Learning from Model Rankings Improves Blind Super-Resolution Image Quality Assessment

Junlin Chen[1], Peibei Cao[2], Guangtao Zhai[1], Xiaokang Yang[1], Weixia Zhang[1*]

***Abstract*— Image super-resolution (SR) aims to generate a high-resolution (HR) image from a low-resolution (LR) input. Traditionally, full-reference image quality assessment (FR-IQA) models have been widely used to evaluate the perceptual quality of super-resolved images, relying on pristine reference images as the gold standard. However, in real-world SR applications, such reference images are often unavailable, posing challenges for the use of FR-IQA. While blind image quality assessment (BIQA) models can assess the perceptual quality of super-resolved images without requiring a reference, there remains a lack of comprehensive studies evaluating the effectiveness of existing BIQA models for real-world SR tasks. This dilemma can largely be attributed to the high cost of subjective testing required to collect sufficient human quality annotations, which in turn hinders the development of effective SR-IQA models. In this study, we tackle this challenge with a data-efficient approach. We first generate super-resolved images from LR inputs using state-of-the-art real-world SR methods. Then, we use the maximum differentiation competition (MAD) to select a diverse set of images for subjective testing, allowing us to efficiently gather human preferences and assess the alignment between BIQA model predictions and human judgments. The resulting global ranking of SR methods not only indicates the relative performance of recent real-world SR models, but also gives us an opportunity to develop a new BIQA model tailored for real-world SR-IQA. By utilizing the global rankings of SR algorithms as prior knowledge, we can refine pretrained BIQA models using vast amounts of super-resolved images without any supervisory signal. Experimental results show that our approach substantially enhances IQA performance for real-world SR while preserving robust predictive accuracy across various distortion scenarios. The dataset and the code are available at `https://github.com/cschenjunlin/SR-IQA-SMC25`.**

I. INTRODUCTION

Image super-resolution (SR) aims at reconstructing a high-resolution (HR) image from its low-resolution (LR) counterpart, which is inherently challenging due to its ill-posed nature, particularly when dealing with large scaling factors (*e.g.*, $\times 4$ SR). Classical SR algorithms either exploit internal similarities within an image [1], [2], or learn LR-to-HR mapping functions from external paired images [2], [3]. Since the advent of SRCNN [4], deep neural networks (DNNs) have increasingly dominated the field of SR [5]–[8]. Early DNN-based methods formulate the SR task by assuming simple and known degradation models (*e.g.*, bicubic interpolation), some of which also leverage generative adversarial networks (GANs) to enhance the visual quality of super-resolved images [9], [10]. Despite the significant successes of these methods, they struggle to handle real-world SR where the degradation model is more complex or even unknown. To better simulate real-world degradation, BSRGAN [11] uses a degradation model consisting of randomly shuffled blur, downsampling, and noise operations. Real-ESRGAN [12] further enhances this approach by employing a high-order degradation model. LDL [13] enhances GAN-based methods by generating an artifact map to stabilize the training process.

Recent years have witnessed the emergence of a new paradigm that leverages diffusion models [14] for the real-world SR task. StableSR [15] harnesses diffusion priors from pre-trained text-to-image models, achieving remarkable SR results through a time-aware encoder, fidelity-controllable module, and progressive sampling. DiffBIR [16] decouples the real-world SR problem into a regression-based content reconstruction stage and a diffusion-based details enhancement stage. SeeSR [17] utilizes a degradation-aware model to extract more reliable semantic information from LR images, guiding a diffusion model to generate rich and semantically accurate details. ResShift [18] enhances generation efficiency by modeling the residual shift between LR and HR images using a Markov chain. SinSR [19] further accelerates ResShift by distilling it into a one-step sampling model.

Image quality assessment (IQA) has been widely used to evaluate the quality of super-resolved images, which is generally categorized into subjective IQA and objective IQA. Subjective IQA involves human participants rating the quality of images through formal testing procedures. Although subjective IQA is the most reliable method, it is limited by its high labor and time costs. Objective IQA acts as a proxy for subjective IQA, quantifying image quality using computational models. Full-reference IQA (FR-IQA) metrics (*e.g.*, PSNR and SSIM [20]), which assess the quality of a test image by comparing it to a pristine reference image, have long been the *de facto* standard for evaluating SR algorithms. However, pristine HR images are typically unavailable or do not even exist in real-world SR scenarios. Blind image quality assessment, which directly estimates the quality of a test image, has been utilized as a performance measure for SR algorithms [21]–[24]. However, research on developing reliable blind SR-IQA models remains limited, particularly for the emerging real-world SR task [25]. Previous studies on SR-IQA have primarily relied on datasets containing super-resolved images generated by legacy SR algorithms [26], [27]. As demonstrated in Sec.III, task-specific SR-IQA meth-

[1]Junlin Chen, Guangtao Zhai, Xiaokang Yang, Weixia Zhang are with School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China {chenjunlin233, zhaiguangtao, xkyang, zwx8981}@sjtu.edu.cn

[2]Peibei Cao is with Nanjing University of Information Science and Technology, Nanjing 210044, China peibeicao2-c@my.cityu.edu.hk

*Corresponding author

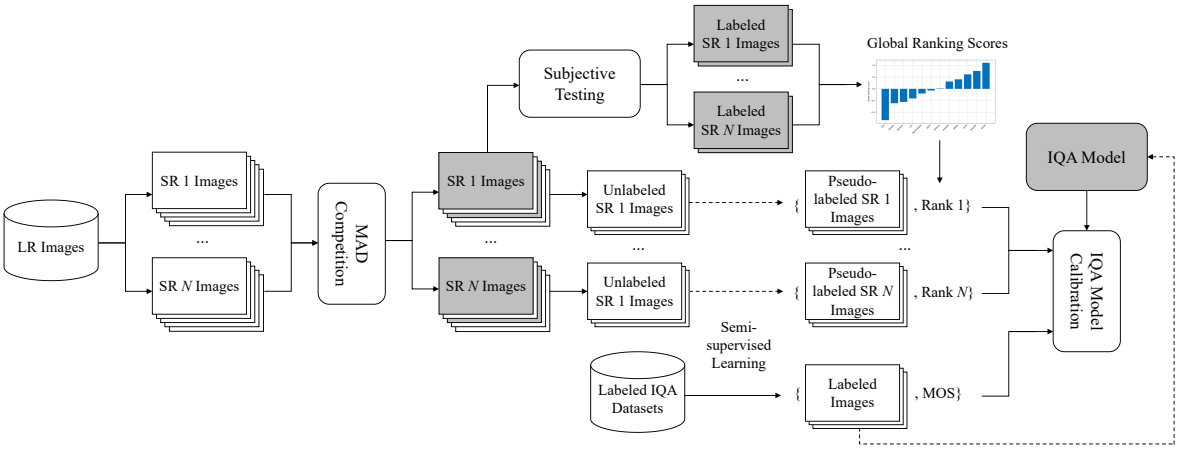


Fig. 1: The schematic of our subjective testing in the SR model ranking, and our training scheme.

ods [27]–[29] trained on such datasets struggle to generalize to the novel distortion patterns introduced by recent SR techniques (*e.g.*, diffusion-based methods). This limitation is largely attributed to the phenomenon of significant subpopulation shifts [30], [31].

While collecting extensive human quality annotations on super-resolved images would be highly beneficial for evaluating SR methods and training blind SR-IQA models, the extremely high annotation costs render this approach impractical [32]. In this work, we take steps to address the above issues. Our primary motivation stems from a common phenomenon: humans can make model-level comparisons based solely on their prior impressions. For example, people can easily conclude that GPT-4 is more powerful than GPT-3.5 without needing to compare every conversation generated by these two models. This insight suggests that if we know the overall relative performance ranking of two SR models, we can readily determine the average quality ranking of the super-resolved image batches produced by these models, without the need for time-consuming subjective evaluations on each individual image. It remains to obtain the model-level ranking information of SR algorithms. This approach automatically samples diverse and adaptive sets of images that best differentiate each pair of SR methods, all while adhering to a limited human annotation budget. The selected image pairs then undergo formal subjective testing, resulting in pairwise preference data that is subsequently used to derive a global ranking of the competing SR methods.

In theory, any BIQA model (pre-trained on existing IQA datasets) can be adapted to our task by regularizing its quality predictions for unannotated super-resolved images to align with the global rankings of SR models. In practice, we achieve this by computing the average quality predictions for super-resolved images generated by different SR methods and minimizing the statistical discrepancy between these predictions and the global model rankings using fidelity loss [33]. In summary, our contributions are three-fold:

- We construct a benchmark by collecting human preferences on super-resolved image pairs generated by different SR algorithms. These pairs are identified by the MAD competition as the most informative to derive the global model ranking of competing SR algorithms.
- We compare a range of BIQA models on our benchmark, revealing their relative strengths and weaknesses as metrics for real-world SR.
- We adapt pre-trained BIQA models for real-world SR by aligning their average quality predictions with global model rankings using a fidelity loss, thereby better handling novel distortion patterns from emerging SR algorithms.

II. METHODOLOGY

In this section, we begin by describing the creation of a large set of super-resolved images using various SR methods. We then outline the procedure of MAD competition [34] for globally ranking these SR methods. Next, we present the calibration process of a computational model for SR-IQA in detail. The overall framework is illustrated in Fig. 1.

A. Subjective Testing

Since all super-resolved images are ultimately viewed by humans, subjective testing remains the most reliable approach to obtain the overall performance rankings of different SR methods. Due to the time-consuming and labor-intensive nature of subjective testing, we adopt an efficient sample selection strategy that automatically chooses adaptive and diverse images for evaluation. Taking inspiration from the MAD competition methodology [32], [35], we start from an input LR image domain $\mathscr{X}$ and choose a set of SR methods $\mathscr{F} = \{f_n\}_{n=1}^{N}$, where each method produces a super-resolved image $f_n(x)$ from an input LR image $x \in \mathscr{X}$. Specifically, for each of two different SR methods f_i and f_j, we iteratively select the k-th image $\hat{x}^{(k)}$ that best differentiates between them by solving the following problem:

$$\hat{x}^{(k)} = \underset{x \in \mathscr{X} \setminus \mathscr{S}}{\arg\max}\, D_1(f_i(x), f_j(x)) + \lambda_1 D_2(x, \mathscr{S}) \qquad (1)$$

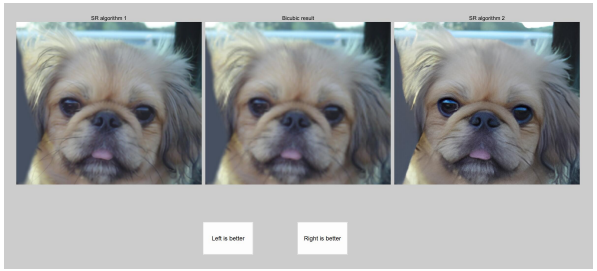


Fig. 2: Screenshot of the graphic user interface used for the subjective testing.

where $\mathscr{S} = \{\hat{x}^{(i)}\}_{i=1}^{k-1}$ is the set of $k-1$ image that have been selected according to Eq. (1). D_1 is a quantitative measure to approximate the perceptual distance between $f_i(x)$ and $f_j(x)$, while D_2 quantifies the semantic distance between x and the set $\mathscr{S}$, with λ_1 governing the trade-off between the two terms.

The identified images are screened to human subjects in formal subjective testing, where the two-alternative forced choice (2AFC) method is employed. In each trial, a subject is presented with a pair of images produced by two different SR methods $\{f_i(x), f_j(x)\}$, and is asked to choose the one with higher quality. A screenshot of the graphic user interface used in our subjective testing is shown in Fig. 2. We select the top-K ($K \ll |\mathscr{X}|$) images for each pair of SR methods $\{f_i, f_j\}$. Given N SR methods, we obtain $\frac{(N^2-N)\times K}{2}$ SR image pairs in total. We organize the human preference data collected from the subjective testing into an $N \times N$ matrix C, where C_{ij} represents the number of votes for f_i over f_j. Finally, we adopt maximum likelihood for multiple options [36] under the Thurstone's model [37] to infer the global ranking of $\mathscr{F}$, maximizing the log-likelihood of the count matrix C:

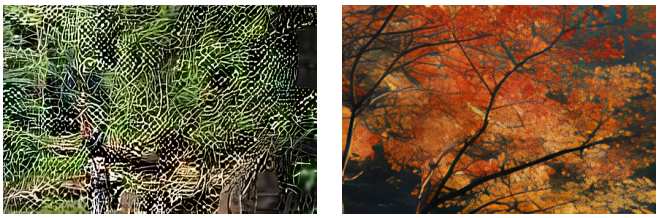
$$L(\mu; C) = \sum_{ij} C_{ij} \log(\Phi(\mu_i - \mu_j)) \tag{2}$$

where $\mu = [\mu_1, \mu_2, ..., \mu_N]$ is the vector of global ranking scores, Φ is the standard Normal cumulative distribution function, and $\sum_i \mu_i = 0$.

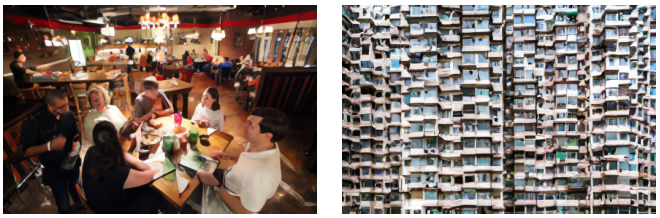
B. IQA Model Calibration

We have selected the most discriminative subset of image pairs from a large pool of super-resolved images through the MAD competition for subjective testing. Next, we aim to leverage the remaining set of unlabeled images, featuring large-scale and rich semantics, to adapt a pre-trained BIQA model tailored for SR.

Given an input LR image x, we have N super-resolved images $\{f_n(x)\}_{n=1}^N$ produced by N different SR methods. We assume a pretrained BIQA model $q_w(\cdot)$ parameterized by w, that has learned prior knowledge about image quality from existing IQA datasets. Observing that distortion patterns from specific SR methods are highly algorithm-dependent (see Fig. 3), directly fine-tuning $q_w(\cdot)$ on these data may cause overfitting to certain algorithms, compromising its learned priors. To address this, we freeze the weights of $q_w(\cdot)$ and



(a) USRGAN (b) BSRGAN



(c) ResShift (d) SinSR

Fig. 3: Distortion patterns from specific SR methods.

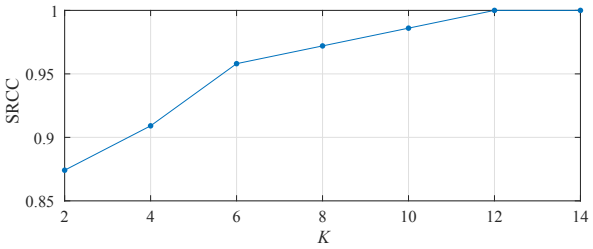


Fig. 4: The SRCC values between the top-16 and other top-K rankings.

introduce a lightweight rectifier $h_\theta(\cdot)$, which takes image features from the visual encoder as inputs and generate a tuple of scaling and shift parameters (α, β) to calibrate the quality predictions of q_w for super-resolved images generated by N SR models as $\alpha q_w(x) + \beta$, aligning them with the global model rankings $[\mu_1, \mu_2, ..., \mu_N]$.

Specifically, given two SR methods $f_i(\cdot)$ and $f_j(\cdot)$, we derive a binary label of their relative model ranking:

$$r_{ij} = \begin{cases} 1 & \text{if } \mu_i > \mu_j \\ 0.5 & \text{if } \mu_i = \mu_j \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

Inspired by [38], [39], we adopt a learning-to-rank method under the Thurstone model [37], which estimates the probability that the average quality of a group of super-resolved images $\{f_i(x_b)\}_{b=1}^B$ generated is higher than another group $\{f_j(x_b)\}_{b=1}^B$ as:

$$\hat{r}_{ij}(\{x_b\}_{b=1}^B) = \Phi\left(\frac{\frac{1}{B}\left(\sum_{b=1}^B Q_i(x_b) - \sum_{b=1}^B Q_j(x_b)\right)}{\sqrt{2}}\right), \tag{4}$$

where $Q_i = h_\theta(q_w(f_i(\cdot)))$, and the variance is fixed to one, Φ is the standard Normal cumulative distribution function. We

TABLE I: Comparison of 2AFC scores on RealSR-1K

General-purpose BIQA								
NIQE	DBCNN	HyperIQA	MUSIQ	MANIQA	CLIPIQA	LIQE	Q-ALIGN	ARNIQA
0.5818	0.6237	0.6386	0.6166	0.6443	0.6294	0.6391	0.6732	0.6451

General-purpose BIQA			Task-specific SR-IQA			Ours	Human (Oracle)
TOPIQ	Compare2Score	QualiCLIP+	NRQM	DISQ	TADSRNet		
0.6402	0.6518	0.6181	0.5995	0.5816	0.5553	**0.7206**	0.7767

TABLE II: 2AFC scores of ablation studies

Model I	Model II	Model III	Model IV	Ours
0.6644	0.7101	0.7091	0.7148	**0.7206**

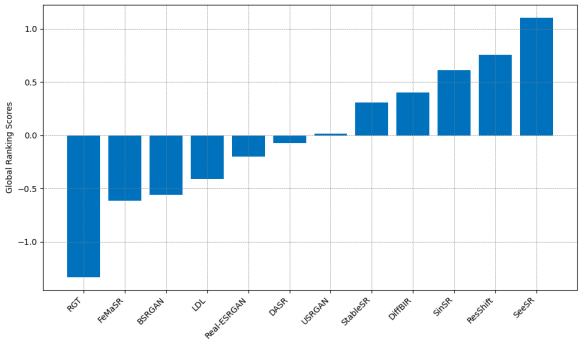
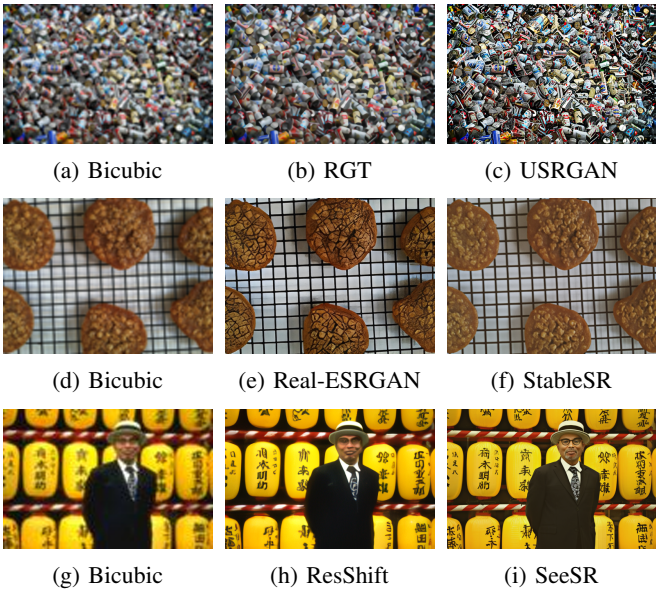


Fig. 5: Global ranking scores of SR methods.



(a) Bicubic (b) RGT (c) USRGAN

(d) Bicubic (e) Real-ESRGAN (f) StableSR

(g) Bicubic (h) ResShift (i) SeeSR

Fig. 6: Visual comparison of super-resolved images produced by different SR methods and the bicubic baseline.

adopt the fidelity loss [40] as the statistical distance measure:

$$\ell_p(f_i(x), f_j(x); w) = 1 - \sqrt{r_{ij}\hat{r}_{ij}(\{x_b\}_{b=1}^B)} - \sqrt{(1-r_{ij})(1-\hat{r}_{ij}(\{x_b\}_{b=1}^B))}. \quad (5)$$

Under the learning-to-rank paradigm, it is straightforward to integrate human-rated IQA datasets for joint training $h_\theta(\cdot)$ across multiple datasets. Following a similar pipeline of Eqs. (3)-(5), we compute an additional fidelity loss term ℓ_t using image pairs sampled from human-rated IQA datasets[1]. This results in a semi-supervised learning loss function:

$$\ell = \ell_t + \lambda_2 \ell_p, \quad (6)$$

where λ_2 controls the trade-off between two terms.

III. EXPERIMENTS

In this section, we first describe the experimental setups. We then present and analyze the results of subjective testing. Finally, we compare BIQA models on our test set along with public IQA datasets.

[1]The only difference is that the binary label is inferred from the ground-truth mean opinion scores (MOSs) instead of the global rankings as in Eq. (3)

A. Experimental Setups

We use 10,000 real-world LR images sourced from [41], [42], and the Internet as the input domain $\mathscr{X}$. These two datasets predominantly contain natural scenes, reflecting authentic degradation conditions encountered in real-world super-resolution tasks.

We select 12 SR methods ranging from 2020 to 2024 to ensure broad coverage of diverse foundational models, including GAN-based models: USRGAN [8], Real-ESRGAN [12], BSRGAN [11], LDL [13], and diffusion-based approaches: DiffBIR [16], ResShift [18], SinSR [19], StableSR [15], and SeeSR [17], as well as other advanced methods: DASR [43], FeMaSR [44], RGT [45]. All methods produce super-resolved images with a scaling factor of 4.

In Eq. (1), DISTs [46] and VGG-16 [47] are employed as D_1 and D_2, and we set the trade-off parameter λ_1 to 0.1. We set $K = 16$, resulting in $\binom{12}{2} \times 16 = 1056$ paired comparisons, which are viewed by 23 subjects (13 males and 10 females) with general knowledge of image processing. We show the Spearman rank order correlation coefficient (SRCC) values between the top-16 ranking and other top-K rankings in Fig. 4, indicating stable ranking results (SRCC > 0.97 when $K >= 8$).

In addition to the global rankings of SR methods, we also use the resulting dataset of approximate 1,000 image pairs to form a benchmark for BIQA models: $\{(f_i(\hat{x}^{(k)}), f_j(\hat{x}^{(k)})), p_{ijk}\}_{k=1}^{K}$, where $i, j = \{1, ..., N, i \neq j\}$, and p_{ijk} denotes the fraction of votes for i-th SR method over the j-th SR method, based on the relative quality of the super-resolved images for the k-th LR input. Following [48], we use the 2AFC score to evaluate BIQA models, which credits a BIQA model with a score of $pq + (1-p)(1-q)$, where p is the percentage of human votes and $q \in \{0, 1\}$ is the preference of the BIQA model. We term the resulting set of image pairs with human annotations as RealSR-1K.

B. Implementation Details

Inspired by the strong generalization ability of the CLIP [49] models, which serve as the vision backbone across a wide range of tasks, we first train a CLIP-based BIQA model q_w following a similar approach to [39]. We adopt ViT-B/32 [50] as the visual encoder along and work with a textual template $a photo with c quality$, where $c \in \mathscr{C} = \{1, 2, 3, 4, 5\} = \{$"bad", "poor", "fair", "good", "perfect"$\}$, corresponding to a Likert-scale of five quality levels. By applying a softmax function to the cosine similarities between the visual embedding and all textual embeddings, we obtain the probability distribution $\hat{p}(c|x)$ over the five quality levels. We then relate the discrete c to continuous quality scores by $q_w(x) = \sum_{c=1}^{C} \hat{p}(c|x) \cdot c$, where $C = 5$ is the number of quality levels. We train q_w on a combination of KonIQ-10k [51] and PIPAL [52], encouraging the model to learn prior knowledge of image quality from both photos captured in the wild and those processed by perceptual image processing algorithms. We train q_w using the AdamW optimizer [53] with an initial learning rate of 5×10^{-6}, and a weight decay of 0.001. To adjust the learning rate throughout training, we employ a cosine annealing scheduler with a period of 5 epochs.

After filtering out all images that overlap in content with RealSR-1K, we are left with 8,377 unannotated super-resolved images, which we denote as $\mathscr{D}$. This naturally results in $8,377 \times \binom{12}{2} = 552,882$ image pairs available for IQA model calibration as illustrated in Sec. II-B. Once the training of q_w is complete, we freeze its weights and attach a multi-layer perceptron (MLP) h_θ on top, which is then jointly trained using $\mathscr{D}$ along with the same datasets used for training q_w. During the IQA model calibration process, we set the batch size of $\mathscr{D}$ to 64, *i.e.*, $B = 64$ in Eq. (4) and Eq. (5). The rectifier h_θ is optimized for 8 epochs with an initial learning rate of 1×10^{-3}. We set the trade-off parameter λ_2 to 1. The training process took approximately 12 hours on a single NVIDIA RTX 4090 GPU.

C. Subjective Testing Results

We present the global ranking scores of SR models in Fig. 5, where we have several useful observations. First, the top-5 methods are all based on diffusion models, exactly matching the recent trends in this field. Second, the leading method, SeeSR [17], largely benefits from its degradation-aware prompt extractor module, enabling it to reconstruct semantically meaningful results from inputs with severe degradation (see Fig. 6 (i)). Third, among GAN-based methods, USRGAN [8] attains the highest ranking, despite being the earliest one. This underscores the effectiveness of integrating the flexibility of model-based methods into learning-based methods. Fourth, despite the remarkable performance on synthetic datasets, RGT [45] fails to generalize to the real-world SR (see Fig. 6 (b)), suggesting the importance of incorporating generative priors for real-world super-resolution.

D. Evaluation of BIQA Models

Comparison of 2AFC scores on RealSR-1K In Table I, we compare our method with 12 general-purpose BIQA models: NIQE [21], DBCNN [56], HyperIQA [57], MUSIQ [22], LIQE [39], MANIQA [24], Q-ALIGN [58], CLIPIQA [23], TOPIQ [59], QualiCLIP [60], ARNIQA [61], Compare2Score [62] and three tasks-specific SR-IQA methods: NRQM [27], DISQ [28], TADSRNet [29]. From these results, we derive several key insights. First, although NRQM, DISQ, and TADSRNet are specifically designed for the SR-IQA task, their performance is inferior to that of general-purpose BIQA models, with the exception of NIQE. We attribute this to the novel distortion patterns introduced by the latest SR models, which hinder task-specific SR-IQA methods from transferring the knowledge acquired from legacy SR distortions to unknown degradations. Second, Q-Align and Compare2Score achieve higher 2AFC scores than the other methods, highlighting the significant advantages of multi-modal large language model (MLLM)-based approaches in terms of model capacity and training data scale. Third, our method attains the highest performance despite having a considerably smaller model capacity and less training data compared to MLLM-based methods, demonstrating the effectiveness of our proposed semi-supervised learning approach in leveraging unlabeled data to enhance quality prediction for real-world SR.

Ablation Study We conduct ablation studies for different model variants as follows:

Model I: Directly applying q_w to predict the quality of images in RealSR-1K without further calibration.

Model II: Training the rectifier h_θ solely on $\mathscr{D}$, without leveraging other datasets.

Model III: Using our full model configuration but setting the batch size of $\mathscr{D}$ to 1.

Model IV: Training a BIQA model q_w directly on the combination of KonIQ-10k, PIPAL, and $\mathscr{D}$, bypassing the IQA model calibration process.

The results of these ablation experiments are presented in Table II, leading to several key observations. First, Model I exhibits limited generalization to RealSR-1K. This highlights the need for an additional adaptation step to address the subpopulation shift caused by the novel distortion patterns introduced by emerging SR models. Second, with exposure to $\mathscr{D}$, Model II achieves a significantly higher 2AFC score than Model I, demonstrating the effectiveness of using global model rankings as pseudo labels to adapt a pre-trained BIQA model to novel distortion scenarios. However, its

TABLE III: PLCC and SRCC results of BIQA models. The highest and second-highest results are indicated using **bold** and underline, respectively. The symbol "–" indicates that the full dataset is used for evaluation but the corresponding BIQA model is (partially) exposed to the same dataset

BIQA Model	QADS [26]		Ma17 [27]		SRIQA-Bench [54]		KonIQ-10k [51]		KADID-10k [55]		PIPAL [52]	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
NIQE	0.327	0.394	0.643	0.639	0.602	0.558	0.309	0.375	0.390	0.379	0.124	0.108
DBCNN	0.602	0.630	0.607	0.579	0.749	0.697	0.884	0.875	0.497	0.484	0.428	0.443
HyperIQA	0.654	0.660	0.712	0.687	0.733	0.677	0.917	0.906	0.506	0.468	0.475	0.489
MUSIQ	0.768	0.772	0.781	0.762	0.763	0.728	0.924	0.929	0.575	0.556	0.426	0.457
MANIQA	0.869	0.872	**0.879**	**0.870**	**0.817**	**0.799**	0.686	0.682	0.560	0.512	0.676	0.695
CLIPIQA	0.578	0.577	0.800	0.789	0.708	0.660	0.723	0.701	0.520	0.501	0.291	0.325
LIQE	0.780	0.790	0.751	0.714	0.757	0.721	0.912	0.928	0.667	0.662	0.550	0.564
Q-ALIGN	0.802	0.821	0.856	0.844	0.765	0.696	**0.941**	**0.940**	–	–	0.543	0.548
ARNIQA	0.802	0.807	0.753	0.720	0.712	0.663	0.890	0.870	0.717	0.725	0.382	0.422
TOPIQ	0.749	0.756	0.779	0.776	0.751	0.694	0.896	0.900	0.546	0.511	0.484	0.505
Compare2Score	0.830	0.865	0.863	0.847	0.726	0.689	0.939	0.931	–	–	0.533	0.545
QualiCLIP+	0.716	0.723	0.788	0.770	0.745	0.671	0.901	0.898	0.618	0.610	0.449	0.473
NRQM	0.722	0.727	–	–	0.679	0.566	0.479	0.438	0.299	0.168	0.176	0.200
DISQ	0.423	0.407	0.448	0.488	0.398	0.385	0.468	0.492	0.278	0.171	0.105	0.034
TADSRNet	–	–	0.724	0.639	0.569	0.474	0.146	0.155	0.302	0.231	0.124	0.138
Ours	**0.871**	**0.875**	0.870	0.856	0.795	0.771	0.918	0.925	**0.745**	**0.746**	**0.680**	**0.699**

performance remains inferior to our full model, further validating the efficacy of the proposed semi-supervised learning process across multiple datasets. Third, the relatively low performance of Model III highlights the challenge of directly training a model on multiple datasets with significantly different distributions. In contrast, the proposed two-stage training scheme effectively addresses this challenge. Fourth, due to the inherent noise in pseudo labels, the global model rankings do not strictly apply to each individual list of super-resolved images. Consequently, Model IV, which is trained with a batch size of one, exhibits suboptimal performance. In contrast, our full model leverages a significantly larger batch size, allowing the averaged predictions to better align with the global model rankings.

Further Testing We aim to verify the generalizability of the BIQA models. Specifically, we evaluate BIQA models on the full sets of: three SR-IQA datasets (QADS [26], Ma17 [27], and SRIQA-bench [54]) and a dataset consisting of various distortion types (KADID-10k [55]). We also evaluate BIQA models on the test set of KonIQ-10k [51] and the validation set of PIPAL. We present the SRCC and Pearson linear correlation coefficient (PLCC) results in Table III, drawing the following interesting observations: First, MANIQA generalizes well to datasets containing super-resolved images, but it exhibits relatively weak performance on KonIQ-10k and KADID-10k, highlighting a significant domain shift between algorithm-dependent distortions and photorealistic or synthetic distortions. Second, general-purpose BIQA models, particularly MLLM-based methods, demonstrate significantly stronger generalization ability compared to task-specific SR-IQA methods. This aligns with the prevalent use of general-purpose BIQA models—rather than task-specific SR-IQA models—as performance metrics in recent real-world SR studies [15]–[18]. Third, the proposed method presents strong generalization across various distortion scenarios, validating the effectiveness of our model design.

TABLE IV: 2AFC scores of more BIQA models and our versions

DBCNN	DBCNN (ours)	HyperIQA	HyperIQA (ours)
0.6391	0.6730	0.6386	0.6742

Application to Existing BIQA Models We apply the two-stage IQA model calibration method to two additional BIQA models: DBCNN [56] and HyperIQA [57]. We present the 2AFC scores on RealSR-1K in Table IV, from which we observe that our method introduces significant performance improvement over the baseline models. This underscores the strong generalizability of our framework and its orthogonality to innovations in model architecture.

IV. CONCLUSIONS

We propose an SR-IQA framework that uses the MAD competition to efficiently select representative samples for subjective evaluation, enabling global SR model ranking with minimal data. Leveraging this, we introduce a semi-supervised calibration method to adapt a pre-trained BIQA model to the distortion patterns of modern SR algorithms. Experiments on the RealSR-1K benchmark show superior performance and strong generalization across diverse distortions. The training scheme also allows existing BIQA models to be adapted to new distortion scenarios. In future work, we will extend this framework to other image processing applications like low-light enhancement, deblurring, denoising, and dehazing.

ACKNOWLEDGMENTS

This work was supported in part by Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), the Fundamental Research Funds for the Central Universities, and the National Natural Science Foundation of China under Grant 62371283.

REFERENCES

[1] Daniel Glasner, Shai Bagon, and Michal Irani, "Super-resolution from a single image," in *IEEE/CVF International Conference on Computer Vision*, 2009, pp. 349–356.

[2] Jianchao Yang, John Wright, Thomas S. Huang, and Yi Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[3] Hong Chang, Dit-Yan Yeung, and Yimin Xiong, "Super-resolution through neighbor embedding," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2004, pp. I–I.

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*, 2014, pp. 184–199.

[5] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 136–144.

[6] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita, "Deep back-projection networks for super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1664–1673.

[7] Muhammad Waleed Gondal, Bernhard Scholkopf, and Michael Hirsch, "The unreasonable effectiveness of texture transfer for single image super-resolution," in *European Conference on Computer Vision Workshops*, 2019, pp. 80–97.

[8] Kai Zhang, Luc Van Gool, and Radu Timofte, "Deep unfolding network for image super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3217–3226.

[9] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *European Conference on Computer Vision Workshops*, 2018, pp. 63–79.

[10] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao, "RankSRGAN: Generative adversarial networks with ranker for image super-resolution," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3096–3105.

[11] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4791–4800.

[12] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *IEEE/CVF International Conference on Computer Vision Workshops*, 2021, pp. 1905–1914.

[13] Jie Liang, Hui Zeng, and Lei Zhang, "Details or artifacts: A locally discriminative learning approach to realistic image super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5657–5666.

[14] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10850–10869, 2023.

[15] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy, "Exploiting diffusion prior for real-world image super-resolution," *International Journal of Computer Vision*, vol. 132, pp. 5929–5949, 2024.

[16] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong, "DiffBIR: Toward blind image restoration with generative diffusion prior," in *European Conference on Computer Vision*, 2025, pp. 430–448.

[17] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang, "SeeSR: Towards semantics-aware real-world image super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25456–25467.

[18] Zongsheng Yue, Jianyi Wang, and Chen Change Loy, "ResShift: Efficient diffusion model for image super-resolution by residual shifting," in *Advances in Neural Information Processing Systems*, 2023, pp. 13294–13307.

[19] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C. Kot, and Bihan Wen, "SinSR: Diffusion-based image super-resolution in a single step," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2024, pp. 25796–25805.

[20] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[21] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[22] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang, "MUSIQ: Multi-scale image quality transformer," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5148–5157.

[23] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy, "Exploring CLIP for assessing the look and feel of images," in *AAAI Conference on Artificial Intelligence*, 2023, pp. 2555–2563.

[24] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang, "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022, pp. 1191–1200.

[25] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3086–3095.

[26] Fei Zhou, Rongguo Yao, Bozhi Liu, and Guoping Qiu, "Visual quality assessment for super-resolved images: Database and method," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3528–3541, 2019.

[27] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang, "Learning a no-reference quality metric for single-image super-rolution," *Computer Vision and Image Understanding*, vol. 158, pp. 1–16, 2017.

[28] Tiesong Zhao, Yuting Lin, Yiwen Xu, Weiling Chen, and Zhou Wang, "Learning-based quality assessment for image super-resolution," *IEEE Transactions on Multimedia*, vol. 24, pp. 3570–3581, 2022.

[29] Xing Quan, Kaibing Zhang, Hui Li, Dandan Fan, Yanting Hu, and Jinguang Chen, "Tadsrnet: A triple-attention dual-scale residual network for super-resolution image quality assessment," *Applied Intelligence*, vol. 53, no. 22, pp. 26708–26724, 2023.

[30] Weixia Zhang, Dingquan Li, Chao Ma, Guangtao Zhai, Xiaokang Yang, and Kede Ma, "Continual learning for blind image quality assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 2864–2878, 2022.

[31] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang, "Task-specific normalization for continual learning of blind image quality models," *IEEE Transactions on Image Processing*, vol. 33, pp. 1898–1910, 2024.

[32] Peibei Cao, Zhangyang Wang, and Kede Ma, "Debiased subjective assessment of real-world image enhancement," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 711–721.

[33] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma, "FRank: a ranking method with fidelity loss," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 383–390.

[34] Kede Ma, Zhengfang Duanmu, Zhou Wang, Qingbo Wu, Wentao Liu, Hongwei Yong, Hongliang Li, and Lei Zhang, "Group maximum differentiation competition: Model comparison with few samples," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 851–864, 2020.

[35] Zhou Wang and Eero P. Simoncelli, "Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities," *Journal of Vision*, vol. 8, no. 12, pp. 8–8, 2008.

[36] K. Tsukida and M. R. Gupta, "How to analyze paired comparison data," Tech. Rep., Washington University Seattle, Department of Electrical Engineering, 2011.

[37] L. L. Thurstone, "A law of comparative judgment," *Psychological Review*, vol. 34, no. 4, pp. 273, 1927.

[38] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, 2021.

[39] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma, "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14071–14081.

[40] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma, "FRank: A ranking method with fidelity loss," in *International*

ACM SIGIR Conference on Research and Development in Information Retrieval, 2007, pp. 383–390.

[41] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 1004–1016, 2017.

[42] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 606–615.

[43] Jie Liang, Hui Zeng, and Lei Zhang, "Efficient and degradation-adaptive network for real-world image super-resolution," in *European Conference on Computer Vision*, 2022.

[44] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo, "Real-world blind super-resolution via feature matching with implicit high-resolution priors," in *ACM International Conference on Multimedia*, 2022, pp. 1329–1338.

[45] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, and Xiaokang Yang, "Recursive generalization transformer for image super-resolution," in *International Conference on Learning Representations*, 2024.

[46] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2022.

[47] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[48] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[50] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

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

[52] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S. Ren, and Dong Chao, "PIPAL: A large-scale image quality assessment dataset for perceptual image restoration," in *European Conference on Computer Vision*, 2020, pp. 633–651.

[53] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," 2019.

[54] Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang, "Toward generalized image quality assessment: Relaxing the perfect reference quality assumption," *arXiv preprint arXiv:2503.11221*, 2025.

[55] Hanhe Lin, Vlad Hosu, and Dietmar Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *International Conference on Quality of Multimedia Experience*, 2019, pp. 1–3.

[56] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020.

[57] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3667–3676.

[58] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin, "Q-ALIGN: Teaching LMMs for visual scoring via discrete text-defined levels," in *International Conference on Machine Learning*, 2024, pp. 54015–54029.

[59] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin, "Topiq: A top-down approach from semantics to distortions for image quality assessment," *IEEE Transactions on Image Processing*, vol. 33, pp. 2404–2418, 2024.

[60] Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini, "Quality-aware image-text alignment for opinion-unaware image quality assessment," 2025.

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

[62] Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang, "Adaptive image quality assessment via teaching large multimodal model to compare," 2024.