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 8 (2023), 6833–6846.

[14] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. 2019. KADID-10k: A large-scale artificially distorted IQA database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–3.

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

[16] Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. 2021. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems* 34 (2021), 21808–21820.

[17] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. DeepSeek-VL: Towards Real-World Vision-Language Understanding. arXiv:2403.05525 [cs.AI]

[18] Kede Ma and Yuming Fang. 2021. Image quality assessment in the modern age. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5664–5666.

[19] Xiaoyu Ma, Yaqi Wang, Chang Liu, Suiyu Zhang, and Dingguo Yu. 2022. ADGNet: Attention discrepancy guided deep neural network for blind image quality assessment. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1309–1318.

[20] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. 2022. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing* 31 (2022), 4149–4161.

[21] Anush Krishna Moorthy and Alan Conrad Bovik. 2011. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing* 20, 12 (2011), 3350–3364.

[22] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2408–2415.

[23] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. 2023. Towards Stable Test-Time Adaptation in Dynamic Wild World. In *Internetional Conference on Learning Representations*.

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

[25] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal processing: Image communication* 30 (2015), 57–77.

[26] Siddharth Roheda, Amit Satish Unde, Loay Rashid, and Abhishek Ameta. 2023. Degradation Aware Multi-Scale Approach to No Reference Image Quality Assessment. In *Proceedings of the Fourteenth Indian Conference on Computer Vision, Graphics and Image Processing*. 1–9.

[27] Subhadeep Roy, Shankhanil Mitra, Soma Biswas, and Rajiv Soundararajan. 2023. Test time adaptation for blind image quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16742–16751.

[28] H Sheikh. 2005. LIVE image quality assessment database release 2. *http://live. ece. utexas. edu/research/quality* (2005).

[29] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing* 15, 11 (2006), 3440–3451.

[30] Yongyi Su, Xun Xu, and Kui Jia. 2024. Towards real-world test-time adaptation: Tri-net self-training with balanced normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 15126–15135.

[31] Wei Sun, Weixia Zhang, Yuqin Cao, Linhan Cao, Jun Jia, Zijian Chen, Zicheng Zhang, Xiongkuo Min, and Guangtao Zhai. 2024. Assessing UHD Image Quality from Aesthetics, Distortions, and Saliency. *arXiv preprint arXiv:2409.00749* (2024).

[32] Sanjot Sagar Totade, Nithin C Babu, Shika Rao, and Rajiv Soundararajan. 2024. Internal Embeddings of Multi-modal LLMs as Generalizable Representations for Image Quality Assessment. In *Proceedings of the Fifteenth Indian Conference on Computer Vision Graphics and Image Processing*. 1–9.

[33] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726* (2020).

[34] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 2555–2563.

[35] Puyi Wang, Wei Sun, Zicheng Zhang, Jun Jia, Yanwei Jiang, Zhichao Zhang, Xiongkuo Min, and Guangtao Zhai. 2024. Large multi-modality model assisted ai-generated image quality assessment. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 7803–7812.

[36] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. 2023. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181* (2023).

[37] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. 2024. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 25490–25500.

[38] Haihang Wu and Bohan Zhuang. 2024. Fast and Accurate Continual Test Time Domain Adaptation. In *Proceedings of the 1st on Continual Learning meets Multimodal Foundation Models: Fundamentals and Advances*. 14–22.

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

[40] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024).

[41] Junfeng Yang, Jing Fu, Zhen Zhang, Limei Liu, Qin Li, Wei Zhang, and Wenzhi Cao. 2024. Align-IQA: aligning image quality assessment models with diverse human preferences via customizable guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 10008–10017.

[42] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. 2020. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3575–3585.

[43] Jiquan Yuan, Fanyi Yang, Jihe Li, Xinyan Cao, Jinming Che, Jinlong Lin, and Xixin Cao. 2024. PKU-AIGIQA-4K: A Perceptual Quality Assessment Database for Both Text-to-Image and Image-to-Image AI-Generated Images. *arXiv preprint arXiv:2404.18409* (2024).

[44] Longhui Yuan, Binhui Xie, and Shuang Li. 2023. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition. 15922–15932.

[45] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[46] Marvin Zhang, Sergey Levine, and Chelsea Finn. 2022. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems* 35 (2022), 38629–38642.

[47] Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. 2023. Quality-aware pre-trained models for blind image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22302–22313.

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