TuningIQA: Fine-Grained Blind Image Quality Assessment for Livestreaming Camera Tuning

Xiangfei Sheng[1*], Zhichao Duan[1*], Xiaofeng Pan[1], Yipo Huang[2], Zhichao Yang[1], Pengfei Chen[1†], Leida Li[1,3†]

[1]School of Artificial Intelligence, Xidian University,
[2]School of Data Science and Artificial Intelligence, Chang'an University,
[3]State Key Lab. of Electromechanical Integrated Manufacturing of High-Performance Electronic Equipments, Xidian University
xiangfeisheng@gmail.com, zach@stu.xidian.edu.cn, panxf@stu.xidian.edu.cn, yphuang@chd.edu.cn, yangzhichao@stu.xidian.edu.cn, chenpengfei@stu.xidian.edu.cn, ldli@xidian.edu.cn

Abstract

Livestreaming has become increasingly prevalent in modern visual communication, where automatic camera quality tuning is essential for delivering superior user Quality of Experience (QoE). Such tuning requires accurate blind image quality assessment (BIQA) to guide parameter optimization decisions. Unfortunately, the existing BIQA models typically only predict an overall coarse-grained quality score, which cannot provide fine-grained perceptual guidance for precise camera parameter tuning. To bridge this gap, we first establish **FGLive-10K**, a comprehensive fine-grained BIQA database containing 10,185 high-resolution images captured under varying camera parameter configurations across diverse livestreaming scenarios. The dataset features 50,925 multi-attribute quality annotations and 19,234 fine-grained pairwise preference annotations. Based on FGLive-10K, we further develop **TuningIQA**, a fine-grained BIQA metric for livestreaming camera tuning, which integrates human-aware feature extraction and graph-based camera parameter fusion. Extensive experiments and comparisons demonstrate that TuningIQA significantly outperforms state-of-the-art BIQA methods in both score regression and fine-grained quality ranking, achieving superior performance when deployed for livestreaming camera tuning.

Introduction

With the rapid expansion of mobile livestreaming across entertainment and e-commerce platforms (*e.g.*, TikTok Live, YouTube Shorts), automatic camera parameter tuning has become increasingly important for delivering superior Quality of Experience (QoE) (Gilstrap and Gilstrap 2023; Hamilton, Garretson, and Kerne 2014). Unlike traditional photography where post-capture editing can compensate for suboptimal settings (Kai et al. 2025), livestreaming demands real-time optimization during capture, making automated camera tuning essential for both professional streamers and casual users lacking photography expertise.

Figure 1: Illustration of the fine-grained challenge in IQA. Coarse-grained (a) and fine-grained quality difference (b) caused by inappropriate camera parameters. SOTA score-based methods LIQE (Zhang et al. 2023) and MUSIQ (Ke et al. 2021) fail to capture fine-grained quality differences.

Image quality assessment (IQA) constitutes the core of effective camera tuning. In practice, automated camera tuning follows a common approach: algorithms perform parameter adjustments within consistent scenes, followed by an objective metric to predict the output image quality and provide feedback to the parameter tuning, producing the best quality image. Therefore, the fundamental challenge lies in precise blind image quality assessment (BIQA), particularly during the fine-grained parameter optimization process where quality variations are often subtle and may not dramatically change overall image appearance. In these scenarios, fine-grained quality assessment becomes essential for distinguishing between parameter configurations that produce visually similar but still *noticeable* perceptual differences.

Existing BIQA methods (Zhang et al. 2023; Ke et al. 2021; Wang, Chan, and Loy 2023; Sheng et al. 2025b) designed for score-based evaluation, face fundamental limitations when applied to camera tuning scenarios. As illustrated in Figure 1, while these methods demonstrate accuracy for coarse-grained quality discrimination, they struggle

*These authors contributed equally.
†Corresponding authors

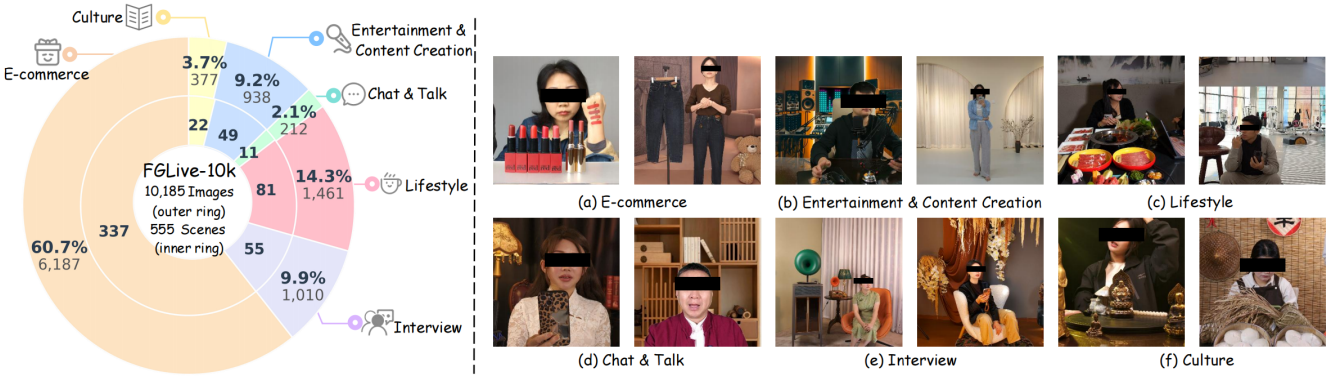

Figure 2: Statistics of the FGLive-10K dataset and representative samples covering diverse livestreaming scenarios.

with fine-grained assessment, which is essential for parameter optimization (Kai et al. 2024). When parameter adjustments produce subtle quality changes, existing BIQA methods often fail to provide reliable guidance, restricting their practical utility in automated camera tuning systems.

Livestreaming camera tuning also presents two specific requirements, which are inadequately addressed by general-purpose BIQA methods: (i) **Human-region Priority**, emphasizing image quality evaluation on human-centric regions such as faces and bodies, reflecting the heightened sensitivity of the human visual system (HVS) (Grimson 1981) towards these areas over background elements. (ii) **Parameter Interdependency Modeling**, capturing complex physical relationships among camera settings that jointly influence image quality. Camera parameters exhibit intricate interdependencies rooted in photographic principles—adjusting one parameter often necessitates compensating changes in others to maintain desired quality. Current BIQA methods fail to model these parameter interactions and lack the specialized architectural designs needed for camera-guided quality optimization.

To address the above limitations, we present multi-fold technical contributions spanning dataset construction, method development, and experimental analysis:

• **Dataset.** We construct FGLive-10K, the first-of-its-kind fine-grained BIQA dataset specifically designed for livestreaming camera tuning. As shown in Figure 2, the dataset comprises 10,185 high-resolution images from 555 distinct scenes, each containing systematic camera parameter variations. Beyond conventional Mean Opinion Scores (MOS), we provide fine-grained pairwise preference annotations, enabling reliable discrimination of subtle quality differences essential for parameter optimization.

• **Method.** We propose TuningIQA, a fine-grained BIQA framework with two key innovations. First, we design a Human-aware Feature Extraction (HFE) module that prioritizes human-centric regions through aesthetics-guided feature learning. Second, we introduce an optional Graph-based Camera Parameter Fusion (GCPF) module that models physical relationships among camera settings through graph attention networks. The framework unifies multi-attribute regression and fine-grained rank learning for comprehensive quality assessment and camera tuning guidance.

• **Experimental Analysis.** Extensive experiments demonstrate TuningIQA's superior performance over state-of-the-art methods in both coarse-grained quality score prediction and fine-grained quality ranking. Our analysis reveals fundamental limitations in existing score-based approaches for fine-grained assessment, while practical camera tuning experiments show 74-76% win rates against leading BIQA methods, confirming the effectiveness of our specialized design for camera tuning applications.

Dataset	#Img	#Param	#Attr	Annotation
BID	585	0	0	MOS
CID2013	480	0	4	MOS
LIVE Challenge	1,162	0	0	MOS
KonIQ-10K	10,073	0	4	MOS
SPAQ	11,125	0	5	MOS
FGLive-10K	10,185	7	4	MOS+Rank

Table 1: Comparison of popular BIQA datasets with authentic distortions.

Related Work

IQA Benchmarks

Benchmark datasets have driven significant progress in image quality assessment research. Early datasets focused on synthetic distortions like LIVE (Sheikh, Sabir, and Bovik 2006), using simulated degradations on reference images to enable full-reference IQA development. However, obtaining pristine references proves impractical in real scenarios, motivating research toward authentic distortions. Pioneering efforts include BID (Ciancio et al. 2011) with 585 DSLR-captured blur images, LIVE Challenge (Ghadiyaram and Bovik 2016) featuring 1,162 mobile-captured images with crowdsourced annotations, and KonIQ-10k (Hosu et al. 2020) expanding diversity with 10,000 multimedia images. SPAQ (Fang et al. 2020) incorporated EXIF metadata and laboratory-controlled annotations for smartphone imaging scenarios. A brief comparison of authentic BIQA datasets is summarized in Table 1.

IQA Method

Early IQA methods relied on statistical priors (Zhang, Zhang, and Bovik 2015; Mittal, Soundararajan, and Bovik 2013) and handcrafted natural scene statistics (Mittal, Moorthy, and Bovik 2012), achieving reasonable success on synthetic distortions but struggling with authentic degradations (Ghadiyaram and Bovik 2016). The shift toward real-world assessment drove CNN-based approaches (Bosse et al. 2016; Zhang et al. 2020) that learned hierarchical quality representations directly from data. Recent advances include architectural innovations like HyperIQA's semantic-quality disentanglement (Su et al. 2020) and Transformer-based models with attention mechanisms (Zhang et al. 2023; Ke et al. 2021; Sheng et al. 2023; Li et al. 2025). More recently, MLLM-based methods (Wu et al. 2023; Sheng et al. 2025a) have emerged, demonstrating superior generalization capabilities and interpretability.

Despite considerable advancements, existing methods exhibit fundamental limitations when applied to livestreaming camera tuning. (1) Most methods target general-purpose quality evaluation, overlooking key characteristics of livestreaming environments such as human-centric visual focus and systematic camera parameter variations. (2) Existing methods typically produce single coarse-grained quality scores, which are inherently limited in identifying subtle differences essential for fine-grained perception. These limitations underscore the need for a specialized fine-grained BIQA framework tailored for livestreaming camera tuning, motivating the development of TuningIQA.

FGLive-10K

Dataset Construction

We describe the image collection and annotation processes for FGLive-10K, designed specifically for livestreaming camera tuning scenarios. It should be noted that all depicted individuals provided consent for use in this research.

Image Collection. To capture camera parameter-induced distortions, we employed three internally developed livestreaming cameras in authentic usage scenarios. Field Application Engineers first optimized reference parameters for each scenario, then generated distorted images by introducing systematic deviations to aperture, shutter speed, ISO, white balance, contrast, saturation, and sharpness. Figure 3(a) demonstrates parameter-induced variations within a single scene. The collected images span typical livestreaming applications including e-commerce, entertainment, and creative content. After filtering, FGLive-10K comprises 10,185 high-resolution (1920×1080) images from 555 distinct scenes.

To explore Parameter Interdependency Modeling, we preserved the complete 7-parameter metadata for a subset of images, creating **FGLive-p** with 5,559 training images and 1,148 test images. This subset follows the same annotation protocol as the main dataset.

Image Annotation. We invited 25 volunteers with diverse backgrounds (IQA researchers, photography enthusiasts, art students) to participate in subjective experiments following ITU-R BT.500-15 recommendations (Series 2023).

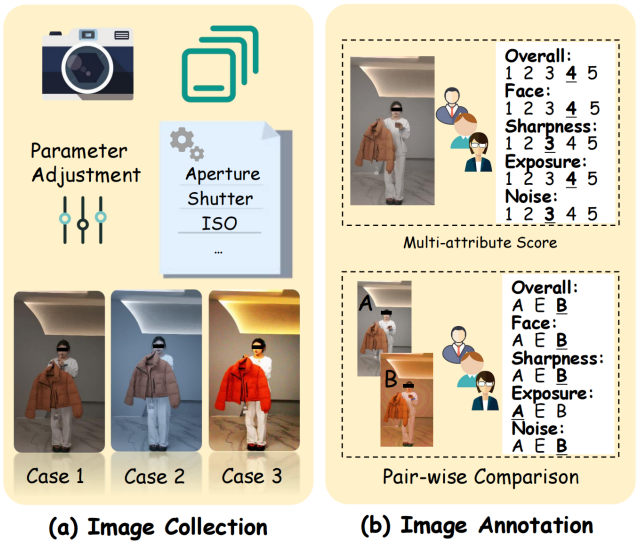

Figure 3: Details of the FGLive-10K dataset construction. (a) Various distorted images are collected by adjusting the camera parameters. (b) Multi-attribute quality score annotation and pairwise comparison of fine-grained image pair.

Multi-attribute MOS. Annotators assigned discrete scores (1-5: bad to excellent) for overall quality and four critical attributes: *face quality* (visual quality of human facial regions), *sharpness* (image clarity and detail preservation), *exposure* (brightness and contrast appropriateness), and *noise* (absence of visual artifacts and grain). After post-screening based on Pearson correlation, each image received at least 16 annotations, yielding Mean Opinion Scores:

$$s^{\text{attr}} = \frac{1}{N}\sum_{i=1}^{N} s_i^{\text{attr}}, \tag{1}$$

where s^{attr} denotes the MOS for a specific attribute, N is the number of annotators.

Pairwise Comparison. To address the limitation of MOS in fine-grained quality discrimination, we constructed image pairs with subtle differences within each scene and conducted pairwise comparisons. Initial preferences derived from MOS comparisons ($c^*_{pq} = \mathbb{I}(s_p > s_q)$) were refined for pairs with $\Delta s = |s_p - s_q| \leq 0.8$, as these exhibit significant annotation uncertainty (Katsigiannis et al. 2018). Fine-grained pairs underwent human verification where annotators selected preferred images or marked equivalence. The final preference was:

$$c_{pq} = \begin{cases} c^*_{pq}, & \text{if } \Delta s > 0.8, \\ \frac{1}{K}\sum_{k=1}^{K} \psi_k(I_p, I_q), & \text{if } \Delta s \leq 0.8, \end{cases} \tag{2}$$

where K denotes annotations per pair and $\psi_k \in \{0, 0.5, 1\}$ represents individual judgments (worse, equivalent, better quality of I_p vs I_q). This strategy produced 91,946 annotated pairs, with 21% (19,234 pairs) refined through pairwise comparison.

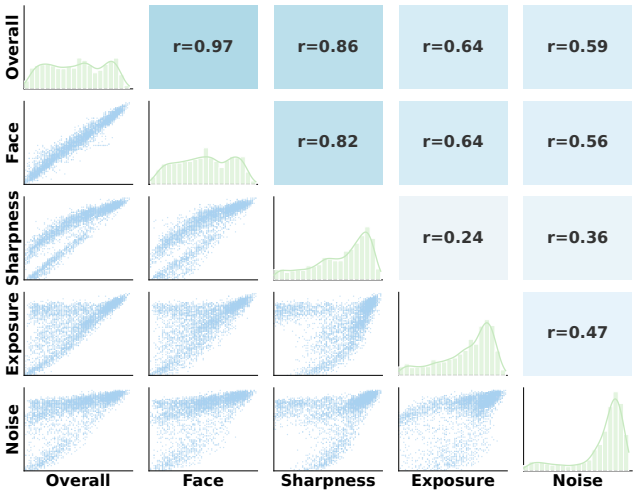

Figure 4: Relationships among quality dimensions: scatterplot (bottom left), Pearson correlation (top right), annotation distribution (diagonal).

Statistics and Analysis

FGLive-10K contains 555 scenes with 10,185 images annotated across five quality attributes. The dataset is split into 451 training scenes (8,148 images, 72,843 pairs) and 104 test scenes (2,037 images, 3,705 fine-grained pairs).

Annotation Distribution. The FGLive-10K dataset features comprehensive multi-attribute quality annotations, where distributions are visualized along the diagonal in Figure 4. The inter-attribute variations reveal their distinct perceptual value. The uniform distribution of overall quality scores confirms our dataset's balanced coverage of diverse quality levels across various scenarios.

Attribute Correlation. To study FGLive-10K from the correlation perspective, we visualize the Pearson Linear Correlation Coefficient (PLCC) and scatterplot among each dimension in Figure 4. It can be observed that the correlation between overall and face is extremely high, which indicates that face quality can largely affect overall perception. Low correlations among sharpness, exposure, and noise dimensions suggest their complementary yet independent contributions. This divergence highlights the necessity of multi-attribute evaluation for precise distortion perception.

Individual Annotation Consistency. We investigate annotator reliability through correlation analysis in Figure 5. Post-screened annotators strongly agree with MOS (average PLCC=0.85), confirming annotation reliability at the coarse-grain level. However, when analyzing within smaller MOS tiers (*e.g.* Figure 5 rows 2-6), the correlations drop significantly (*i.e.* PLCC=0.24 in the 4.2-5.0 tier). Which highlights single-stimulus limitations in discriminating subtle quality differences. Such findings validate the necessity of pairwise comparison refinement for fine-grained image pairs.

TuningIQA Metric

Based on insights from the FGLive-10K dataset, we further propose TuningIQA, a fine-grained quality assessment framework for livestreaming camera tuning. The framework supports two operational modes: *single-image mode* for direct quality scoring and *pairwise mode* for fine-grained comparison. The pipeline is illustrated in Figure 6. TuningIQA is built upon two core modules: *human-centric feature extraction* which prioritizes perceptually critical regions, and *graph-based camera parameter fusion* that models camera setting interdependencies.

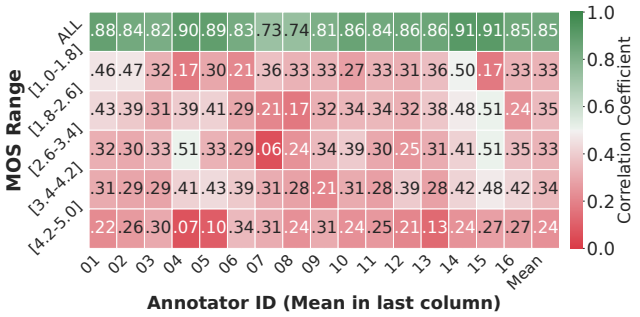

Figure 5: Correlation between Individual Annotation and MOS within specific quality tiers (MOS range).

Human-aware Feature Extraction

To address the human-region priority, we design a Human-aware Feature Extraction (HFE) module that emphasizes human-centric regions. Specifically, we first employ Faster R-CNN (Ren et al. 2016) to detect human subjects in input images, obtaining human bounding boxes BBox_h that guide subsequent feature extraction. To understand human-centric visual quality, we leverage a Human Aesthetics Network built upon the EfficientNetV2-M (Tan and Le 2021) backbone, pre-trained on the human subset of the TAD66K dataset (He et al. 2022). This backbone extracts a multi-scale feature map $\mathbf{M}$ with C channels that captures aesthetic-aware representations.

Given the detected human bounding boxes, we extract human region features $\mathbf{R}^h$ using ROI Align operations. The dimension of $\mathbf{R}^h$ is reduced to $\frac{C}{2}$ and then appended to each spatial location of $\mathbf{M}$, creating an enhanced feature representation that integrates both global image context and human-centric information. We then partition the enhanced feature map into nine spatial regions $\{\hat{\mathbf{M}}_k\}_{k=1}^9$ centered around the human subjects. To explicitly handle different partitions, we employ nine individual nonlinear transformations with residual learning:

$$\mathbf{F}_k = \Phi_k(\hat{\mathbf{M}}_k) + \hat{\mathbf{M}}_k, \quad k \in \{1, 2, \ldots, 9\}, \tag{3}$$

where $\Phi_k(\cdot)$ is a partition-specific C-channel convolutional layer.

The updated partitions $\{\mathbf{F}_k\}_{k=1}^9$ are combined to form the human-aware feature map. Then we employ a cross-attention mechanism that fuses human-aware features $\mathbf{F}_h$ (obtained through Global Average Pooling applied to the combined partitions) with basic backbone features $\mathbf{F}_b$:

$$\mathbf{F}_q = \text{CrossAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{4}$$

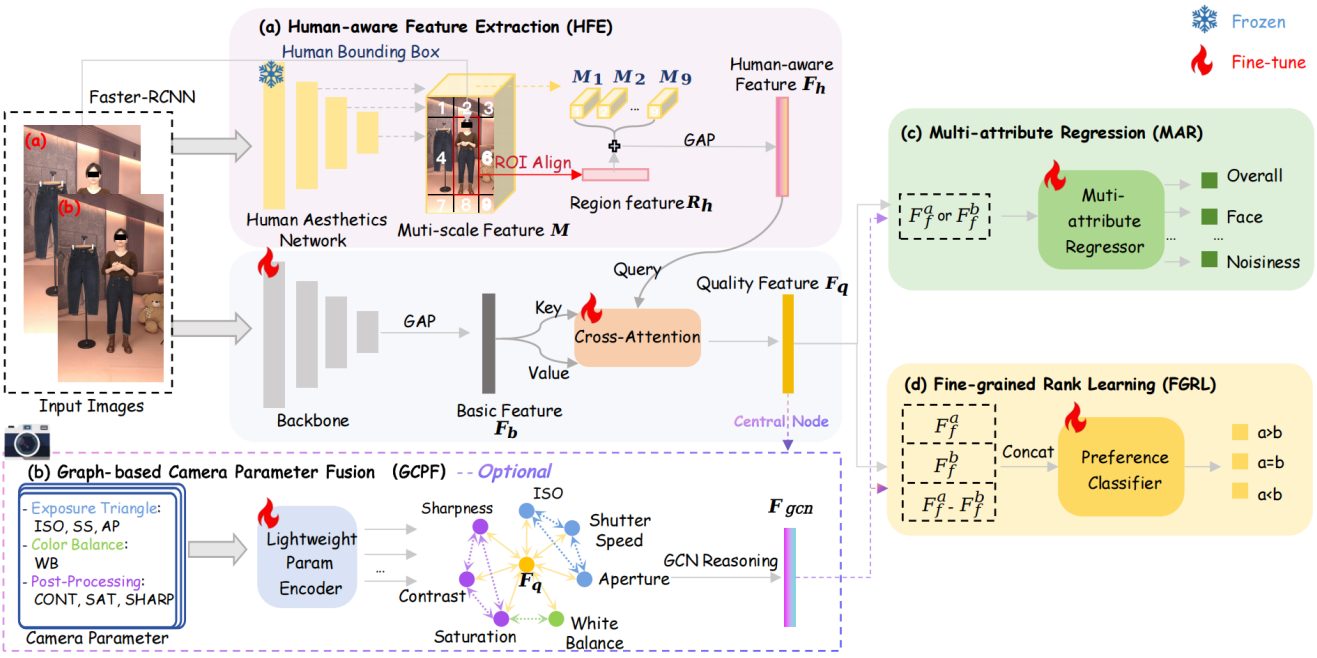

Figure 6: Overview of the proposed TuningIQA framework.

where $\mathbf{Q} = \mathbf{F}_h\mathbf{W}_Q, \mathbf{K} = \mathbf{F}_b\mathbf{W}_K, \mathbf{V} = \mathbf{F}_b\mathbf{W}_V$ are learned projections. The resulting quality-aware features $\mathbf{F}_q$ effectively integrate human-centric visual cues with global image context for comprehensive quality assessment.

Graph-based Camera Parameter Fusion

For images with available camera metadata, we introduce an *optional* Graph-based Camera Parameter Fusion (GCPF) module that explicitly models these parameter interactions through graph neural networks.

Graph Construction and Node Encoding. We construct a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where nodes $\mathcal{V}$ represent both visual features and camera parameters. The visual node serves as the central hub that aggregates information from all parameter nodes, enabling rich cross-modal interactions. Specifically, we define eight nodes: one visual node encoding $\mathbf{F}_q$ and seven parameter nodes. Each parameter node is encoded through individual linear transformations:

$$\mathbf{v}_{\text{visual}} = \mathbf{W}_v\mathbf{F}_q, \quad \mathbf{v}_i = \mathbf{W}_p p_i, \tag{5}$$

where $\mathbf{W}_v$ and $\mathbf{W}_p$ are learned projection matrices, and p_i represents the i-th camera parameter value.

Physical Relationship Modeling. The edge set $\mathcal{E}$ encodes both cross-modal connections and physical relationships based on photographic principles. We establish connections following relationships: (1) *Cross-modal connections*: the visual node connects to all parameter nodes, enabling parameter-visual feature interaction; (2) *Exposure Triangle*: ISO, shutter speed, and aperture form a strongly connected subgraph as they jointly determine exposure—increasing ISO compensates for faster shutter or smaller aperture; (3) *Post-processing Chain*: contrast, saturation, and sharpness exhibit mutual dependencies in image enhancement, where higher contrast often requires adjusted saturation; (4) *Color Correlation*: white balance directly influences saturation perception due to color temperature effects.

Graph Reasoning. The graph reasoning process employs two-layer Graph Attention Networks (GAT) (Velickovic et al. 2017) to propagate information across the parameter-visual graph:

$$\mathbf{H}^{(1)} = \text{GAT}_1(\mathbf{X}, \mathcal{E}), \quad \mathbf{H}^{(2)} = \text{GAT}_2(\mathbf{H}^{(1)}, \mathcal{E}), \tag{6}$$

where $\mathbf{X} = [\mathbf{v}_{\text{visual}}, \mathbf{v}_1, \ldots, \mathbf{v}_7]$ contains all encoded node features. The first GAT layer employs 4 attention heads with concatenation to capture diverse relationship patterns, while the second layer uses single-head attention for feature integration. The final parameter-aware representation $\mathbf{F}_{GCN}$ is extracted from the updated central visual node:

$$\mathbf{F}_{gcn} = \mathbf{H}^{(2)}[0, :]. \tag{7}$$

where $\mathbf{H}^{(2)}[0, :]$ denotes the first row of matrix $\mathbf{H}^{(2)}$. This approach enables the model to reason about parameter interactions and their combined effects on image quality, providing richer representations for subsequent quality prediction.

Multi-attribute Regression and Fine-grained Ranking

Multi-attribute Regressor. For comprehensive quality assessment, we implement five specialized prediction heads that estimate quality scores across different perceptual dimensions:

$$\hat{s}^{\text{attr}} = \text{MLP}_{\theta_s^{\text{attr}}}(\mathbf{F}_f), \tag{8}$$

where $\mathbf{F}_f$ represents the fused features (either $\mathbf{F}_q$ or $\mathbf{F}_{gcn}$ when parameters are available), and attr $\in$ {overall, face, sharpness, exposure, noise}.

Method	Backbone	#Param	Input Size	Score Regression		FG Ranking
				SRCC↑	PLCC↑	FG-ACC↑
NIQE (Mittal, Soundararajan, and Bovik 2013)	–	–	Full res.	0.4512	0.5065	0.4327
ILNIQE (Zhang, Zhang, and Bovik 2015)	–	–	Full res.	0.5706	0.5412	0.4089
BRISQUE (Mittal, Moorthy, and Bovik 2012)	–	–	Full res.	0.3411	0.4280	0.4583
DBCNN (Zhang et al. 2020)	VGG16	15.31M	224×224	0.7507	0.7495	0.5667
HyperIQA (Su et al. 2020)	ResNet-50	27.38M	224×224	0.8472	0.8518	0.5980
MT-A (Fang et al. 2020)	ResNet-50	23.57M	512×512	0.8821	0.8829	0.6223
CLIP-IQA+ (Wang, Chan, and Loy 2023)	CLIP (ResNet-50)	149.70M	Full res.	0.7056	0.7039	0.5570
MUSIQ (Ke et al. 2021)	Transformer	157.23M	Full res.	0.8662	0.8687	0.6861
SARQUE (Huang et al. 2023)	MobileNetV3-S	8.85M	224×224	0.8581	0.8591	0.6095
LIQE (Zhang et al. 2023)	CLIP (ViT-B/32)	151.28M	224×224	0.9235	0.9206	0.6533
Q-Align (Wu et al. 2023)	mPLUG-Owl2-7B	8197.86M	448×448	0.8721	0.8315	0.6311
Compare2Score (Zhu et al. 2024)	mPLUG-Owl2-7B	8197.86M	448×448	0.8502	0.8190	0.6266
TuningIQA	MobileNetV3-S	58.29M	224×224	0.9308	0.9302	0.7065
TuningIQA	ResNet-50	84.62M	224×224	**0.9385**	**0.9364**	**0.7284**

Table 2: Performance of TuningIQA and the state-of-the-art BIQA methods in two tasks: quality score regression and fine-grained ranking on the FGLive-10K database.

Fine-grained Preference Classifier. To capture subtle quality differences, we incorporate a pairwise comparison mechanism:

$$\hat{c}^{\text{attr}} = \sigma\left(\text{MLP}_{\theta_c^{\text{attr}}}([\mathbf{F}_a \oplus \mathbf{F}_b \oplus \mathbf{F}_a - \mathbf{F}_b])\right), \quad (9)$$

where $\oplus$ denotes concatenation, σ is the sigmoid function, and the difference term captures relative quality variations.

Learning Objectives

We optimize TuningIQA using a multi-task learning framework with carefully designed loss functions.

Confidence-Weighted Regression Loss. To handle annotation uncertainty in fine-grained assessment, we design variance-based confidence weighting:

$$\mathcal{L}_{reg} = \frac{1}{|\mathcal{D}|} \sum_{(I_i, s_i)\in\mathcal{D}} \exp(-v_i) \sum_{\text{attr}} |\hat{s}_i^{\text{attr}} - s_i^{\text{attr}}|, \quad (10)$$

where v_i denotes the annotation variance, adaptively reducing the impact of unreliable quality scores.

Fine-grained Ranking Loss. For rank learning, we employ the binary cross-entropy:

$$\mathcal{L}_{rank} = -\frac{1}{|\mathcal{P}|} \sum_{(I_p, I_q)\in\mathcal{P}} \sum_{\text{attr}} BCE(c_{pq}^{\text{attr}}, \hat{c}_{pq}^{\text{attr}}), \quad (11)$$

where $BCE(c, \hat{c})$ denotes the binary cross-entropy function. The combined objective: $\mathcal{L} = \lambda_{reg}\mathcal{L}_{reg} + \lambda_{rank}\mathcal{L}_{rank}$ with empirically determined weights $\lambda_{reg} : \lambda_{rank} = 1 : 2$.

Experiments

Performance Evaluation on FGLive-10K

Evaluation Protocol. We compare the proposed TuningIQA model against handcrafted and state-of-the-art learning-based BIQA models. All learning-based models are trained on FGLive-10K following strict scene-level division. We evaluate using SRCC, PLCC for score prediction, and fine-grained accuracy for pairwise ranking.

Method	Metric	Sharp.	Noise	Exp.	Face
CLIP-IQA	SRCC	0.532	0.472	0.199	0.510
	PLCC	0.528	0.434	0.208	0.512
	FG-ACC	0.649	0.622	0.571	0.518
MT-A	SRCC	0.896	0.804	0.782	0.807
	PLCC	0.943	0.898	0.863	0.802
	FG-ACC	0.688	0.665	0.621	0.582
SARQUE	SRCC	0.868	0.743	0.700	0.754
	PLCC	0.930	0.876	0.804	0.754
	FG-ACC	0.666	0.645	0.602	0.547
TuningIQA	SRCC	**0.935**	**0.805**	**0.897**	**0.885**
	PLCC	**0.963**	**0.905**	**0.949**	**0.880**
	FG-ACC	**0.773**	**0.746**	**0.776**	**0.725**

Table 3: Multi-attribute performance comparison.

Implementation Details. We implement TuningIQA using PyTorch and train on RTX4090 GPUs. The model is optimized end-to-end using AdamW with cosine annealing at a maximum learning rate of 1×10^{-5} and batch size of 64. Input images are resized to 256×256 and randomly cropped to 224×224 during training. The model is trained for 5 epochs on all image pairs from the training set. We implement TuningIQA based on lightweight backbones to ensure application efficiency.

Overall Performance. Table 2 summarizes the performance comparison between TuningIQA and state-of-the-art BIQA methods on the FGLive-10K dataset. It is easily observed that while existing methods achieve reasonable score regression performance, they exhibit significantly poor fine-grained ranking capabilities, highlighting the inadequacy of score-based methods for subtle quality discrimination, which is essential in camera tuning. In contrast, TuningIQA achieves superior performance in both tasks, demonstrating the effectiveness of joint regression and ranking learning.

Multi-attribute Evaluation. To evaluate multi-attribute performance, we further compare TuningIQA with three

Dataset	Config.	SRCC↑	PLCC↑	FG-ACC↑
FGLive-10K	Baseline	0.8757	0.9201	0.7258
	w/ HFE	**0.8917**	**0.9267**	**0.7496**
FGLive-p	Baseline	0.8206	0.8768	0.7092
	w/ GCPF	0.8552	0.8963	0.7308
	w/ GCPF+HFE	**0.8613**	**0.9009**	**0.7382**

Table 4: Ablation study. Results report average performance across overall score and attributes.

SOTA methods that support quality attribute prediction: CLIP-IQA, MT-A, and SARQUE. As shown in Table 3, face quality assessment proves particularly challenging for prior methods, where TuningIQA achieves superior performance through human-aware feature extraction across all attributes, proving the efficacy of our human-centric design.

Qualitative Analysis. Figure 7 shows gMAD competition results (Ma et al. 2020), exposing perceptual inconsistencies among models. LIQE and MUSIQ produce similar predictions for significantly different images or exaggerate minimal differences, while TuningIQA maintains correct discrimination, aligning better with human visual sensitivity.

Ablation Study. Table 4 presents ablation study results on two datasets: FGLive-10K and FGLive-p (subset with camera parameters). We systematically analyze the contribution of the Human-aware Feature Extraction (HFE) module and Graph-based Camera Parameter Fusion (GCPF) module. The results show that HFE consistently improves performance across all metrics on FGLive-10K, with notable gains in FG-ACC. On the FGLive-p subset, GCPF demonstrates substantial improvements, validating the importance of modeling parameter interdependencies. The combination of both modules achieves the best performance, confirming their complementary effectiveness.

Application for Camera Tuning

To validate the practical effectiveness of IQA metrics in guiding livestreaming camera parameter tuning, we simulate fine-grained parameter adjustment scenarios by incrementally adjusting camera parameters with minimal step sizes. We capture images after each adjustment and employ different metrics to rank the quality of these images, thereby identifying optimal parameters. TuningIQA performs ranking based on pairwise comparisons, while other methods rely on score-based ranking. As illustrated in Figure 8, we demonstrate the fine-grained quality differences resulting from small adjustments to Exposure Value and ISO, along with the ranking results from different methods. The results reveal the inaccuracy of score-based methods in fine-grained evaluation and their insufficient sensitivity to facial region quality. Additionally, we conduct a subjective experiment to compare the top-1 image quality between TuningIQA and LIQE/MUSIQ ranking results. Evaluation across 44 diverse scenarios with 12 volunteers shows that TuningIQA achieves **76%** and **74%** win rates against LIQE and MUSIQ respectively. These significant improvements validate that fine-grained quality assessment capabilities are essential for practical camera tuning applications, where subtle parameter adjustments can substantially impact user experience despite minimal changes in overall image appearance.

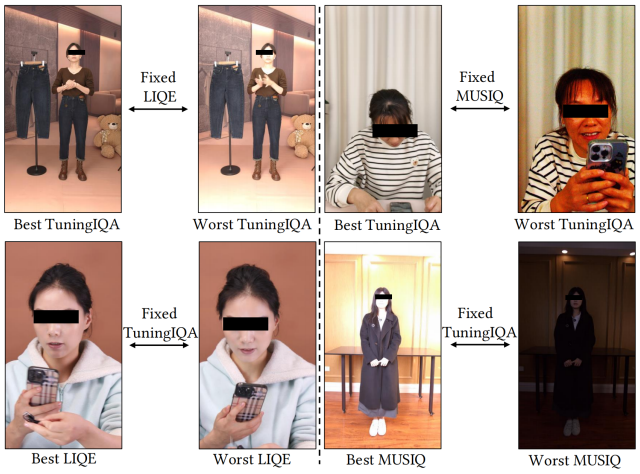

Figure 7: gMAD competition results on the FGLive-10K against LIQE and MUSIQ.

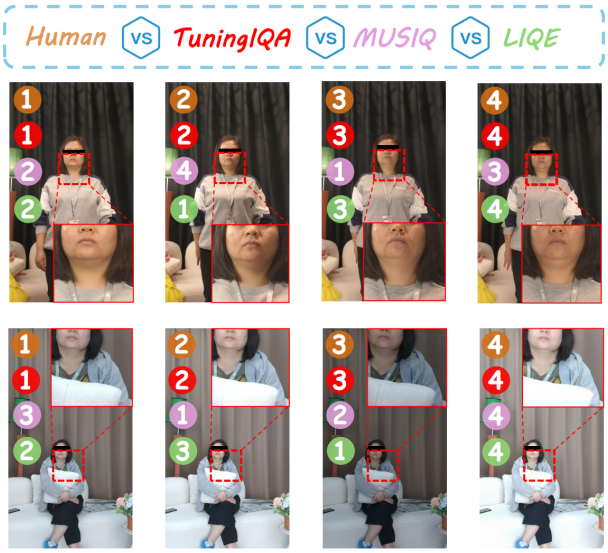

Figure 8: Qualitative comparison of IQA-guided livestreaming camera tuning results (best viewed zoomed in).

Conclusion

This work presents FGLive-10K dataset and TuningIQA, advancing fine-grained BIQA for livestreaming camera tuning. Our findings reveal that existing score-based BIQA methods, while achieving reasonable score regression, fundamentally struggle with fine-grained discrimination essential for camera optimization. Through human-aware feature extraction and graph-based parameter fusion, TuningIQA jointly model multi-attribute regression and fine-grained ranking. Extensive experiments demonstrate TuningIQA's superior performance over existing methods and its practical effectiveness in providing actionable guidance for automated camera parameter optimization.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grants 62471349, 625B2142, 62301378, 62501080 and 62171340, Fundamental Research Funds for the Central Universities under Grants YJSJ25004 and QTZX25076, and partly supported by the China Postdoctoral Science Foundation under Grant 2024M762553.

References

Bosse, S.; Maniry, D.; Wiegand, T.; and Samek, W. 2016. A deep neural network for image quality assessment. In *Proc. IEEE Int. Conf. Image Process.*, 3773–3777.

Ciancio, A.; Targino da Costa, A. L. N. T.; da Silva, E. A. B.; Said, A.; Samadani, R.; and Obrador, P. 2011. No-Reference Blur Assessment of Digital Pictures Based on Multifeature Classifiers. *IEEE Transactions on Image Processing*, 20(1): 64–75.

Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; and Wang, Z. 2020. Perceptual Quality Assessment of Smartphone Photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3674–3683.

Ghadiyaram, D.; and Bovik, A. C. 2016. Massive Online Crowdsourced Study of Subjective and Objective Picture Quality. *IEEE Transactions on Image Processing*, 25(1): 372–387.

Gilstrap, C. A.; and Gilstrap, C. M. 2023. Mobile Technologies and Live Streaming Commerce: A Systematic Review and Lexical Analysis. In *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, 36–44.

Grimson, W. E. L. 1981. *Analysis And Development*, 63–100. n.p.

Hamilton, W. A.; Garretson, O.; and Kerne, A. 2014. Streaming on twitch: fostering participatory communities of play within live mixed media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1315–1324.

He, S.; Zhang, Y.; Xie, R.; Jiang, D.; and Ming, A. 2022. Rethinking Image Aesthetics Assessment: Models, Datasets and Benchmarks. In Raedt, L. D., ed., *Proceedings of the International Joint Conference on Artificial Intelligence*, 942–948.

Hosu, V.; Lin, H.; Sziranyi, T.; and Saupe, D. 2020. KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment. *IEEE Transactions on Image Processing*, 29: 4041–4056.

Huang, Y.; Li, L.; Yang, Y.; Li, Y.; and Guo, Y. 2023. Explainable and Generalizable Blind Image Quality Assessment via Semantic Attribute Reasoning. *IEEE Transactions on Multimedia*, 25: 7672–7685.

Kai, D.; Lu, J.; Zhang, Y.; and Sun, X. 2024. EvTexture: Event-driven Texture Enhancement for Video Super-Resolution. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Kai, D.; Zhang, Y.; Wang, J.; Xiao, Z.; Xiong, Z.; and Sun, X. 2025. Event-Enhanced Blurry Video Super-Resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Katsigiannis, S.; Scovell, J.; Ramzan, N.; Janowski, L.; Corriveau, P.; Saad, M. A.; and Van Wallendael, G. 2018. Interpreting MOS scores, when can users see a difference? Understanding user experience differences for photo quality. *Quality and User Experience*, 3: 1–14.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157.

Li, L.; Sheng, X.; Chen, P.; Wu, J.; and Dong, W. 2025. Towards Explainable Image Aesthetics Assessment With Attribute-Oriented Critiques Generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2): 1464–1477.

Ma, K.; Duanmu, Z.; Wang, Z.; Wu, Q.; Liu, W.; Yong, H.; Li, H.; and Zhang, L. 2020. Group Maximum Differentiation Competition: Model Comparison with Few Samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4): 851–864.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing*, 21(12): 4695–4708.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters*, 20(3): 209–212.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149.

Series, B. 2023. Methodology for the subjective assessment of the quality of television pictures. *Recommendation ITU-R BT*, 500(15).

Sheikh, H.; Sabir, M.; and Bovik, A. 2006. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE Transactions on Image Processing*, 15(11): 3440–3451.

Sheng, X.; Li, L.; Chen, P.; Wu, J.; Dong, W.; Yang, Y.; Xu, L.; Li, Y.; and Shi, G. 2023. AesCLIP: Multi-attribute contrastive learning for image aesthetics assessment. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1117–1126.

Sheng, X.; Xie, P.; Zou, W.; Chen, P.; Zhu, T.; and Li, L. 2025a. InstructCrop: Teaching Multimodal Large Language Models to Crop Aesthetic Images. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 6830–6839.

Sheng, X.; Zou, W.; Chen, P.; Cai, L.; He, C.; and Li, L. 2025b. Text-to-Image Diffusion Models are AI-Generated Image Quality Scorers. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.

Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; and Zhang, Y. 2020. Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3664–3673.

Tan, M.; and Le, Q. 2021. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, 10096–10106. PMLR.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2017. Graph attention networks. *stat*, 1050(20): 10–48550.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2, 2555–2563.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Li, C.; Liao, L.; Wang, A.; Zhang, E.; Sun, W.; Yan, Q.; Min, X.; Zhai, G.; and Lin, W. 2023. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. *arXiv preprint arXiv:2312.17090*.

Zhang, L.; Zhang, L.; and Bovik, A. C. 2015. A Feature-Enriched Completely Blind Image Quality Evaluator. *IEEE Transactions on Image Processing*, 24(8): 2579–2591.

Zhang, W.; Ma, K.; Yan, J.; Deng, D.; and Wang, Z. 2020. Blind Image Quality Assessment Using a Deep Bilinear Convolutional Neural Network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1): 36–47.

Zhang, W.; Zhai, G.; Wei, Y.; Yang, X.; and Ma, K. 2023. Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14071–14081.

Zhu, H.; Wu, H.; Li, Y.; Zhang, Z.; Chen, B.; Zhu, L.; Fang, Y.; Zhai, G.; Lin, W.; and Wang, S. 2024. Adaptive image quality assessment via teaching large multimodal model to compare. *Advances in Neural Information Processing Systems*, 37: 32611–32629.