SageIQ: Scene-Graph-Guided Blind Image Quality Assessment

Renwei Yang, Zhengjie Yang, Yun Wang, Dapeng Oliver Wu, *Fellow, IEEE*, and Shiqi Wang, *Senior Member, IEEE*

***Abstract*—Conventional Blind Image Quality Assessment (BIQA) methods typically assess the entire image quality, which is suboptimal for tasks like autonomous driving that concern specific Task-Aligned Region (TAR). Moreover, we observe that advanced Multimodal Large Language Model (MLLM)-based BIQA models exhibit bias when evaluating small regions, leading to inaccurate perceptual judgments. To address these issues, we propose SageIQ (Scene-graph-guided Evaluation for Image Quality), a pipeline SageIQ-P for TAR localization, an approach consisting of an MLLM-based BIQA model SageIQ-M, and a dataset SageIQ-D. SageIQ-P is designed to automatically identify and evaluate TARs, with the advantages of being training-free and allowing plug-in integration of off-the-shelf BIQA models without retraining. It operates in three stages: scene-graph-based triplet construction, LLM-driven triplet analysis, and integration of weighted BIQA scores into a final assessment. Since SageIQ-P can produce small-sized TAR crops that may encounter small-region scoring bias in existing BIQA models, we propose SageIQ-M to alleviate this bias by injecting scale information through scale-aware images and size-prompted cues, achieving size awareness across both visual and textual modalities. In addition, we develop a fully automated approach to construct a region-level test set SageIQ-D, significantly reducing the human effort needed. Experimental results demonstrate that our methods achieve superior BIQA performance.**

***Index Terms*—Image quality assessment, multimodal, large language model**

I. Introduction

BLIND Image Quality Assessment (BIQA) has been a long-standing research topic and holds significance across various fields. BIQA plays an important role for safer and more reliable autonomous driving systems, as it filters high quality images to ensure accurate decision-making. Different from general BIQA scenarios, where the full image is assessed, in autonomous driving task, only specific regions of the image are concerned, e.g., the moving car (concerned) on the road (not concerned). We term such a concerned region as Task-Aligned Region (TAR), while other unconcerned/uncritical region is Non-TAR (N-TAR). Simply assessing the full image including both TAR and N-TAR potentially results in an unreliable quality score. For example, in Fig. 1, the overall image may appear blurry due to motion distortion, yet the front vehicle remains relatively clear. Using Q-Instruct [1], a popular BIQA algorithm, the overall image quality score is merely 0.14, whereas the front car region alone achieves a remarkably higher score of 0.48. Therefore, the ability to identify TARs is important for BIQA approaches, as it directly affects the decision-making of autonomous driving systems.

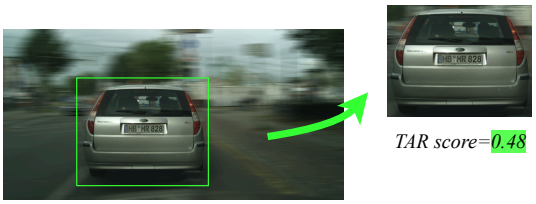

Fig. 1. The overall image has a considerable degree of blurriness, resulting in a quality score of merely 0.14. However, when the car (TAR) is evaluated independently, its quality score increases significantly to 0.48.

With the rapid advancements in machine learning, BIQA methods have migrated from traditional metric based methods [2]–[6] to deep learning based methods [1], [7]–[17], capable of handling more complex visual distortions from pixel-level features. In recent years, Multimodal Large Language Models (MLLMs) like GPT-4v [18] feature the unique ability to process multimodal information and describe image quality in natural language. Thus, MLLMs-based BIQA methods show great potential and attract significant attention from the research community. Early work introduces a vision-language BIQA framework [19]. Q-Instruct [1] constructs task-specific datasets for fine-tuning MLLMs on IQA tasks, followed by IQA benchmarks specifically for MLLMs [17], [20]. Q-Align [21] adopts discrete quality levels for score regression, and DeQA-Score [22] employs distributed soft labels to improve scoring accuracy. Recent works further integrate reinforcement learning for enhanced quality assessment [23], [24]. However, their inability to automatically focus on TAR limits their reliability in tasks like autonomous driving. By assessing the entire image, they become susceptible to interference from N-TARs, which can mislead the system and compromise safety. Furthermore, we observe another critical issue that the state-of-the-art MLLM-based BIQA methods exhibit a pronounced scoring bias towards small regions. This bias stems from the image resizing process required by their Vision Transformer (ViT) [25] encoders. When small TARs are upscaled to a standard input size, the interpolation introduces artificial blur that the model misinterprets as inherent low quality, and thus leading to unfairly suppressed scores.

To address these issues, we introduce SageIQ (Scene-graph-guided Evaluation for Image Quality), an approach that

Renwei Yang, Yun Wang, Dapeng Oliver Wu, and Shiqi Wang are with the Department of Computer Science, City University of Hong Kong, Hong Kong, China. Dapeng Oliver Wu is the corresponding author (dpwu@ieee.org). Zhengjie Yang is with the Hong Kong Generative AI Research and Development Center, The Hong Kong University of Science and Technology, Hong Kong, China. This work was partially supported by the Hong Kong Research Grants Council (RGC) grant #11205424 and Hong Kong Innovation and Technology Commission (ITC) grant MHP/061/23.

includes a training-free pipeline SageIQ-P, an MLLM-based BIQA model SageIQ-M, and a dataset SageIQ-D with region-level labels. We develop **SageIQ-P** to automatically identify and evaluate TARs through a structured three-stage process. Stage 1: a Scene Graph Generation (SGG) module constructs a list of graph triplets (e.g., woman [subject], on [predicate], bike [object]) as a semantic representation of the image, while also producing the spatial locations of these triplets. This list not only identifies entities but also highlights their spatial relationships, providing a comprehensive understanding of the scene required for TAR-level analysis. Stage 2: given the user's task, our triplet analyzer determines triplets most relevant to autonomous driving as TARs. It further evaluates the relative importance weights of the subject and the object within each triplet. Stage 3: the subject and object regions are cropped from the original image. A BIQA model (including SageIQ-M) is then applied to obtain their individual quality scores. Finally, these scores are integrated according to importance weights to produce the triplet quality score. For multiple triplets, their scores are averaged as the image quality score. By implementing the above three stages, our SageIQ-P enables a comprehensive and adaptive evaluation of an image. SageIQ-P allows off-the-shelf BIQA models to be directly plugged into the pipeline in Stage 3 without retraining, which already constitutes an effective TAR-aware evaluation solution. However, SageIQ-P does not directly mitigate the inherent small-region scoring bias of existing BIQA models, and TARs are often small regions in practice. To improve robustness, we therefore propose **SageIQ-M**, which injects scale information to enable more reliable and less biased quality evaluation for small regions. SageIQ-M first generates a Size-Aware (SA) image by embedding multi-scale spatial patterns to retain robust scale information. Then, Learnable Size-Queries (LSQ) and a Size Feature Extractor (SFE) are developed to capture fine-grained size-related characteristics. In the textual branch, image dimensions are explicitly incorporated into the prompt. By incorporating these mechanisms, size information is seamlessly injected into the scoring process, enabling fairer and more robust quality assessments.

We also note that, in the current BIQA literature, datasets with TAR-level labels are lacking. Constructing such a dataset using conventional methods is costly due to (1) intensive manual labor for TAR definition, (2) score annotation requirements, and (3) the need for a dedicated test image collection. In response, based on autonomous driving datasets, we develop a low-cost approach to construct **SageIQ-D**. For (1) and (2), we leverage existing bounding-box annotations in autonomous driving datasets as TARs, with each sample in the dataset containing at least one bounding box. A BIQA model (including SageIQ-M) outputs scores of these predefined TARs as quality score label. For (3), test images are generated by applying diverse types of degradations to non-TAR regions while keeping TARs clear, thereby emphasizing quality disparities to examine whether the BIQA model effectively focuses on TARs. In summary, our contributions are threefold:

- We propose the model SageIQ-M to alleviate scoring bias for small images by leveraging size-related features from both visual and textual modalities. It achieves a fairer and more robust quality evaluation for TARs.
- We propose the pipeline SageIQ-P, which leverages scene graphs and LLM to enable TAR localization and evaluation. Moreover, SageIQ-P is training-free and allows plug-in integration of off-the-shelf BIQA models without retraining.
- We develop an automated method for constructing TAR-level BIQA datasets, addressing the issue of data scarcity with a cost-effective solution.

II. Related Work

A. BIQA methods

1) Traditional Metric Based Methods: Traditional image quality assessment metrics, such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [2], while effective, they have limited applicability in real-world scenarios as they require reference images. To address this, BIQA, which operates without reference images, has gained significant attention. Early methods usually utilized Natural Scene Statistics (NSS) to detect image distortions. domains [3]–[6].

2) Deep Learning Based Methods: With the increasing complexity of visual distortions, hand-crafted BIQA approaches becomes insufficient. Machine learning has enabled BIQA solutions to handle more complex distortions. Starting with LBIQ [26], combining low-level features with regression models to predict quality score. After that, Bosse et al. [7] has proposed a Convolutional Neural Network (CNN)-based BIQA model, significantly improving performance [8], [10], [12], [14], [27]. DB-CNN [8] utilizes dual-stream CNNs to handle synthetic and authentic distortions. Su et al. designs a ResNet-based method [10] to integrate multi-scale and semantic features for prediction. Zhang et al. [12], [14] introduces continual learning for adapting models to various datasets. Transformer-based models [9], [11], [13], [15], [16], [28], [29] have further advanced BIQA by capturing global dependencies and adapting to diverse distortions. In recent times, MLLMs like GPT-4v [18] have demonstrated remarkable capabilities in visual perception tasks, and they exhibit promising generalization ability to handle various distortions. Besides, these models have a unique capability to assess image quality in natural language. As such, they have attracted significant attention from the BIQA research community. Early work by [19] introduces a vision-language framework for BIQA. Subsequently, [1] develops Q-Instruct, which constructed specialized datasets to fine-tune MLLMs for IQA tasks. Meanwhile, benchmarks like Q-Bench [17] and Q-Bench-Video [20] are proposed to quantitatively measure MLLM IQA ability. The field witnessed further advancement with Q-Align [21], which employs discrete quality levels expressed in natural language to train MLLMs for visual scoring. DeQA-Score [22] further enhances this method by incorporating continuous score distribution labels to improve assessment accuracy. More recently, approaches [23], [24] have begun integrating reinforcement learning and semantic reasoning mechanisms, demonstrating promising IQA results.

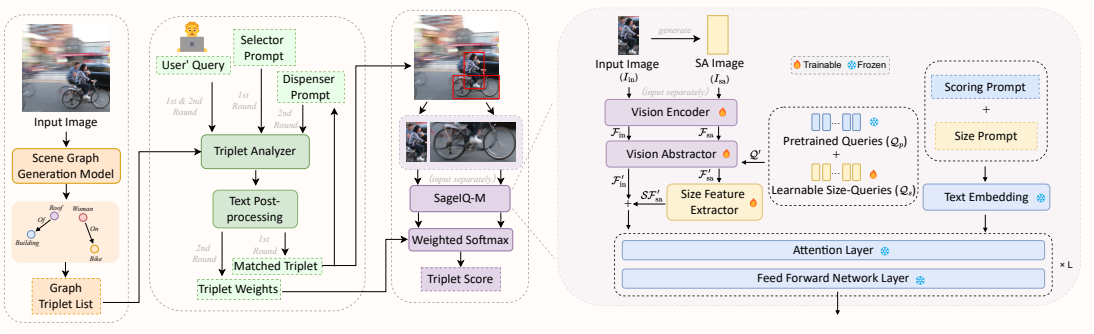

Fig. 2. The left presents our proposed pipeline SageIQ-P, consisting of three steps: triplet construction, triplet analysis, and triplet scoring. For the triplet analyzer, two rounds of analysis are conducted. During triplet scoring, the subject and object are input into the SageIQ-M separately. The right shows SageIQ-M architecture, where the input and SA images are fed into the vision encoder separately.

Nevertheless, their inability to selectively attend to TARs undermines their robustness in applications such as autonomous driving. Because they are likely to be influenced by N-TARs, which may introduce misleading cues and potentially jeopardize system safety. Note that although some prior BIQA methods, such as PaQ-2-PiQ [27] and MANIQA [30], involve region-level scoring, their primary objective remains the prediction of full-image quality. The region-level scores mainly serve as intermediate cues for aggregating global quality estimates. Compared to our work, they lack the capability to automatically identify TARs that are semantically relevant to a given task query.

B. SGG

Scene Graph Generation (SGG) is a key task in computer vision. It identifies entities and their relationships in images, representing them as graphs with nodes for entities and edges for relationships. SGG models are categorized into two-stage and single-stage approaches. Two-stage models separately detect entities and predict relationships [31]–[33], offering strong performance but higher complexity. Single-stage models, such as FCSGG [34], Structured Sparse R-CNN [35], RelTR [36], and SGTR+ [37], integrate entity detection and relationship prediction for greater efficiency, using end-to-end architectures like CNNs and Transformers. By leveraging SGG to extract image semantic information, our method enables TAR-level BIQA. To the best of our knowledge, our work is the first one to integrate SGG into BIQA.

C. LLMs and MLLMs

Large Language Models (LLMs) have shown exceptional linguistic reasoning abilities, driven by advancements in data and model scaling [18], [38]–[42]. Their limitations in visual understanding have spurred research into combining LLMs with Large Vision Models (LVMs) [25], [43], resulting in Multimodal Large Language Models (MLLMs) [44], [45]. MLLMs integrate LLMs' linguistic reasoning with LVMs' visual processing capabilities, enabling multimodal understanding for tasks involving complex visual data. In this work, LLMs are employed to support reasoning in our proposed SageIQ-P, while the novel SageIQ-M, adopting MLLM backbone, is designed to evaluate image quality effectively.

III. Proposed Method

This section presents our methodology in a top-down manner: outlining the pipeline SageIQ-P, describing the BIQA model SageIQ-M, and introducing the datasets SageIQ-D.

A. SageIQ-P

SageIQ-P provides the ability to localize and evaluate the quality of critical TARs. As shown in Fig. 2, it includes three key steps: 1. triplet construction, 2. triplet analysis, and 3. triplet scoring.

1) Triplet Construction: To help determine TARs for an image, we first implement Scene Graph Generation (SGG) using an advanced RelTR [36] model. In Fig. 2, yellow zone, SGG outputs several triplets to form the graph triplet list, and note that the SGG module also outputs the spatial locations of these triplets. Each triplet contains a subject, an object, and a predicate, which encapsulates the semantic and structural information of the scene, providing a comprehensive understanding of the image. The triplet also provides contextual information that goes beyond isolated entities, capturing how they relate to each other. For example, *car on road* carries a different semantic meaning than *car in garage*, even though the subject is the same. These triplets serve as the essential input for the subsequent triplet analysis step.

TABLE I
Triplet Analyzer Text Generation Configuration Parameters

Parameter	Value
do_sample	True
temperature	0.1
top_k	20
top_p	0.2
max_new_tokens	256

2) Triplet Analysis: After obtaining the structural information of an image in SGG, the next step is to analyze it with our proposed triplet analyzer, as shown in Fig. 2, green zone. Triplet analyzer is built on top of LLaMA3 model [42], upon which we develop selector and dispenser prompts, generation configurations, and text post-processing module, guiding the model to select the matched triplet and dispense triplet weights. The triplet analyzer configuration is shown in Table I. Specifically, these parameters regulate the randomness and diversity of the generated text: `do_sample=True` activates stochastic sampling instead of

TABLE II
PROMPTS USED IN SAGEIQ

Prompt Type	Prompt Content
Selector Prompt	You are an advanced AI assistant specializing in analyzing data from car cameras. Your task is to evaluate the relevance of observed items captured by the camera during a drive. The captured items are: {graph_triplet_list}. Identify the items that are most relevant to {user_query}, the most significant items should be listed first. Please return the numbered indices. Provide your response in the following format: Selected items: [X, Y, Z].
Dispenser Prompt	You are an AI assistant for {user_query}. Analyze two entities, {triplet_subject} and {triplet_object}, and assign weights value according to importance(1 = minimal, 10 = critical). Consider: 1. Safety: Effects on driving safety. 2. Context: Interaction with the vehicle or each other. 3. Relevance: Impact on driving decisions. Pay special attention to entities such as 'pedestrian' or 'vehicle'. Your response format: Object: {triplet_subject}, Weight: X; Object: {triplet_object}, Weight: Y.
Scoring prompt	Rate the quality of this image <|image|>.
Size prompt	The image size is {Image size}.

greedy decoding, and the `temperature` parameter (0.1) controls the degree of randomness, where lower values lead to more deterministic outputs. The `top_k` parameter (20) limits the candidate token pool to the K most probable next tokens, effectively filtering out low-probability noise, while the nucleus sampling `top_p` parameter (0.2) further constrains the selection to the smallest set of tokens whose cumulative probability mass does not exceed 0.2, ensuring both fluency and coherence. This configuration allows the model to generate stable and reliable outputs that emphasize stability and safety in autonomous driving scenarios, with the maximum output length (`max_new_tokens=256`) preventing overly long or redundant responses. These hyper-parameters follow widely adopted configurations for stable text generation in instruction-following LLMs, as recommended in prior works (e.g., the LLaMA-3 official implementation [42]).

The triplet analyzer conducts 2 rounds of analysis. In the first round, it is instructed by the selector prompt shown in Tab. II, to select the matched triplet by analyzing the user's query and graph triplet list. Note that in this step, the complete triplet structure, including the subject, predicate, and object, is fed into the triplet analyzer to provide comprehensive semantic information. In this work, user's query is set as *autonomous driving*. In the second round, we develop a weights dispensing mechanism. The dispenser prompt in Tab. II instructs the triplet analyzer to dispense subject and object weights, aiming to reflect their semantic and functional significance. Each specific triplet has its weights $\omega_i, i \in \{1, 2\}$, corresponding to subject and object respectively. The weights indicate the contribution of each entity to the final quality score. Please note in the case of multiple matched triplets, the second round is performed iteratively for each triplet.

Despite the specifically designed prompts and generation configurations, the issue of extraneous tokens may still manifest. To this end, we develop a text post-processing module. It first removes redundancy by employing regular expressions to precisely extract required information as shown in Tab. III. Specifically, the first regular expression is designed for triplet selection, capturing the list of selected items enclosed in square brackets (`[...]`, where parentheses `()` capture the content and `?` ensures minimal matching). The second pattern is used for weight dispensing, extracting the object name and its corresponding numeric score, where `\w+` matches words and `\d+` matches numbers. Then, the information is converted from string into numerical format, making it usable as input for subsequent modules.

TABLE III
REGULAR EXPRESSIONS FOR TEXT POST-PROCESSING

Task	Regular Expression
Triplet Selection	`Selected items:\s*\[(.*?)\]`
Weight Dispensing	`Object:\s*(\w+),\s*Score:\s*(\d+)`

Algorithm 1 Triplet Scoring

Input:
1) Subject x_1 and object x_2.
2) Scoring prompt $\mathcal{S}$.
3) SageIQ-M Θ.
4) Quality keyword token indices $\mathcal{P} = \{p_g, p_a, p_p\}$.
5) Quality score vector $\mathbf{q} = [1, 0.5, 0]^\top$.
6) Subject weight ω_1 and object weight ω_2 for a specific triplet.

1: Use the SageIQ-M Θ with the input triplet and scoring prompt $\mathcal{S}$ to obtain logits distributions:

$$\mathcal{L}_i = \Theta(\mathbf{x}_i, \mathcal{S}) \in \mathbb{R}^{\mathcal{V}}, i \in \{1, 2\}.$$

2: **for** $i \in \{1, 2\}$ **do**
3: Extract $\mathbf{v}_i = \mathcal{L}_i|_{\mathcal{P}} \in \mathbb{R}^3, i \in \{1, 2\}$.
4: Probability distribution: $\mathbf{p}_i = \text{softmax}(\mathbf{v}_i)$
5: *where:*

$$p_i^{(j)} = \frac{\exp(v_i^{(j)})}{\sum_{k=1}^{3} \exp(v_i^{(k)})}, \ j \in \{1, 2, 3\}.$$

6: Compute quality score: $s_i = (\mathbf{p}_i)^\top \mathbf{q}$
7: **end for**
8: Normalize weights using softmax:

$$\widehat{\omega}_i = \frac{\exp(\omega_i)}{\sum_{j=1}^{2} \exp(\omega_j)}, i \in \{1, 2\}.$$

9: Compute final triplet score:

$$s = \sum_{i=1}^{2} \widehat{\omega}_i s_i.$$

10: **return** s.

3) Triplet Scoring: Triplet analysis has provided triplet weights for all triplets. Now we are ready to calculate the final score according to triplet weights, as shown in Fig. 2, purple zone. For clear explanation, we elaborate on the case where a single triplet is matched. For the case of multiple matched triplets, the final score is obtained by averaging.

In Alg. 1. The triplet scoring process begins by separately feeding subject x_1 and object x_2, along with a scoring prompt $\mathcal{S}$ (a string as illustrated in Tab. II), into the SageIQ-M Θ to generate logits distributions $\mathcal{L}_i \in \mathbb{R}^{\mathcal{V}}, i \in \{1, 2\}$ (Line 1 in Alg. 1). SageIQ-M Θ is our proposed model, explained in Sec.III-B. $\mathcal{L}_i$ is the output of the last feed forward network layer of Θ, representing the predicted probabilities of words in

the vocabulary. $\mathcal{V}$ denotes the vocabulary size. $\mathcal{S}$ instructs the model to evaluate image quality. The indices $\mathcal{P}$ correspond to the positions of the quality-related keywords *good* (p_g), *average* (p_a), and *poor* (p_p) in the vocabulary, while the quality score vector $\mathbf{q}$ quantifies these keywords with scores of 1, 0.5, and 0, respectively. Then, following the method described in [17], logits vectors $\mathbf{v}_i$ are extracted from $\mathcal{L}_i$ according to $\mathcal{P}$ (Line 3). Next, they are softmax normalized to produce probability distributions $\mathbf{p}_i \in \mathbb{R}^3$ (Line 4). By performing inner products with $\mathbf{q}$ (Line 6), quality scores $s_i, i \in \{1, 2\}$ are obtained. To aggregate these scores, we softmax normalize triplet weights ω_i obtained from triplet analyzer to produce $\widehat{\omega}_i$ (Line 8). The softmax function provides adaptive weighting: it amplifies the contribution of the entity more likely to be a TAR (with higher weight) and suppresses the other one. When both entities are likely TARs, their normalized weights become comparable, resulting in a more balanced influence. Then, quality scores are weighted combined to compute the final triplet quality score s (Line 9).

The proposed scoring mechanism integrates triplet weights in triplet scoring process, aiming to adaptively combine subject and object scores based on their importance, ensuring s accurately reflects the triplet quality from the perspective of user's query.

B. SageIQ-M

Existing MLLM-based BIQA models exhibit a scoring bias toward small images because the CLIP vision encoder [46] built on ViT [25] requires uniform image resizing, causing small images to be stretched and blurred. This undermines assessment fairness and accuracy. To address this, SageIQ-M integrates size information across image and text modalities to capture scale-aware features and mitigate this bias.

As shown Fig. 2, SageIQ-M is an MLLM-based BIQA model, receiving inputs of image and text modalities. It adopts mPLUG-Owl2 [47] as the backbone, with weights pretrained on both mPLUG-Owl2 and Q-Instruct [1] datasets. On top of the backbone, we propose brand-new SA image, Learnable Size-Queries (LSQ), Size Feature Extractor (SFE), and size prompt to enhance the ability to extract size-related features. Input image I_{in} and a generated SA image I_{sa} are separately fed into the vision encoder, producing image features $\mathcal{F}_{\text{in}}$ and $\mathcal{F}_{\text{sa}}$, respectively. They are then individually passed to the vision abstractor, to perform cross-attention with the combination of Pretrained Query and LSQ, generating $\mathcal{F}'_{\text{in}}$ and $\mathcal{F}'_{\text{sa}}$. Subsequently, $\mathcal{F}'_{\text{sa}}$ is passed to the SFE, yielding the more refined size feature $\mathcal{SF}'_{\text{sa}}$. It is then combined with $\mathcal{F}'_{\text{in}}$ via element-wise addition, so that the size related features seamlessly integrated with the general image features. On the textual modality, size information is explicitly provided to the model via size prompt, as shown in Tab. II. Then, they are embedded into textual features, combined with image features, sent into the pretrained attention layers and feed forward layers to obtain scoring results.

1) SA image: We develop SA image I_{sa}, aiming to provide size features to model. I_{sa} is generated by embedding multi-scale spatial patterns that vary smoothly with position and frequency, as inspired by ViT position encoding [25]. As illustrated in Alg. 2, the process begins by computing a scaling factor array $div[k]$ for $k \in [0, \lfloor C/2 \rfloor]$ (Line 1 in Alg. 2), which controls the frequency of the patterns across channels. For each channel $i \in [0, C-1]$, the algorithm assigns pixel values based on the channel index. For even channels, the pixel values are the sum of sine functions applied to the spatial coordinates (y, x) (Line 4). For odd channels, cosine functions are used instead (Line 6). Alternating sine and cosine functions across channels introduces phase diversity, enhancing inter-channel distinctiveness in terms of size information. This SA image design provides two significant advantages. First, it ensures that size-related information is effectively preserved even after preprocessing operations, such as resizing and rescaling, because images of different original sizes produce distinguishable pattern responses after transformation, allowing the encoded size features to remain robust and distinguishable. Second, the SA image provides only size-related information while eliminating all semantic content. Its processing pathway focuses solely on size-related features, simplifying the task and improving efficiency.

2) LSQ: To address the limitation of Pretrained Query that solely extracts general features, LSQ guides the model to focus on size related features via cross attention. This is particularly effective for I_{sa}, which exclusively contains size information.

Algorithm 2 SA image Generation

Input: Input image $I_{\text{in}} \in \mathbb{R}^{H \times W \times C}$

1: Compute scaling factor:

$$div[k] = \exp\left(-\frac{k \cdot \log(10000)}{C}\right),$$

where $k \in [0, \lfloor C/2 \rfloor]$.

2: **for** $i \in [0, C-1]$ **do**

3: **if** $i \,\%\, 2 = 0$ **then** ▷ Even channels

4: Fill I_{sa} with sine values:

$$I_{\text{sa}}(y, x, i) = \sin\left(y \cdot div[\lfloor i/2 \rfloor]\right) + \sin\left(x \cdot div[\lfloor i/2 \rfloor]\right),$$

5: **else** ▷ Odd channels

6: Fill I_{sa} with cosine values:

$$I_{\text{sa}}(y, x, i) = \cos\left(y \cdot div[\lfloor i/2 \rfloor]\right) + \cos\left(x \cdot div[\lfloor i/2 \rfloor]\right),$$

7: **end if**

where $y \in [0, H-1]$ and $x \in [0, W-1]$

8: **end for**

9: **return** I_{sa}.

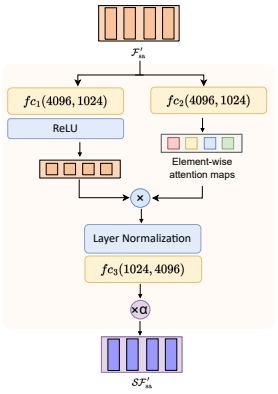

Fig. 3. The SFE architecture. fc_1(4096,1024) indicates the fully connected layers, where input and output dimensions are 4096 and 1024, respectively. Similar for fc_2 and fc_3.

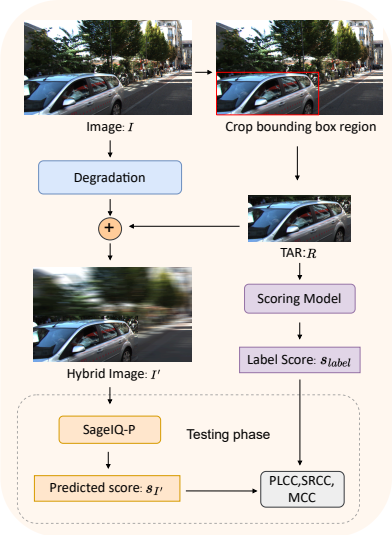

Fig. 4. The procedure to build SageIQ-D.

To elaborate, Pretrained Query ($\mathcal{Q}_p$) and LSQ ($\mathcal{Q}_s$) have aligned shape $[b, \mathcal{H}_n, \mathcal{H}_s]$, where b is the batch size, $\mathcal{H}_n$ is hidden state number 64, and $\mathcal{H}_s$ is hidden state size 1024. To fuse size related features with general image features, $\mathcal{Q}_s$ are combined with $\mathcal{Q}_p$ via element-wise addition to form a new query $\mathcal{Q}'$. It serves as a query in Vision Abstractor, conducting cross attention with the Vision Encoder output $\mathcal{F}_{\text{in}}$ or $\mathcal{F}_{\text{sa}}$, producing the cross-attention output $\mathcal{F}'_{\text{in}}$ or $\mathcal{F}'_{\text{sa}}$ as the Vision Abstractor output.

3) SFE: SFE is designed to further extract and refine size-related features, to enhances the model's robustness and accuracy in evaluating images of varying sizes. Its architecture is shown in Fig. 3. To elaborate, input $\mathcal{F}'_{\text{sa}}$ is fed into a fully connected layer fc_1 with ReLU to generate refined feature vectors. Meanwhile, $\mathcal{F}'_{\text{sa}}$ is also processed by fc_2 to produce element-wise attention map to highlight size related positions. It is then combined with refined feature vectors through element-wise multiplication, emphasizing size-relevant characteristics. Next, layer normalization is employed to stabilize training and ensure consistent feature scaling. Following this, fc_3 projects the features back into the original dimensional space. Eventually, with a learnable parameter α adjusting feature influence, the SFE output $\mathcal{SF}'_{\text{sa}}$ is produced.

4) Scoring and Size Prompts: As shown in Tab. II, the scoring prompt straightforwardly instructs the SageIQ-M to assess the image. Within this prompt, <|image|> is a special placeholder for multimodal inputs, which is replaced by the actual image feature embedding during the token embedding stage. Besides, we further design the size prompt to explicitly inject size related information into the model's text feature space. It enables size-aware assessment, which is often overlooked in previous approaches. In this prompt, the `Image size` serves as a dynamic text variable, dynamically retrieving the width and height of the image in real time.

C. SageIQ-D Construction

We propose a fully automated SageIQ-D construction method, as illustrated in Fig. 4. The process begins with an input image I sourced from widely used autonomous driving datasets: KITTI [48], Cityscapes [49], and Waymo [50]. These datasets provide annotated bounding boxes, which we leverage as predefined TARs. Each image contains at least one bounding box, and at most five bounding boxes are used per image. Each TAR is passed through a pretrained scoring model (including SageIQ-M), yielding a label score denoted as s_{label} for the image I. The scoring model used here is the same one employed in the third step Triplet Scoring of SageIQ-P. In the case of multiple TARs, their individual scores are averaged to obtain a single label. Concurrently, the input image I is subjected to diverse degradations to emulate realistic low-quality conditions, after which the original clear TAR is overlaid back onto it, producing a hybrid image I'. Each type of degradation is individually employed to construct a dedicated dataset for testing. Specifically, the degradation types are as follows. Horizontal motion blur (kernel size $= 30$), Gaussian blur (kernel size $= 31$, $\sigma = 10$), mean blur (kernel size $= 20$), median blur (kernel size $= 21$), additive Gaussian noise (mean $= 0$, $\sigma = 50$), strong JPEG compression (quality factor $= 5$), salt-and-pepper noise (noise amount $= 0.1$, and salt-to-pepper ratio $= 0.5$). Such a construction introduces a pronounced quality disparity between the TAR and its degraded context, thereby serving as a means to assess whether the model focuses specifically on TAR quality. During the testing phase, I' serves as input to SageIQ-P, predicting a quality score $s_{I'}$. A point to emphasize is that, the goal of SageIQ-P is to ensure $s_{I'}$ closely matches s_{label}, highlighting its ability to accurately assess TAR while effectively disregarding background distortions. To quantitatively evaluate the correlation between $s_{I'}$ and s_{label}, we computed the widely used Spearman's Rank Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), and their mean value, Mean Correlation Coefficient (MCC).

To conclude, this automated methodology not only eliminates the need for costly manual annotation, but also provides a robust testbed for evaluating IQA methods in TAR-focused quality assessment tasks.

IV. Experiments

A. Experimental Setups

1) Dataset: LIVE-itw [51], a common IQA dataset comprising 1,162 naturally distorted images, is used for SageIQ-M training and testing in Section IV-B. The dataset is split into approximately 80% for training (930 images) and 20% for testing (232 images). KITTI [48], Cityscapes [49], and Waymo [50] are widely used in autonomous driving tasks and primarily focus on street scenes. We construct the SageIQ-D from these publicly available sources: 7,481 images from KITTI, 5,000 images from Cityscapes, and the first frame from 1,000 scenes in Waymo. These three datasets are used as test datasets to evaluate the SageIQ-P pipeline in Section IV-C.

2) Training Settings: The experiments use an NVIDIA RTX 5000 Ada GPU (32GB) with the PyTorch 2.0.1 deep learning framework. The Adam optimizer [52] is employed with a learning rate of 1e-5, and the model is trained for a total of 5 epochs. LSQ, SFE, and bias term of the vision encoder

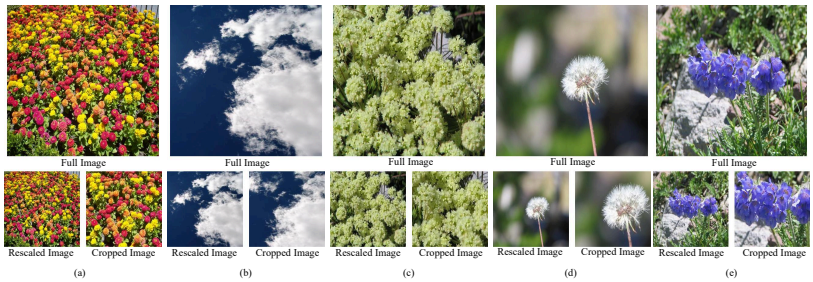

Fig. 5. The illustrative images for showing scoring bias. The first row shows the original full images, and the second row shows the corresponding rescaled and cropped images.

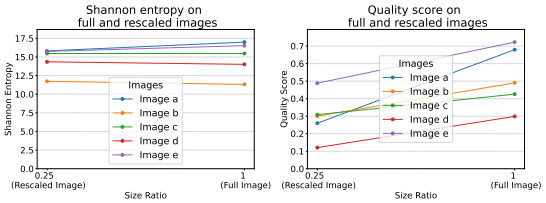

Fig. 6. Shannon entropy and quality scores for full and rescaled images. The horizontal axis represents the size ratio, where 1 corresponds to the original full image and 0.25 indicates the rescaled image with both its width and height reduced to half.

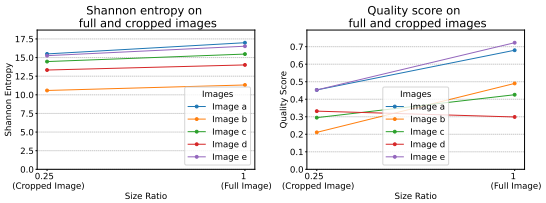

Fig. 7. Similar to Fig. 6, but applied to cropped images.

and abstractor are trainable. The training phase utilizes the 16-bit brain floating point data type, which maintains wide value range while effectively accelerates the training process.

B. Model SageIQ-M

1) Size Problem Investigation: The existing MLLM-based BIQA models exhibit a notable scoring bias toward small images, showing a tendency to assign them lower scores. In this section, we aim to investigate this issue through a series of experiments. To evaluate the scoring bias, we generate smaller images from an original image using two resizing strategies: rescaling and cropping. In the rescaling strategy, the image is proportionally reduced to half its original width and height. In contrast, the cropping strategy extracts the central region of the image while preserving half its original dimensions.

Shannon entropy [53] is adopted as a quantitative metric, as it reflects the structural richness and detail that are key attributes of perceptual quality [2]. We then compare the Shannon entropy with the quality scores predicted by the baseline model (Q-Instruct). Concretely, Shannon entropy H is defined as:

$$H = -\sum_{i=1}^{m} p_i \log_2 p_i, \tag{1}$$

where p_i is the probability of the i-th unique intensity value in the image, and m is the number of unique intensity values. The probability p_i is computed as

TABLE IV
SHANNON ENTROPY AND QUALITY SCORE CORRELATIONS FOR RESCALED AND CROPPED IMAGES

	Rescaled Image		Cropped Image	
	Shannon Entropy	Quality Score	Shannon Entropy	Quality Score
SRCC	1.000	0.600	1.000	0.564
PLCC	0.978	0.759	0.996	0.683
MCC	0.989	0.680	0.998	0.624

$$p_i = \frac{n_i}{N}, \forall i = 1, 2, \ldots, m, \tag{2}$$

where n_i is the occurrences count of the i-th intensity value and N is the total number of pixels in the image. Shannon entropy measures the diversity and randomness of pixel intensities within an image, with higher entropy values indicating greater complexity and lower values reflecting simpler images.

For illustration, Fig. 5 presents the images along with their corresponding rescaled and cropped versions. Left line charts of Fig. 6 and Fig. 7 present the Shannon entropy values for full images and their size reduced versions (rescaled and cropped). The results indicate that size reduction introduces only minor fluctuations in entropy, while overall stability is maintained. Conversely, the right line charts display the quality scores predicted by the baseline model, and these scores are notably sensitive to changes in image size. Please note that an exception occurs for image d in Fig. 7, where the quality score improves due to the removal of large blurry regions in the original image. This aligns with our previously discussed limitations that, existing IQA models often fail to accurately reflect the quality of user-interest regions and are primarily influenced by large background areas.

To further analyze the size effects, Tab. IV provides the correlation coefficients for the size reduced images (rescaled and cropped). To elaborate, we computed the SRCC, PLCC, and MCC by comparing the Shannon entropy and quality scores of the full and size reduced images. The results show a high correlation for Shannon entropy. In particular, its SRCC reaches the maximum value 1, and PLCC and MCC are close to 1, indicating that Shannon entropy remains relatively stable after size reduction. In contrast, the quality scores exhibited a significant decrease in correlation, suggesting that the baseline model struggles to maintain consistency when image sizes are altered.

2) Datasets Similarity Analysis: In this study, we intentionally adopt the LIVE-itw dataset, a non-autonomous-driving dataset, to train SageIQ-M, rather than leveraging the widely used autonomous driving datasets (ADDs) such as Cityscapes,

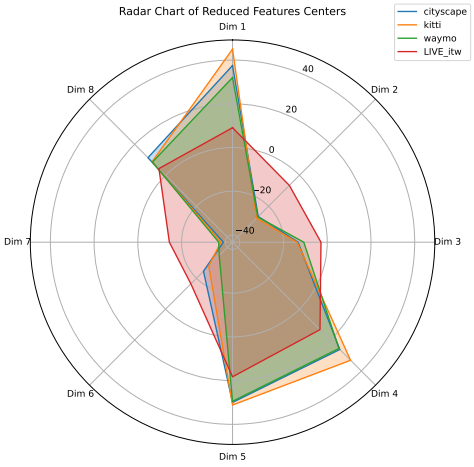

Fig. 8. Feature centers of 4 datasets, with GRP applied to reduce feature dimensions. The LIVE-itw dataset exhibits a distinct distribution pattern, while the other three datasets are similar.

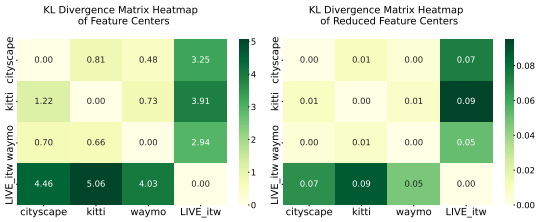

Fig. 9. The KL divergence matrices for feature centers of 4 datasets. The left is for the original feature centers, the right is for the reduced version after applying GRP. Darker colors represent higher values, indicating greater disparities.

KITTI, or Waymo, which will be to build SageIQ-D. While this choice introduces greater challenges, it serves to more effectively assess the generalization capability of the model. The rationale behind this decision is grounded in the observation that existing ADDs exhibit a high degree of feature-domain similarity, which makes it difficult to disentangle genuine model robustness from performance gains due to distributional overlap. Training and testing on similarly structured datasets may thus lead to overly optimistic evaluations that fail to reflect a model's true capacity. In the following analysis, we show that these ADDs have high similarity, while LIVE-itw stands out as distinct.

The dataset feature center analysis, as visualized in Fig. 8, reveals an obvious contrast between ADDs and LIVE-itw. The analysis procedure is outlined as follows. Firstly, images from each dataset are resized and normalized, with their feature vectors extracted using a pretrained ResNet50 model [55]. These feature vectors are flattened into 1-dimensional representations, and then within each dataset, the average feature vector across all images is computed as the feature center, denoted as $c_i \in \mathbb{R}^L$. Here, $L = 2048$ is ResNet50 feature vector length. $i \in [1, N]$ is dataset index, with $N = 4$. Then, the feature center matrix $C = \begin{bmatrix} c_1 & \cdots & c_N \end{bmatrix}^\top \in \mathbb{R}^{N \times L}$.

To facilitate comparisons and visualizations, we employ Gaussian Random Projection (GRP) [56] to reduce the dimensionality. This technique is based on random linear transformation. To achieve this, we firstly construct a random projection matrix $R \in \mathbb{R}^{L \times L'}$, where L' is the reduction feature vector length. Each element in R is drawn independently from a

TABLE V
PERFORMANCE COMPARISON ON RESCALED AND CROPPED IMAGES

Model	Rescaled Images			Cropped Images		
	SRCC	PLCC	MCC	SRCC	PLCC	MCC
Q-Instruct	0.643	0.693	0.668	0.857	0.974	0.916
Q-Align	0.648	0.654	0.651	0.850	0.853	0.852
DeQA-Score	0.653	0.663	0.658	0.871	0.867	0.869
SageIQ-M	**0.797**	**0.754**	**0.775**	**0.975**	**0.988**	**0.981**

Fig. 10. SageIQ-D visual examples. The first to third rows are original images, hybrid images, and hybrid images with detected TARs. The fourth to sixth rows are label scores, scores predicted without and with SageIQ-P, using SageIQ-M, Q-instruct [1], and TRES [54].

standard normal distribution $\mathcal{N}(0, 1)$. Then, the dimensionality reduction is performed by:

$$C' = C \cdot R = \begin{bmatrix} c'_1 & \cdots & c'_N \end{bmatrix}^\top, \tag{3}$$

where $C' \in \mathbb{R}^{N \times L'}$ represents the feature center matrix after dimensionality reduction. According to the Johnson-Lindenstrauss lemma [57], GRP can preserve the pairwise Euclidean distances between points with high probability. Therefore, for $\forall i, j \in [1, N]$ the following inequality holds:

$$(1 - \varepsilon)\|c_i - c_j\|^2 \leq \|c'_i - c'_j\|^2 \leq (1 + \varepsilon)\|c_i - c_j\|^2 \tag{4}$$

where $\varepsilon \in [0, 1]$ is a distortion parameter. This property ensures that the reduced-dimension matrix C' retains the essential characteristics of the original matrix C. C' is visualized in Fig. 8. As shown in this radar chart, the feature distributions of Cityscapes, KITTI, and Waymo demonstrate significant overlap and exhibit highly similar shapes. Particularly, they have pronounced peaks in dimension 1, 4, and 5, and valleys in dimension 2 and 7. In contrast, the distribution of the LIVE-itw dataset is comparatively more balanced and distinct from the other three datasets.

To further quantitatively evaluate the similarity, we compute the Kullback-Leibler (KL) divergence [58] matrix for C and C', as shown in Fig. 9. The KL divergence is used to measure how one probability distribution diverges from another. To formulate the discrete probability distribution, we first normalize C to $\widehat{C}$:

$$\widehat{C} = \begin{bmatrix} \widehat{c}_1 & \cdots & \widehat{c}_N \end{bmatrix}^\top = \begin{bmatrix} \frac{c_1}{\|c_1\|} & \cdots & \frac{c_N}{\|c_N\|} \end{bmatrix}^\top. \tag{5}$$

Then, the KL divergence $D_{KL}(\widehat{c}_i \| \widehat{c}_j)$ is computed as:

$$D_{KL}(\widehat{c}_i \| \widehat{c}_j) = \sum_{k=1}^{L} \widehat{c}_{ik} \log \frac{\widehat{c}_{ik}}{\widehat{c}_{jk}}, i, j \in [1, N], \tag{6}$$

TABLE VI
SAGEIQ-M ABLATION STUDY RESULTS ON RESCALED AND CROPPED IMAGES

Model	Modules			Rescaled Images						Cropped Images					
	Learnable Size-Queries	Size Feature Extractor	Size Prompt	SRCC	PLCC	MCC	ΔSRCC	ΔPLCC	ΔMCC	SRCC	PLCC	MCC	ΔSRCC	ΔPLCC	ΔMCC
Baseline Model	✗	✗	✗	0.643	0.693	0.668	-	-	-	0.857	0.974	0.916	-	-	-
Ours	✗	✗	✔	0.786	0.748	0.767	0.143	0.055	0.099	0.865	0.981	0.923	0.008	0.007	0.007
	✗	✔	✗	0.714	0.722	0.718	0.071	0.029	0.050	0.964	0.986	0.975	0.107	0.012	0.059
	✗	✔	✔	0.786	0.711	0.748	0.143	0.017	0.080	0.893	0.980	0.936	0.036	0.006	0.021
	✔	✗	✗	0.750	0.704	0.727	0.107	0.011	0.059	0.971	0.986	0.978	0.114	0.011	0.063
	✔	✗	✔	0.739	0.733	0.736	0.096	0.040	0.068	0.883	0.983	0.933	0.026	0.009	0.017
	✔	✔	✗	0.714	0.736	0.725	0.071	0.043	0.057	0.964	0.987	0.976	0.107	0.013	0.060
	✔	✔	✔	0.797	0.754	0.775	**0.154**	**0.061**	**0.107**	0.975	0.988	0.981	**0.118**	**0.013**	**0.066**

TABLE VII
TRIPLET ANALYSIS WORKFLOW

Round of Analysis	Round 1: Select matched Triplet	Round 2: Dispense Triplet Weights
Input	Graph triplet list: [1. window on building, 2. car on street]	Triplet Subject: car, Triplet Object: street
Prompt	Selector prompt	Dispenser prompt
Analyzer Response	Based on the options provided, **selected items: [2]**. This is because the car on the street is a critical component of the autonomous driving system, as it provides the necessary data for the system to make decisions. Therefore, option 2 is important for autonomous driving. The number index is 2.	I have analyzed the objects car and street, and assigned triplet weights based on their importance to autonomous driving: **- Subject: car, Weight: 8**, **- Object: street, Weight: 5**. The car is assigned a score of 8 because it is directly relevant to autonomous driving decisions, such as determining the vehicle's speed and trajectory. The street is assigned a score of 5 because it is indirectly relevant to autonomous driving decisions, such as determining the vehicle's position and orientation.
Output	**Triplet Index: 2**	**Weights Scores: [8, 5]**

where $\widehat{c}_{ik}$ represents the k-th elements of $\widehat{c}_i$. Next, using pairwise KL divergence values, we construct an KL divergence matrix $\mathbf{D} \in \mathbb{R}^{N\times N}$ defined as:

$$\mathbf{D} = \left[D_{KL}(\widehat{c}_i \| \widehat{c}_j)\right]_{i,j=1}^{N}. \tag{7}$$

The same procedure is also applied to C' (normalization and matrix construction), acquiring its KL divergence matrix $\mathbf{D}'$. Fig. 9 visualizes $\mathbf{D}$ and $\mathbf{D}'$ as heat maps. The color intensity of each grid cell represents the $D_{KL}(\widehat{c}_i \| \widehat{c}_j)$ value, where darker colors indicate larger values. The diagonal elements are zero since they represent self KL divergence. The light grid cells between the Cityscape, KITTI, and Waymo datasets suggest their features are distributed in a similar manner. In contrast, the dark cells between LIVE-itw and other datasets reflect pronounced differences in their feature distributions.

3) Experimental Results and Analysis: This section evaluates the performance of several MLLM-based IQA models in handling image scale variations on LIVE-itw dataset. We compare models including Q-Instruct (mPLUG-Owl2 base), Q-Align, DeQA-Score, and our proposed SageIQ-M. The comparison is conducted by feeding both the full-resolution image and its rescaled or cropped version into the same model, and then evaluating the correlation between the two predicted quality scores. During SageIQ-M training, the original full-resolution image is input into Q-Instruct to generate s_{full} as labels. The image is then rescaled or cropped to half size and input into SageIQ-M to obtain predicted score s_{resc} or s_{crop}. The loss is computed as the mean squared error between labels and predicted scores. This strategy encourages the model to produce scale-consistent quality scores, ensuring robustness to input resolution.

As shown in Tab. V, existing MLLM-based IQA models exhibit lower performance than SageIQ-M, indicating that their quality predictions are more sensitive to image resolution. This problem arises from the image resizing operation required by their ViT encoders. When small images are upscaled to a fixed input size, the interpolation process introduces artificial blur. The models tend to misinterpret this blur as inferior image quality, leading to biased and underestimated scores. In contrast, SageIQ-M achieves higher correlation values, suggesting the effectiveness of our proposed approach: by explicitly incorporating size-aware features, SageIQ-M has mitigated the scoring bias, resulting in more consistent and robust quality assessment. We also find that rescaled images yield lower correlations than cropped images, this is probably because the two interpolation steps (scalding down and up) cause more detail information loss.

4) SageIQ-M Ablation Study: In Tab. VI, we conduct ablation studies to compare the baseline model and SageIQ-M, by systematically enabling or disabling three key modules introduced in our approach: LSQ, SFE, and the size prompt. When all three modules are disabled, the model reduces to the baseline model, i.e. Q-Instruct. Notably, the SA image is used only with SFE enabled to extract size-related features; otherwise, it is not used. We can observe in Tab. VI that, for both rescaled and cropped images, by integrating the proposed modules, all three metrics have been improved. When all modules are enabled, the SageIQ-M has achieved the best performance. These experiments validate the effectiveness of the proposed modules in improving the model's robustness to size variations introduced by rescaling and cropping. Consequently, the proposed modules successfully enhance the model's ability to mitigate scoring bias on small images.

C. Pipeline SageIQ-P

1) Experiment Examples: Tab. VII provides an illustrative example of triplet analysis. In the first round, the graph triplet list serves as the input, which is derived from the first step of SageIQ-P. This list, along with selector prompt, is fed into the triplet analyzer to select the matched triplet. In the second round, the matched triplet and dispenser prompt are

TABLE VIII
BIQA MODEL PERFORMANCE WITHOUT AND WITH SAGEIQ-P

Model	Dataset	SRCC w/o SageIQ-P	PLCC w/o SageIQ-P	MCC w/o SageIQ-P	SRCC w/ SageIQ-P	PLCC w/ SageIQ-P	MCC w/ SageIQ-P	ΔSRCC	ΔPLCC	ΔMCC
ARNIQA [59]	Cityscapes	0.185	0.156	0.170	0.466	0.180	0.323	0.281	0.024	0.153
	Kitti	0.597	0.614	0.606	0.605	0.554	0.579	0.008	-0.060	-0.026
	Waymo	0.274	0.223	0.248	0.468	0.509	0.488	0.194	0.286	0.240
	Average	0.352	0.331	0.341	0.513	0.414	0.463	0.161	0.083	0.122
TOPIQ [60]	Cityscapes	0.191	0.080	0.136	0.239	0.120	0.179	0.048	0.039	0.044
	Kitti	0.275	0.458	0.367	0.641	0.648	0.644	0.365	0.190	0.278
	Waymo	0.574	0.621	0.597	0.735	0.726	0.731	0.162	0.105	0.134
	Average	0.347	0.386	0.366	0.538	0.498	0.518	0.192	0.111	0.152
TRES [54]	Cityscapes	0.012	0.143	0.078	0.580	0.620	0.600	0.568	0.477	0.523
	Kitti	0.071	0.046	0.058	0.605	0.690	0.647	0.534	0.643	0.589
	Waymo	0.521	0.557	0.539	0.641	0.585	0.613	0.121	0.028	0.074
	Average	0.201	0.249	0.225	0.609	0.632	0.620	0.408	0.383	0.395
MANIQA [30]	Cityscapes	0.398	0.554	0.476	0.135	0.156	0.146	-0.263	-0.398	-0.331
	Kitti	0.235	0.035	0.135	0.552	0.672	0.612	0.317	0.637	0.477
	Waymo	0.494	0.568	0.531	0.641	0.652	0.647	0.147	0.084	0.116
	Average	0.376	0.386	0.381	0.443	0.493	0.468	0.067	0.108	0.087
MUSIQ [9]	Cityscapes	0.773	0.788	0.780	0.744	0.733	0.739	-0.029	-0.055	-0.042
	Kitti	0.364	0.592	0.478	0.713	0.713	0.713	0.349	0.121	0.235
	Waymo	0.738	0.628	0.683	0.841	0.841	0.841	0.103	0.213	0.158
	Average	0.625	0.669	0.647	0.766	0.762	0.764	0.141	0.093	0.117
DBCNN [8]	Cityscapes	0.341	0.309	0.325	0.654	0.712	0.683	0.313	0.404	0.358
	Kitti	0.262	0.658	0.460	0.687	0.671	0.679	0.426	0.013	0.219
	Waymo	0.447	0.525	0.486	0.532	0.805	0.669	0.085	0.280	0.183
	Average	0.350	0.497	0.424	0.625	0.730	0.677	0.275	0.232	0.253
PaQ2PiQ [27]	Cityscapes	-0.029	-0.134	-0.081	0.158	-0.128	0.015	0.186	0.005	0.096
	Kitti	0.083	0.188	0.135	0.720	0.649	0.685	0.638	0.461	0.549
	Waymo	0.553	0.537	0.545	0.382	0.392	0.387	-0.171	-0.145	-0.158
	Average	0.202	0.197	0.200	0.420	0.304	0.362	0.218	0.107	0.163
HyperIQA [10]	Cityscapes	0.310	0.385	0.348	0.489	0.328	0.408	0.179	-0.057	0.061
	Kitti	-0.232	-0.028	-0.130	0.592	0.710	0.651	0.824	0.739	0.781
	Waymo	0.565	0.722	0.643	0.665	0.754	0.709	0.100	0.032	0.066
	Average	0.214	0.360	0.287	0.582	0.598	0.590	0.368	0.238	0.303
NIMA [61]	Cityscapes	0.465	0.520	0.492	0.325	0.260	0.292	-0.140	-0.260	-0.200
	Kitti	-0.266	-0.342	-0.304	0.675	0.621	0.648	0.941	0.964	0.953
	Waymo	0.344	0.217	0.280	0.229	0.265	0.247	-0.115	0.048	-0.033
	Average	0.181	0.132	0.156	0.410	0.382	0.396	0.229	0.250	0.240
Q-Instruct [1]	Cityscapes	0.095	0.204	0.150	0.283	0.404	0.343	0.188	0.200	0.194
	Kitti	0.409	0.056	0.232	0.396	0.874	0.635	-0.013	0.818	0.403
	Waymo	0.326	0.127	0.227	0.794	0.878	0.836	0.468	0.751	0.609
	Average	0.277	0.129	0.203	0.491	0.719	0.605	0.214	0.590	0.402
Q-Align [21]	Cityscapes	0.082	0.204	0.143	0.294	0.407	0.351	0.213	0.203	0.208
	Kitti	0.387	0.094	0.241	0.423	0.878	0.651	0.036	0.784	0.410
	Waymo	0.301	0.163	0.232	0.805	0.884	0.845	0.504	0.721	0.613
	Average	0.256	0.154	0.205	0.508	0.723	0.615	0.251	0.569	0.410
DeQA-Score [22]	Cityscapes	0.104	0.209	0.156	0.289	0.417	0.353	0.186	0.208	0.197
	Kitti	0.397	0.068	0.233	0.419	0.881	0.650	0.022	0.813	0.418
	Waymo	0.346	0.145	0.245	0.794	0.872	0.833	0.447	0.727	0.587
	Average	0.282	0.141	0.212	0.501	0.724	0.612	0.218	0.583	0.401
SageIQ-M	Cityscapes	0.355	0.455	0.405	0.519	0.734	0.626	0.164	0.278	0.221
	Kitti	0.308	0.045	0.177	0.731	0.959	0.845	0.422	0.914	0.668
	Waymo	0.285	-0.049	0.118	0.802	0.892	0.847	0.517	0.941	0.729
	Average	0.316	0.151	0.233	0.684	0.862	0.773	0.368	0.711	0.539

subsequently input into the triplet analyzer to obtain triplet weights for subject and object. The prompts are in Tab. II. It can be seen in Tab. VII that triplet analyzer responses accurately adhere to the requested format, so the information can be extracted accurately. Furthermore, the reasoning process is aligned with human common sense for autonomous driving, ensuring that the triplet analyzer decision is reliable.

Fig. 10 illustrates SageIQ-D visual examples. The first row displays the original images, whereas the second row presents the hybrid images constructed by our method. The third row highlights their TARs, where entities with higher weights are shown as they are dominant for scoring. For instance, given triplet weights of [8, 5] in Tab. VII, the softmax normalized weights are [0.952, 0.047]. The fourth to sixth rows depict the quality scores with and without SageIQ-P, denoted as "w/" and "w/o", respectively. These results showcase that, by incorporating SageIQ-P, the predicted scores are closer to labels scores. This is because model can accurately localize TARs and assess their quality, rather than relying on the entire image assessment. It demonstrates the effectiveness of SageIQ-P.

TABLE IX
SAGEIQ-M AND Q-INSTRUCT CONFIGURATION

Parameter	Value
do_sample	True
temperature	0.9
top_k	80
top_p	0.95
max_new_tokens	256

2) Experimental Results and Analysis: We evaluate the BIQA performance under two settings: without (w/o) and with (w/) the proposed pipeline SageIQ-P. Incorporating SageIQ-P means to employ the corresponding model as the scoring model. During the construction of the SageIQ-D, we integrated a diverse set of degradation types. For motion blur, the experimental results are summarized in Table VIII. Performance is assessed using SRCC, PLCC, and their mean value (MCC). Positive gains ($\Delta > 0$) achieved after applying SageIQ-P are highlighted in green for clarity. Fig. 11 illustrates the performance of MLLM-based IQA methods under a wider variety of degradation types, including Gaussian blur, mean blur, median blur, Gaussian noise, JPEG compression, and salt-and-pepper noise. Each bar represents the averaged MCC.

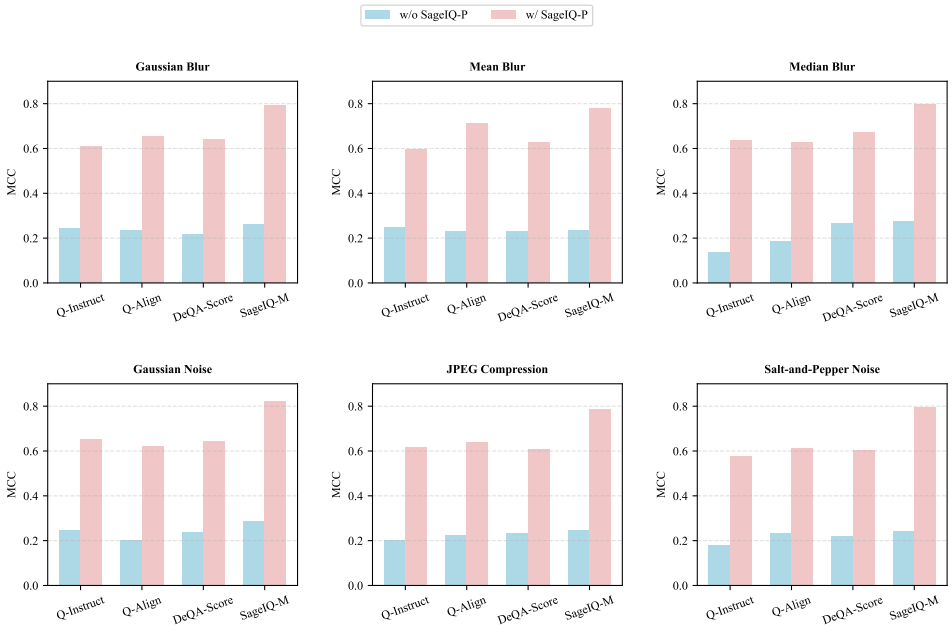

Fig. 11. Comparison of MLLM-based IQA methods under different image degradations. Each subfigure reports the averaged MCC, with blue and pink bars denoting results without and with SageIQ-P, respectively.

From the results presented in Tab. VIII, it can be observed that using existing BIQA methods directly yields unsatisfactory results. This is mainly because these methods perform quality assessment at a global level, where the influence of N-TARs often dominates the final score due to their large spatial area, thereby masking the true quality of TARs. In contrast, incorporating SageIQ-P effectively mitigates the influence of N-TARs, enabling more precise and region-aware quality assessment. This demonstrates the capability of SageIQ-P to focus on the perceptually relevant regions and deliver more reliable evaluations. Interestingly, in certain cases, SageIQ-P produces slightly lower results. We speculate that this might be caused by two main factors. (1) Scoring bias. Although the full image is degraded, its large size causes BIQA models to output scores close to the clean TARs, thereby yielding seemingly higher correlation coefficients. (2) Image homogeneity. Most BIQA models are trained on images with globally uniform quality distributions. However, the predicted TARs in our experiments may contain both clean and degraded areas. Under these non-homogeneous quality conditions, the models may produce unstable quality estimations.

We further evaluate SageIQ-P across diverse distortion types on MLLM-based IQA methods. As shown in Fig. 11, it consistently delivers substantial improvements in the average MCC, demonstrating enhanced robustness and perceptual consistency of MLLM-based IQA models under various degradation conditions. Moreover, across all evaluated scenarios, our proposed SageIQ-M demonstrates consistent and significant improvements. Such superiority can be attributed to its enhanced robustness against variations in image sizes, enabling it to effectively suppress distortions and maintain stable scoring performance. Overall, in the vast majority of cases, incorporating the proposed SageIQ-P pipeline leads to notable gains in SRCC, PLCC, and MCC values, verifying its effectiveness in accurately and robustly evaluating the quality of TARs.

3) SageIQ-P Ablation Study: In Fig. 12, we compare the performance of fine-tuning (FT) MLLMs, applying object detection (OD) to replace SGG, and our SageIQ-P, along with two SageIQ-P ablation variants, w/o predicate, w/o triplet and w/o weights. The experiments are conducted on MLLM-based IQA models using the SageIQ-D dataset with motion blur degradation. Fig. 12 reports the averaged MCC values across all test sets, where the baseline refers to directly feeding the full image into the model.

For the FT setting, the SageIQ-D is split into training and testing subsets with an 8:2 ratio. The input prompt is: "Rate the quality of this image, considering regions critical to autonomous driving, such as vehicles and pedestrians." Other training settings follow those in Section IV-A, including learning rate of 1e-5 and training epochs of 5. As shown in Fig. 12, FT improves model performance effectively. This indicates that the MLLM learns to follow the prompt to identify TARs and assess their quality. However, the improvement achieved through FT is limited, remaining lower than SageIQ-P. This can be attributed to several factors: (1) MLLMs-based IQA models are pre-trained on specific IQA datasets, which strengthens their IQA ability but introduces a mild form of catastrophic forgetting [62], [63], weakening other capabilities needed in this complex task, such as logical reasoning, spatial localization, and instruction following. (2) During training, the model must simultaneously learn to locate TARs of varying sizes and positions while assigning quality scores, but these irregular and varying patterns are inherently difficult for MLLMs to learn. (3) N-TARs are still included in the model input, introducing noise and interfering for the quality assessment. In contrast, the training-free SageIQ-P avoids additional training overhead and provides better interpretability in localizing TAR according to specific requirements, thereby achieving superior overall performance.

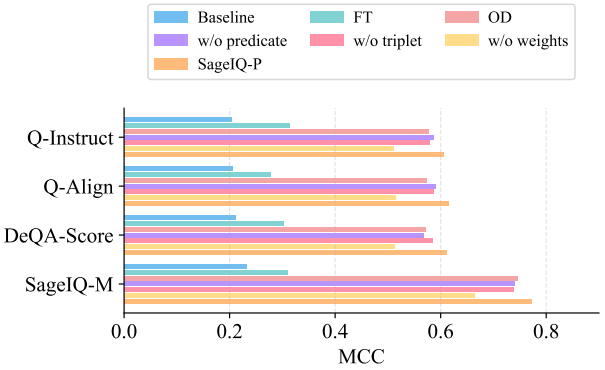

Fig. 12. Comparison of the performance of Baseline, FT, and SageIQ-P variants, evaluated using average MCC.

For OD, we use the widely adopted object detection model DETR [64] to replace the SGG module RelTR in the first stage of SageIQ-P. DETR directly outputs an entity list to replace the original triplet list. As shown in Fig. 12, OD exhibits a performance drop compared with SageIQ-P. This is mainly because the relational information is lost, and entities sharing the same name become indistinguishable in the list, making it difficult for the model to correctly identify and select appropriate entities. For the SageIQ-P ablation variants, the w/o predicate variant removes the predicate terms from each triplet, while the "w/o triplet" variant only retains the entities, which is roughly equivalent to OD. As illustrated in Fig. 12, both variants lead to a noticeable performance drop compared with the full SageIQ-P. For the "w/o predicate" variant, removing the predicate terms weakens the triplet analyzer, as it receives pairs of entities without meaningful semantic relationships. This lack of relational context introduces ambiguity during triplet selection and results in less accurate weight assignments. For the w/o triplet variant, its behavior and underlying cause are similar to those observed in OD. In contrast, SageIQ-P preserves the complete structural information, enabling the triplet analyzer to more accurately identify TARs for quality assessment. In Fig. 12, the "w/o weights" variant replaces the weighting mechanism for subject and object quality scores with a simple averaging strategy. This leads to a notable performance drop compared to the full SageIQ-P, because the score of the less relevant entity can be over-emphasized, weakening the final score's ability to reflect the quality of the critical entity. As a result, the final prediction is more easily biased by irrelevant regions, reducing its correlation with the ground truth.

4) Performance without Blur: As shown in Fig. 13, removing background blur yields a noticeable MCC improvement under both w/o and w/ SageIQ-P settings. However, it does not necessarily indicate more accurate prediction of TAR quality. Instead, removing blur makes the evaluation less discriminative, as the task becomes less sensitive to whether the model truly focuses on TARs. For w/o SageIQ-P, once blur is removed, the overall image quality becomes closer to the TAR quality, so even global-level assessment can better agree with the TAR-level ground truth. The improvement does not imply that the model can automatically focus on TARs. For w/ SageIQ-P, the improvement is largely driven by the near-

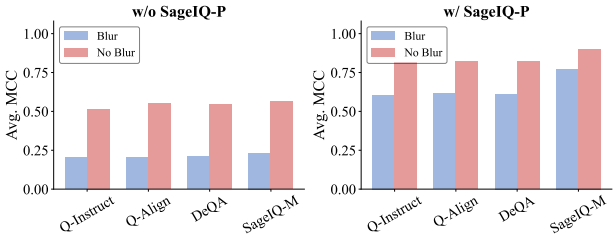

Fig. 13. Performance comparison with and without background blur.

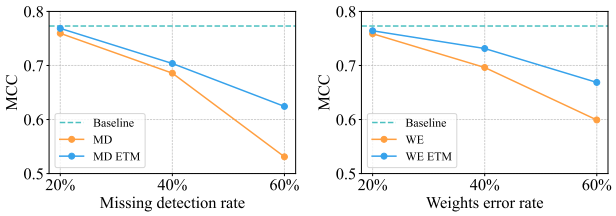

Fig. 14. Error robustness analysis of SageIQ-P under missing detections and weights error.

homogeneous quality distribution across regions. That is, even if the selected region is not the TAR, its quality can still be close to that of the TAR and lead to a high MCC.

5) Error Robustness Analysis: In this subsection, we analyze the effects of SGG missing detections (MD) and triplet weights error (WE) on the overall system performance, and propose the corresponding ensemble-based Error-Tolerant Mechanisms (ETM) for each case. The experiments are conducted using SageIQ-M model on the motion-blurred SageIQ-D dataset. The results are shown in Fig. 14, where the evaluation metric is the averaged MCC.

For MD, we randomly remove 20%, 40%, and 60% of the triplets from the triplet list. When no triplet results remain, the performance is evaluated on the entire image. As shown in the figure, the performance drop is minor at low missing rates since the removed triplets often do not correspond to TARs. At higher missing rates, more TARs are omitted, leading to a clear decline in performance. To address this issue, we propose an MD ETM that runs the SGG stage five times, ensembles the outputs into a union set, and removes duplicates to obtain a refined triplet list. This mechanism enables SageIQ-P to recover potential missing TARs without requiring ground-truth labels, improving performance particularly under high MD rates. For WE, Gaussian noise with zero mean and standard deviations of 20%, 40%, and 60% is added to the triplet weights. As shown in the figure, small noise levels have minimal impact because a subsequent softmax normalization allows higher weights to contribute more to the final score. When the noise becomes large enough to make the weights comparable, a noticeable performance degradation occurs. To mitigate this, we propose a WE ETM based on ensemble averaging. The weight scoring process is repeated five times, and the averaged result is adopted as the final score. This strategy effectively reduces randomness and stabilizes the identification of key entities. As illustrated in the figure, the WE ETM effectively enhances the overall system robustness.

6) Analysis of the Impact of TAR Quantity: In this section, we investigate the impact of the number of TARs on the overall performance. Four MLLM-based IQA models are evaluated

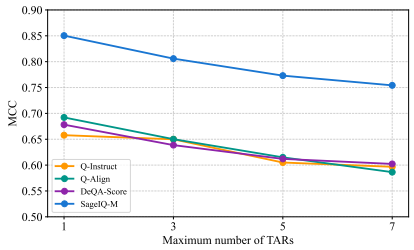

Fig. 15. Effect of the number of TARs on the overall performance, analyzed by applying 4 MLLM-based IQA models within SageIQ-P.

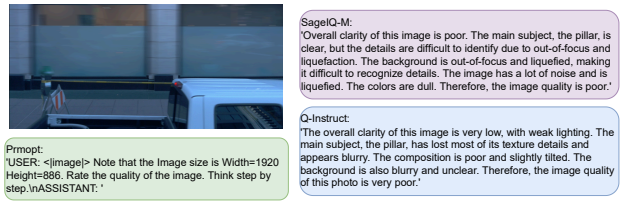

Fig. 16. An example showing the NLQA capability.

under SageIQ-P using SageIQ-D with motion blur, and the averaged MCC is the evaluation metric. The results are shown in Fig. 15. Recall that during the construction of the SageIQ-D, the TARs are derived from the labels of autonomous-driving datasets, in which each image is annotated with at least one label. This design ensures that every image in the SageIQ-D contains at least one TAR. In the previous experiments, the maximum number of TARs per image was set to 5 by default. Here, we vary this upper limit to 1, 3, 5, and 7 to explore the effect of TAR quantity. As shown in Fig. 15, model performance improves as the maximum limit decreases. This is probably because the localization task of SageIQ-P becomes inherently less complex with fewer TARs. Specifically, fewer localization targets lead to smaller cumulative errors, resulting in predictions from the BIQA model that more closely align with the ground-truth labels.

D. Natural Language Quality Assessment

Although incorporating multiple proposed modules to mitigate size-related issues, SageIQ-M successfully retains strong Natural Language Quality Assessment (NLQA) capabilities, a unique feature of MLLMs that distinguishes them from traditional BIQA models, as illustrated in Fig. 16. We verify SageIQ-M's NLQA performance by comparing its response similarity to Q-Instruct, a representative MLLM with exceptional NLQA capabilities. The model configurations, outlined in Tab.IX, offers a balance between generation randomness (*do_sample*=True, *temperature*=0.9, *top_p*=0.95) and determinism (*top_k*=80, *max_new_tokens*=256).

Tab. X shows the Latent Cosine Similarity (LCS) results. Particularly, responses are processed using the CLIP tokenizer and embedding module to extract latent features, followed by two reduction methods: (1) mean pooling across all tokens to obtain Mean token and (2) selecting the first class (CLS) token. The results indicate that both SageIQ-M and Q-Instruct assess quality in a semantically coherent manner. Mean token similarity is slightly lower and fluctuating, probably due to token-level deviations introduced by the generation randomness. To

TABLE X
COMPARISON OF LATENT COSINE SIMILARITY (LCS)

Metric	Cityscapes	KITTI	Waymo
LCS - Mean token w/o SageIQ-P	0.858	0.878	0.856
LCS - CLS token w/o SageIQ-P	1.000	1.000	1.000
LCS - Average w/o SageIQ-P	0.929	0.939	0.928
LCS - Mean token w/ SageIQ-P	0.822	0.860	0.841
LCS - CLS token w/ SageIQ-P	1.000	1.000	1.000
LCS - Average w/ SageIQ-P	0.911	0.930	0.921

TABLE XI
COMPARISON OF WORD2VEC SIMILARITY (WVS)

Metric	Cityscapes	KITTI	Waymo
WVS - Mean Pooling w/o SageIQ-P	0.910	0.927	0.912
WVS - Max Pooling w/o SageIQ-P	0.965	0.973	0.963
WVS - Average w/o SageIQ-P	0.938	0.950	0.938
WVS - Mean Pooling w/ SageIQ-P	0.896	0.910	0.889
WVS - Max Pooling w/ SageIQ-P	0.960	0.967	0.964
WVS - Average w/ SageIQ-P	0.928	0.939	0.927

further validate these findings, we employ another popular model Word2Vec [65] to tokenize and embed responses. Next, embeddings are aggregated using two strategies: mean and max pooling. Then, Word2Vec Similarity (WVS) can be computed, as summarized in Tab. XI. In this result, values exceed 0.9 in most cases, demonstrating that responses of two models are closely aligned. The above results highlight that our proposed SageIQ-M retains well NLQA capability. This can be attributed to the careful design of proposed modules, which effectively suppresses scoring bias while preserving the model's ability for general image feature extraction and logical reasoning.

V. LIMITATIONS AND FUTURE WORK

SageIQ demonstrates strong performance, but its complexity is relatively high. In particular, the RelTR module ($\approx$ 63.7 M), the LLaMA3-based triplet analyzer ($\approx$ 8.03 B), together with the mPLUG-Owl2-based SageIQ-M ($\approx$ 8.2 B), contribute to roughly 16 B parameters in total. This level of complexity, however, remains acceptable for three main reasons: (1) Our method operates in inference mode in practical deployment, eliminating the prohibitive computational burden of on-device training. (2) The parameter complexity is well within the capabilities of the latest automotive computing platforms [66], [67]. (3) Our model scale is consistent with, and often smaller than, existing influential large-model-based autonomous driving works [68]–[70]. Despite this acceptable complexity, further optimization remains an important direction in the future work. Possible explorations include: (1) Knowledge distillation to transfer reasoning ability from large models to smaller ones. (2) Quantization using mixed precision and low-bit methods to cut memory and computation costs. (3) Pruning to remove redundant parameters while preserving key capabilities. These optimizations aim to make SageIQ lighter and more practical for autonomous driving.

VI. CONCLUSION

This work overcomes the limitations of traditional BIQA methods by developing the SageIQ-P paradigm, which focuses on critical TARs using an SGG module, an LLM-based triplet analyzer, and a triplet scoring algorithm. To improve scoring accuracy, a MLLM based BIQA model SageIQ-M is proposed,

addressing scoring bias caused by image resizing. An automated annotation pipeline is also developed to tackle data scarcity. Experimental results show the method's capability to accurately assess image quality based on task-specific queries. While validated in autonomous driving, the approach is generalizable to various domains.

REFERENCES

[1] H. Wu, *et al.*, "Q-instruct: Improving low-level visual abilities for multi-modality foundation models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 25 490–25 500.

[2] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[3] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[4] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[5] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[6] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[7] S. Bosse, *et al.*, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[8] W. Zhang, *et al.*, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.

[9] J. Ke, *et al.*, "Musiq: Multi-scale image quality transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 5148–5157.

[10] S. Su, *et al.*, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2020.

[11] H. Wu, *et al.*, "FAST-VQA: Efficient end-to-end video quality assessment with fragment sampling," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 13666, 2022, pp. 538–554.

[12] W. Zhang, *et al.*, "Continual learning for blind image quality assessment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2864–2878, Mar. 2023.

[13] H. Wu, *et al.*, "Neighbourhood representative sampling for efficient end-to-end video quality assessment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15 185–15 202, Dec. 2023.

[14] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Task-specific normalization for continual learning of blind image quality models," *IEEE Trans. Image Process.*, vol. 33, pp. 1898–1910, 2024.

[15] K. Xu, *et al.*, "Boosting image quality assessment through efficient transformer adaptation with local feature enhancement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2024, pp. 2662–2672.

[16] H. Wu, *et al.*, "Discovqa: Temporal distortion-content transformers for video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4840–4854, Apr. 2023.

[17] ——, "Q-bench: A benchmark for general-purpose foundation models on low-level vision," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[18] OpenAI, "GPT-4 technical report," *arXiv:2303.08774*, Mar. 2024, available at https://arxiv.org/abs/2303.08774.

[19] W. Zhang, *et al.*, "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2023, pp. 14 071–14 081.

[20] Z. Zhang, *et al.*, "Q-bench-video: Benchmark the video quality understanding of lmms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 3229–3239.

[21] H. Wu, *et al.*, "Q-align: Teaching lmms for visual scoring via discrete text-defined levels," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2024, pp. 54 015–54 029.

[22] Z. You, *et al.*, "Teaching large language models to regress accurate image quality scores using score distribution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 14 483–14 494.

[23] W. Li, *et al.*, "Q-insight: Understanding image quality via visual reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2025.

[24] T. Wu, *et al.*, "VisualQuality-R1: Reasoning-induced image quality assessment via reinforcement learning to rank," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2025.

[25] A. Dosovitskiy, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[26] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2011, pp. 305–312.

[27] Z. Ying, *et al.*, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2020, pp. 3572–3582.

[28] H. Hu, *et al.*, "Spatio-temporal feature integration for quality assessment of stitched omnidirectional images," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 8, no. 2, pp. 1484–1499, 2024.

[29] H. Wang, *et al.*, "Ppd-iqa: A bottom-up and top-down combined approach for blind image quality assessment via prototype-prompted disentangling," *IEEE Trans. Emerg. Top. Comput. Intell.*, pp. 1–14, 2025.

[30] S. Yang, *et al.*, "Maniqa: Multi-dimension attention network for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1191–1200.

[31] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9905, 2016, pp. 852–869.

[32] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, July 2017, pp. 3097–3106.

[33] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2018, pp. 5831–5840.

[34] H. Liu, N. Yan, M. Mortazavi, and B. Bhanu, "Fully convolutional scene graph generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2021, pp. 11 541–11 551.

[35] Y. Teng and L. Wang, "Structured sparse r-CNN for direct scene graph generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2022, pp. 19 415–19 424.

[36] Y. Cong, M. Y. Yang, and B. Rosenhahn, "RelTR: Relation transformer for scene graph generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 09, pp. 11 169–11 183, Sept. 2023.

[37] R. Li, S. Zhang, and X. He, "SGTR+: End-to-end scene graph generation with transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2191–2205, Apr. 2024.

[38] G. S. Black, B. P. Rimal, and V. M. Vaidyan, "Balancing security and correctness in code generation: An empirical study on commercial large language models," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 9, no. 1, pp. 419–430, 2025.

[39] W. X. Zhao, *et al.*, "A survey of large language models," *arXiv:2303.18223*, Nov. 2023.

[40] C. Wang, *et al.*, "Enhancing result interpretability of neural architecture search-assisted medical ai via large language model," *IEEE Trans. Emerg. Top. Comput. Intell.*, pp. 1–11, 2025.

[41] F. Bu, Z. Wang, S. Wang, and Z. Liu, "An investigation into value misalignment in llm-generated texts for cultural heritage," *IEEE Trans. Emerg. Top. Comput. Intell.*, pp. 1–15, 2025.

[42] H. Touvron, *et al.*, "LLaMA: Open and efficient foundation language models," *arXiv:2303.13971*, Feb. 2023.

[43] A. Kirillov, *et al.*, "Segment anything," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3992–4003.

[44] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

[45] S. Yin, *et al.*, "A survey on multimodal large language models," *arXiv:2306.13549*, Apr. 2024.

[46] A. Radford, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Feb. 2021, pp. 8748–8763.

[47] Q. Ye, *et al.*, "mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 13 040–13 051.

[48] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2012, pp. 3354–3361.

[49] M. Cordts, *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2016, pp. 3213–3223.

[50] P. Sun, *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2020, pp. 2443–2451.

[51] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2015, pp. 1–15.

[53] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, July 1948.

[54] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 3209–3218.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[56] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: Applications to image and text data," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Aug. 2001, pp. 245–250.

[57] W. B. Johnson and J. Lindenstrauss, "Extensions of lipschitz mappings into a hilbert space," *Contemp. Math.*, vol. 26, pp. 189–206, 1984.

[58] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951.

[59] L. Agnolucci, L. Galteri, M. Bertini, and A. Del Bimbo, "Arniqa: Learning distortion manifold for image quality assessment," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2024, pp. 189–198.

[60] C. Chen, *et al.*, "TOPIQ: A top-down approach from semantics to distortions for image quality assessment," *IEEE Trans. Image Process.*, vol. 33, pp. 2404–2418, 2024.

[61] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[62] Y. Zhai, *et al.*, "Investigating the catastrophic forgetting in multimodal large language models," *arXiv preprint arXiv:2309.10313*, 2023.

[63] D. Zhu, *et al.*, "Model tailor: Mitigating catastrophic forgetting in multimodal large language models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024, pp. 62 581–62 598.

[64] N. Carion, *et al.*, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2020, pp. 213–229.

[65] T. Mikolov, *et al.*, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 26, 2013.

[66] XPENG Motors. (2025, Nov.) Xpeng shares achievements in physical ai emergence: Unveils xpeng vla 2.0, robotaxi, next-gen iron, and flying car. XPENG. [Online; accessed Nov. 25, 2025]. [Online]. Available: https://www.xpeng.com/news/019a56f54fe99a2a0a8d8a0282e402b7

[67] NVIDIA Corporation. (2025) In-vehicle computing for autonomous vehicles. NVIDIA. [Online; accessed Nov. 25, 2025]. [Online]. Available: https://www.nvidia.com/en-us/solutions/autonomous-vehicles/in-vehicle-computing/

[68] J. Mao, *et al.*, "Gpt-driver: Learning to drive with gpt," in *Proc. 37th Annu. Conf. Neural Inf. Process. Syst. (NeurIPS) Workshops*, 2023.

[69] Z. Xu, *et al.*, "Drivegpt4: Interpretable end-to-end autonomous driving via large language model," *IEEE Robot. Autom. Lett.*, 2024.

[70] J. Lubberstedt *et al.*, "V3lma: Visual 3d-enhanced language model for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, 2025, pp. 4769–4778.

Renwei Yang received the B.E. degree and the M.E. degree from University of Electronic Science and Technology of China (UESTC), in 2020 and 2023. He is currently pursuing the Ph.D. degree with the Department of Computer Science, City University of Hong Kong. His research interests include multimodal large language model, image quality assessment and enhancement, and video coding.

Zhengjie Yang received the Bachelor of Information Technology degree from Queensland University of Technology and Jinling Institute of Technology in 2015, the Master of Information Technology degree from The University of Sydney in 2017, and the Ph.D. degree from the School of Computer Science, The University of Sydney, in 2023. From 2017 to 2018, he worked as a software engineer at the Garvan Institute of Medical Research. From 2020 to 2024, he worked as a technical lead at Link Group Pty Ltd in Sydney. In 2024, he worked as a postdoctoral researcher at City University of Hong Kong. He is currently a postdoctoral fellow at the Hong Kong Generative AI Research and Development Center, The Hong Kong University of Science and Technology. His research interests include federated learning, edge computing, and distributed machine learning.

Yun Wang received the B.E. degree from China University of Geosciences (CUG) in 2020 and the M.E. degree from Sun Yat-sen University (SYSU), China, in 2023. He is currently pursuing the Ph.D. degree with the Department of Computer Science, City University of Hong Kong, Hong Kong SAR. His current research interests include 3D perception and multimodal learning.

Dapeng Oliver Wu (S'98–M'04–SM'06–F'13) received a Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, in 2003.

He is Yeung Kin Man Chair Professor of Network Science, and Chair Professor of Data Engineering at the Department of Computer Science, City University of Hong Kong. Previously, he was on the faculty of University of Florida, Gainesville, FL, USA and was the director of NSF Center for Big Learning, USA. His research interests are in the areas of networking, communications, signal processing, computer vision, machine learning, and information and network security. He received University of Florida Term Professorship Award in 2017, University of Florida Research Foundation Professorship Award in 2009, AFOSR Young Investigator Program (YIP) Award in 2009, ONR Young Investigator Program (YIP) Award in 2008, NSF CAREER award in 2007, the IEEE Circuits and Systems for Video Technology (CSVT) Transactions Best Paper Award for Year 2001, and the Best Paper Awards in IEEE GLOBECOM 2011 and International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QShine) 2006. // He has served as Editor in Chief of IEEE Transactions on Network Science and Engineering, Editor-at-Large for IEEE Open Journal of the Communications Society, founding Editor-in-Chief of Journal of Advances in Multimedia, and Associate Editor for IEEE Transactions on Cloud Computing, IEEE Transactions on Communications, IEEE Transactions on Signal and Information Processing over Networks, IEEE Signal Processing Magazine, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Wireless Communications and IEEE Transactions on Vehicular Technology. He has served as Technical Program Committee (TPC) Chair for IEEE INFOCOM 2012, and TPC chair for IEEE International Conference on Communications (ICC 2008), Signal Processing for Communications Symposium, and as a member of executive committee and/or technical program committee of over 100 conferences.

Shiqi Wang (Senior Member, IEEE) is currently an Associate Professor with the Department of Computer Science, City University of Hong Kong. He has proposed more than 50 technical proposals to ISO/MPEG, ITU-T, and AVS standards, and has authored or coauthored more than 300 refereed journal articles and conference papers. His research interests include video compression, image and video quality assessment, and image and video search and analysis. He received the Best Paper Award from IEEE VCIP 2019, ICME 2019, IEEE Multimedia 2018, and PCM 2017. His coauthored article received the Best Student Paper Award at IEEE ICIP 2018. He was also a recipient of the IEEE Multimedia Rising Star Award at ICME 2021.