

Understanding Pure Textual Reasoning for Blind Image Quality Assessment

Yuan Li and Shin'ya Nishida

Kyoto University

li.yuan.67n@st.kyoto-u.ac.jp

Abstract—Textual reasoning has recently been widely adopted in Blind Image Quality Assessment (BIQA). However, it remains unclear how textual information contributes to quality prediction and to what extent text can represent the score-related image contents. This work addresses these questions from an information-flow perspective by comparing existing BIQA models with three paradigms designed to learn the image–text–score relationship: Chain-of-Thought, Self-Consistency, and Autoencoder. Our experiments show that the score prediction performance of the existing model significantly drops when only textual information is used for prediction. Whereas the Chain-of-Thought paradigm introduces little improvement in BIQA performance, the Self-Consistency paradigm significantly reduces the gap between image- and text-conditioned predictions, narrowing the PLC-C/SRCC difference to 0.02/0.03. The Autoencoder-like paradigm is less effective in closing the image–text gap, yet it reveals a direction for further optimization. These findings provide insights into how to improve the textual reasoning for BIQA and high-level vision tasks.

Index Terms—Blind Image Quality Assessment, Self-Supervised Learning, Multimodal Model, Interpretable System

I. INTRODUCTION

Early research [1]–[6] in Blind Image Quality Assessment (BIQA) focused mainly on score prediction, extracting visual features and mapping them to quality scores through classification or regression. Although these models achieved reasonable accuracy, their limited ability to capture higher-level cues (e.g., semantics) restricted their interpretability and generalization. With the rise of multimodal large language models (MLLMs) [7]–[10], recent approaches [11]–[16] have begun to incorporate textual representations into BIQA. Works such as Q-Instruct [13], DepictQA [12], and Q-Ground [14] constructed extensive text-annotated datasets, laying a foundation for multimodal BIQA. Q-insight [15] and Q-Ponder [16] take a different direction by avoiding costly human annotations and instead leveraging pretrained knowledge, using reinforcement learning to optimize solely for the final quality score.

However, previous work on MLLM models has not clearly established the role of textual captions in BIQA apart from providing explanations. Briqa [17] has shown that, under supervised fine-tuning, models may bypass the intermediate text altogether when predicting scores. Q-Align [18] even removes the textual reasoning step entirely and directly fits MLLMs for score regression, achieving state-of-the-art (sota) performance. These observations raise an important question: how much do the generated text captions truly contribute to

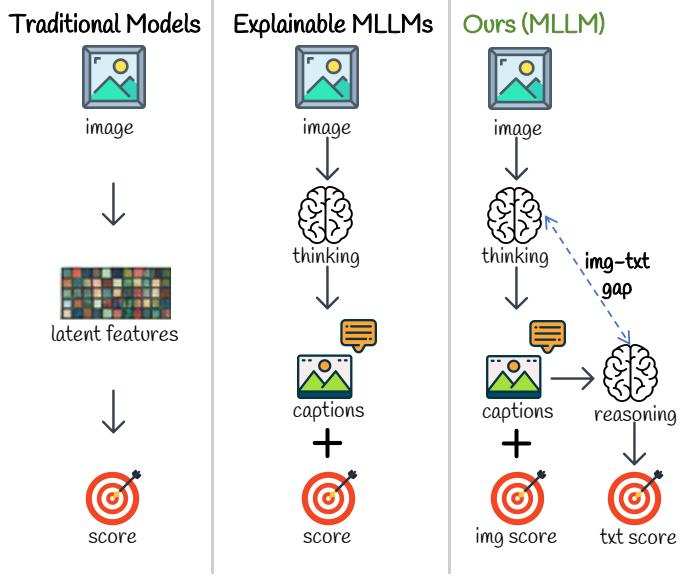


Fig. 1: Image-Text Gap. Traditional BIQA relies on visual features and lacks interpretability. Although MLLM-based BIQA generates both captions and scores, their relationship remains unclear. We examine the performance gap between image-based and text-based score prediction to reveal how textual reasoning contributes to interpretable BIQA.

quality prediction, and to what extent does the model actually engage in textual reasoning rather than merely producing superficial explanations?

Motivated by these questions, we adopt an information-flow perspective to examine how effectively text alone can convey quality-related information and how different learning paradigms influence the image–text gap. To this end, we systematically study three training paradigms: (1) a **Chain-of-Thought** paradigm, (2) a **Self-Consistency** paradigm, and (3) an **Autoencoder-like** paradigm.

We analyze the differences among the three paradigms from three perspectives. First, in terms of score prediction performance, the CoT paradigm provides almost no benefit, whereas both the Self-Consistency and Autoencoder-like paradigms reduce the image–text performance gap, with Self-Consistency showing the most notable improvement. Second, we examine token-level attention patterns during reasoning. The CoT paradigm behaves similarly to the baseline model, focusing on terms such as “focus” and “clear.” In contrast, the Self-

Consistency paradigm shifts attention toward score-related words like “good” and “moderate,” while the Autoencoder-like paradigm highlights cues like “blurry” and “focus.” Finally, to assess whether the score-related words introduce shortcuts, we remove such score-related terms. The resulting performance drop is negligible, indicating that the models rely on additional implicit quality cues and possess stronger reasoning ability. These findings clarify how training paradigms shape internal reasoning and provide insights for developing more reliable BIQA systems.

Our contributions are two-fold:

- From an information-flow perspective, we systematically evaluate three training paradigms for learning textual representations of image quality. Our analysis provides a structured baseline for studying textual reasoning in BIQA and clarifies how different paradigms shape text-conditioned performance.
- Our framework is general and can be applied to other downstream tasks that lack intermediate reasoning annotations. It offers a mechanism to induce task-specific and interpretable textual explanations, enabling broader use in multimodal and vision-centric applications.

II. RELATED WORKS

A. MLLM-based BIQA Systems

With the rapid progress of MLLMs, which demonstrate strong capabilities in textual description and reasoning, recent researches [12]–[16] have begun to explore training MLLMs for BIQA through supervised fine-tuning (SFT) or reinforcement learning (RL). These approaches aim to construct a more interpretable assessment system. Representative works such as DepictQA [12], Q-Instruct [13], Q-Insight [15], and Q-Ponder [16] are all devoted to building text-based interpretable systems for BIQA. Despite the presence of textual reasoning processes, these methods generally lack explicit evaluation of the quality or validity of the generated explanations. HumanIqa [19] supervises the reasoning and compares prediction performance under image- and text-conditioned settings. However, it relies on additional human-annotated reasoning data and provides limited analysis of how text-conditioned learning itself emerges or contributes to BIQA performance.

Motivated by these observations, we aim to investigate how textual tokens contribute to BIQA and how the gap between image- and text-conditioned performance can be effectively bridged, thus providing deeper insight into the role of language in MLLM-based BIQA systems.

B. Interpretable Visual Reasoning

Visual reasoning is a fundamental task of visual question answering (VQA) and MLLMs. Its primary goal is to infer answers by analyzing visual content and following a structured reasoning process. In many standard visual reasoning tasks, models benefit from abundant supervision such as question-answer pairs, attribute annotations, or annotated explanations that explicitly guide the intermediate steps of reasoning. However, for more complex perceptual tasks such

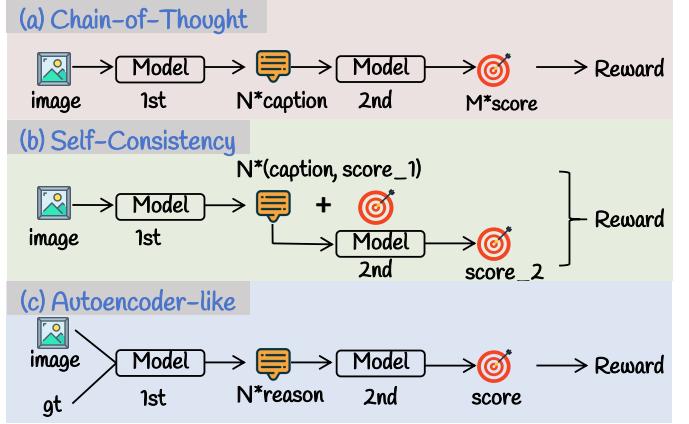


Fig. 2: **Training Paradigms.** All models share the same MLLM backbone and perform two forward passes, where “1st” and “2nd” denote the first and second forward inferences of the same model. The 1st pass is always conditioned on the image. (a) **Chain-of-Thought:** The model first generates N caption candidates from the image; each caption then produces M score predictions through an independent second-stage inference. (b) **Self-Consistency:** The first pass outputs N (caption, score) pairs, and each caption undergoes an additional inference step for score regression, receiving a self-consistency reward. (c) **Autoencoder-like:** The model takes the image and the ground-truth MOS during the first pass to generate reasoning text; the second pass regresses the score solely from this reasoning text.

as BIQA, effective supervision for intermediate reasoning is largely unavailable. To address the broader challenge of missing reasoning supervision, recent general-purpose reasoning systems have explored self-improving or self-rewarding strategies. MM-CoT [20] demonstrates how modality-specific CoT learning in both image and text domains enhances the reasoning capabilities of MLLMs, while Vision-SR1 [21] shows that such models can further improve by generating and evaluating their own reasoning processes without relying on external annotations.

Inspired by these, we propose a related self-consistency strategy tailored for BIQA. Our method leverages the model’s pre-trained image captioning capability as a form of self-supervised signal, encouraging the model to refine its internal reasoning pathway and learn a more coherent quality projection, even in the absence of explicit CoT labels.

III. METHODS

A. Overview of Information-Flows

In this work, we investigate the information flow among image, text, and score in BIQA. We first consider a sequential formulation in which information flows from image to reasoning and then to score, denoted as $I \rightarrow R \rightarrow S$, where each stage is treated independently in a Markov-style manner. We refer to this paradigm as **Chain-of-Thought (CoT)**, assuming that the generated text R can fully represent image quality information and directly support score prediction.

We also introduce a **Self-Consistency** paradigm that relaxes the strict separation between modalities. In this setting, the model performs score prediction in two stages: the first pass follows $\mathbf{I} \rightarrow (\mathbf{R}, \hat{\mathbf{S}})$ with visual information preserved, while the second pass relies solely on textual reasoning, following $\mathbf{R} \rightarrow \hat{\mathbf{S}}$. This design encourages the model to acquire text-based reasoning ability while maintaining consistency with image-conditioned predictions.

Finally, we move beyond the forward formulation by reversing the information flow. In this **Autoencoder-like** paradigm, the model is explicitly provided with the ground-truth score and trained to generate explanations that justify it, following $(\mathbf{I}, S^*) \rightarrow \mathbf{R} \rightarrow \hat{\mathbf{S}}$. This formulation explores how quality-aware supervision shapes textual explanations.

B. Chain-of-Thought Reasoning Learning

As illustrated in Fig. 2 (a), the model generates N reasoning traces. The i -th reasoning trace produces M score predictions, denoted as $s_{i,j}$ for $i = 1, \dots, N$ and $j = 1, \dots, M$. As in (1), each score prediction $s_{i,j}$ receives a reward $r_{i,j}$, where x is the absolute difference between the predicted score and ground-truth MOS, and t controls the tolerance margin. The reward for the i -th reasoning trace is then obtained by averaging the rewards of its M predictions as in (2). Higher values of R_i increase the generation probability of the corresponding sentence by strengthening the loss term defined in (3). In (3), \mathcal{L} denotes the cross-entropy loss, I , R , and S denote the image, reasoning (text), and quality score, respectively, while α and β are hyper-parameters that enable or disable the loss terms associated with different training stages.

$$r_{i,j} = \begin{cases} 0.5(1 + \cos(\pi x/t)), & \text{if } x < t, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

$$R_i = \frac{1}{M} \sum_{j=1}^M r_{i,j}, \quad (2)$$

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}(I, R) + \beta \mathcal{L}(R, S^*). \quad (3)$$

C. Self-Consistency Learning

In the first stage, the model generates a caption sequence and a score prediction directly from the image. In the second stage, the model performs another round of inference using only the generated reasoning sequence. Both predictions are supervised using the same score reward as in (1), encouraging the model to produce reasoning that is not only consistent with the visual input but also predictive when used independently as in Fig. 2 (b). The loss function is defined as:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}(I, S^*) + \beta \mathcal{L}(R, S^*), \quad (4)$$

This formulation allows the model to retain rich visual cues during training while progressively aligning its internal reasoning with textual explanations, ultimately improving its text-only reasoning capability.

D. Autoencoder-like Learning

During training stage one, the model is given the ground-truth quality score S^* and generates a textual explanation R conditioned on both the image and the score: $(I, S^*) \rightarrow R$. In the training stage two, the model is required to perform prediction using only the generated reasoning: $R \rightarrow \hat{S}$. The training stage two evaluates whether the explanation itself is predictive of image quality, functioning analogously to a decoding step in an autoencoder. In the testing stage, the score is masked with placeholders (e.g., “some score”). The loss from the two stages is estimated as:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}(S^*, R) + \beta \mathcal{L}(R, S^*). \quad (5)$$

Compared with the self-consistency paradigm, this Autoencoder-like framework explicitly places the score at the input side of reasoning generation and prohibits visual access during score regression. This encourages the model to encode score-relevant semantics into the reasoning itself, reinforcing the quality-predictive capacity of textual explanations.

E. Training via Group-Relative Policy Optimization Strategy

To reduce the need for human annotations and enable a more scalable training paradigm, we adopt a self-supervised reinforcement learning approach built upon the Group-Relative Policy Optimization (GRPO) [22] framework. During each training iteration, the model generates multiple candidate reasoning chains and corresponding answers. These candidates are then evaluated using a set of designed reward functions, which guide the optimization direction and determine how the model evolves over time. The training objective of a single GRPO process is mathematically expressed in (6).

$$\mathcal{J}_{\text{GRPO}} = \mathbb{E}\left[\frac{1}{N} \sum_i^N \min(d_i A_i, C_{d_i, \epsilon} A_i - \beta \cdot \text{KL})\right], \quad (6)$$

where $d_i = \frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}$, $A_i = \frac{r_i - \text{mean}(r_1, r_2, \dots, r_N)}{\text{std}(r_1, r_2, \dots, r_N)}$, $C_{d_i, \epsilon} = \text{clip}(d_i, 1 - \epsilon, 1 + \epsilon)$, and $\text{KL} = \mathbb{D}_{\text{KL}}(\pi_\theta || \pi_{\text{ref}})$. Note that π_θ , π_{old} and π_{ref} denote the policy model, old policy model and reference model, respectively. r_i denotes rewards, and ϵ and β denote hyper-parameters. The model selectively reweights different sampling losses based on the group reward, thereby reinforcing the more favorable reasoning trajectories.

IV. EXPERIMENTS

A. Datasets and Training Details

Training Dataset. To ensure a fair comparison with prior works, we adopt the default training split of the KonIQ [23] dataset at a resolution of 512×384 . For score supervision, we use the DeQA [24] normalized quality labels following recent MLLM-based BIQA methods. **Test Datasets.** To comprehensively evaluate generalization robustness, we test our models on six datasets, including SPAQ [25], LIVE-W [26], KADID [27], AGIQA [28], CSIQ [29] and the test split of KonIQ [29] dataset. This mixture of real and synthetic benchmarks enables a thorough assessment of the proposed visual-to-text learning paradigms. **Training Details.** We use *Qwen-VL-2.5-7B-Instruct* [10] as our backbone model. In the

TABLE I PLCC / SRCC performance comparisons. All models are assumed to be trained on the KonIQ [23] training set. The best and second-best results are highlighted in red and underlined blue. Text-Only Conditions uses generated captions; the Score-related Words Removed setting evaluates captions with terms like “good,” “moderate,” “average,” “poor,” and “decent” removed. Image-conditioned results of other models are from reported versions.

Model	KonIQ [23]	SPAQ [25]	KADID [27]	LIVE-W [26]	AGIQA [28]	CSIQ [29]	AVG.
<i>Deep-Learning Models</i>							
NIMA [2] (2018)	0.896 / 0.859	0.838 / 0.856	0.532 / 0.535	0.814 / 0.771	0.715 / 0.654	0.695 / 0.649	0.748 / 0.721
DBCNN [5] (2019)	0.884 / 0.875	0.812 / 0.806	0.497 / 0.484	0.773 / 0.730	0.641 / 0.648	0.586 / 0.572	0.714 / 0.689
MUSIQ [3] (2021)	0.924 / 0.929	0.868 / 0.863	0.575 / 0.556	0.789 / 0.830	0.722 / 0.630	0.771 / 0.710	0.775 / 0.753
MANIQA [6] (2022)	0.849 / 0.834	0.768 / 0.758	0.499 / 0.465	0.849 / 0.832	0.723 / 0.636	0.623 / 0.627	0.719 / 0.692
CLIP-IQA+ [31] (2023)	0.909 / 0.895	0.866 / 0.864	0.653 / 0.654	0.832 / 0.805	0.736 / 0.685	0.772 / 0.719	0.795 / 0.770
<i>SFT-based and RL-based MLLMs</i>							
C2Score [32] (2024)	0.923 / 0.910	0.867 / 0.860	0.500 / 0.453	0.786 / 0.772	0.777 / 0.671	<u>0.735 / 0.705</u>	0.765 / 0.729
Q-Align [18] (2024)	<u>0.941 / 0.940</u>	0.886 / 0.887	0.674 / 0.684	0.853 / <u>0.860</u>	0.772 / 0.735	0.671 / 0.737	0.799 / <u>0.807</u>
DeQA [24] (2025)	<u>0.953 / 0.941</u>	<u>0.895 / 0.896</u>	<u>0.694 / 0.687</u>	<u>0.892 / 0.879</u>	<u>0.809 / 0.729</u>	<u>0.787 / 0.744</u>	<u>0.838 / 0.813</u>
Q-Insight-Score [15] (2025)	0.918 / 0.895	<u>0.903 / 0.899</u>	<u>0.702 / 0.702</u>	0.870 / 0.839	<u>0.816 / 0.766</u>	0.685 / 0.640	<u>0.813 / 0.789</u>
Ours (Chain-of-Thought)	0.920 / 0.907	0.886 / 0.884	0.629 / 0.699	<u>0.878 / 0.851</u>	0.806 / <u>0.767</u>	0.687 / 0.634	0.801 / 0.790
Ours (Self-Consistency)	0.917 / 0.905	0.883 / 0.882	0.632 / 0.692	0.874 / 0.843	0.805 / <u>0.766</u>	0.704 / 0.647	0.803 / 0.789
Ours (Autoencoder-like)	0.926 / 0.912	0.884 / 0.882	0.649 / <u>0.700</u>	0.873 / 0.850	0.810 / 0.763	0.683 / 0.636	0.804 / 0.791
<i>Text-Only Conditions</i>							
Q-Insight-Score [15] (2025)	0.859 / 0.827	0.832 / 0.833	0.604 / 0.620	0.778 / 0.776	0.766 / 0.690	0.582 / 0.535	0.737 / 0.713
Ours (Chain-of-Thought)	0.851 / 0.819	0.829 / 0.833	0.604 / 0.620	0.779 / 0.776	0.766 / 0.690	0.582 / 0.535	0.735 / 0.712
Ours (Self-Consistency)	<u>0.900 / 0.879</u>	<u>0.864 / 0.861</u>	<u>0.627 / 0.661</u>	<u>0.838 / 0.815</u>	<u>0.797 / 0.734</u>	<u>0.672 / 0.620</u>	<u>0.783 / 0.762</u>
Ours (Autoencoder-like)	0.877 / 0.861	0.824 / 0.839	0.632 / 0.645	0.761 / 0.767	0.774 / 0.696	0.585 / 0.557	0.742 / 0.725
<i>Text-Only Conditions (Score-related Words Removed)</i>							
Q-Insight-Score [15] (2025)	0.856 / 0.825	0.831 / 0.832	0.611 / 0.621	0.772 / 0.771	0.766 / 0.689	0.583 / 0.535	0.736 / 0.712
Ours (Chain-of-Thought)	0.851 / 0.818	0.831 / 0.831	0.609 / 0.621	0.772 / 0.771	0.766 / 0.689	0.581 / 0.534	0.735 / 0.711
Ours (Self-Consistency)	<u>0.898 / 0.879</u>	<u>0.861 / 0.859</u>	<u>0.632 / 0.660</u>	<u>0.829 / 0.812</u>	<u>0.794 / 0.727</u>	<u>0.665 / 0.621</u>	<u>0.780 / 0.760</u>
Ours (Autoencoder-like)	0.867 / 0.847	0.834 / 0.838	0.638 / 0.644	0.758 / 0.763	0.774 / 0.696	0.589 / 0.557	0.743 / 0.724

pretraining stage, the model is optimized using only discrete score supervision together with a format reward, following Q-Insight-Score [15]. We adopted the Adam optimizer [30], a batch size of 128, and trained for 10 epochs on eight NVIDIA A6000 GPUs, around 27 hours. For the main experiments, we fine-tuned each of the proposed frameworks on the same KonIQ [23] training split for 2 epochs, using the same configuration as in pretraining. Details are reported in Table II.

B. Quality Score Prediction Performance

As shown in Table I, our models achieve performance competitive with current sota BIQA approaches across multiple benchmarks under image conditions, with the average PLCC/SRCC gap limited to 0.03/0.02 over six datasets. Remarkably, in the text-only inference setting, our Self-Consistency model reaches performance comparable to deep learning-based BIQA frameworks, indicating that they have learned genuine reasoning patterns rather than relying solely on visual features. Furthermore, the gap between image- and text-conditioned predictions is reduced to 0.02/0.03 in terms of PLCC/SRCC, marking a substantial step toward self-consistent BIQA systems.

C. Ablation Studies

We conduct ablation experiments by varying the loss weights α and β to control the contribution of each inference stage, as summarized in Table II. We observe that the Chain-of-Thought does not gain any improvement. It may be related to that images typically contain richer and more fine-grained cues than text, and completely removing visual signals during score regression often makes it difficult to learn effective reasoning. Compared to others, the Self-Consistency model achieves the best text-only inference performance when $\alpha = 1$ and $\beta = 0$, whereas the Autoencoder-like model performs best under the combined image- and text-conditioned settings when $\alpha = 0$ and $\beta = 1$. This indicates that Self-Consistency learns the text-to-score mapping primarily when visual information is available during training, while the Autoencoder-like paradigm enhances both image-conditioned and text-conditioned performance by explicitly learning the reasoning-to-score relationship.

V. DISCUSSION

A. Image–Text Gap

To better understand the performance differences between image- and text-conditioned inferences, we further analyze where the score-related information originates in Fig. 3. Under

TABLE II **Ablation studies.** For each model, the first row shows image-conditioned results, and the second row shows text-conditioned results. Training and inference times are measured on the KonIQ [23] dataset. Baseline model is the reproduced version of Q-insight-Score [15].

Setting	KonIQ	SPAQ	KADID	LIVE-W	AGIQA	CSIQ	Avg.	Train (hrs/epoch)	Infer (s/img)
Baseline	0.920 / 0.907 0.859 / 0.827	0.885 / 0.884 0.832 / 0.833	0.629 / 0.698 0.604 / 0.620	0.879 / 0.851 0.778 / 0.776	0.807 / 0.765 0.766 / 0.690	0.687 / 0.634 0.582 / 0.535	0.801 / 0.790 0.737 / 0.714	≈2.7	5.95 / 3.60
Chain-of-Thought ($\alpha = 1, \beta = 1$)	0.920 / 0.907 0.851 / 0.819	0.886 / 0.884 0.829 / 0.833	0.629 / 0.699 0.604 / 0.620	0.878 / 0.851 0.779 / 0.776	0.806 / 0.767 0.766 / 0.690	0.687 / 0.634 0.582 / 0.535	0.801 / 0.790 0.735 / 0.712	≈5.0	6.40 / 2.40
Self-Consistency ($\alpha = 0, \beta = 1$)	0.922 / 0.906 0.849 / 0.810	0.886 / 0.883 0.800 / 0.823	0.642 / 0.707 0.610 / 0.642	0.880 / 0.852 0.762 / 0.783	0.809 / 0.767 0.769 / 0.700	0.700 / 0.642 0.568 / 0.559	0.807 / 0.793 0.726 / 0.720	≈2.6	6.07 / 3.37
Self-Consistency ($\alpha = 1, \beta = 0$)	0.917 / 0.905 0.900 / 0.879	0.883 / 0.882 0.864 / 0.861	0.632 / 0.692 0.627 / 0.661	0.874 / 0.843 0.838 / 0.815	0.805 / 0.766 0.797 / 0.734	0.704 / 0.647 0.672 / 0.620	0.803 / 0.789 0.783 / 0.762	≈2.5	5.76 / 3.20
Self-Consistency ($\alpha = 1, \beta = 1$)	0.919 / 0.907 0.881 / 0.848	0.883 / 0.883 0.854 / 0.853	0.631 / 0.700 0.621 / 0.653	0.879 / 0.849 0.812 / 0.79	0.804 / 0.766 0.784 / 0.704	0.695 / 0.631 0.634 / 0.576	0.802 / 0.789 0.764 / 0.738	≈3.1	5.66 / 3.14
Autoencoder-like ($\alpha = 0, \beta = 1$)	0.926 / 0.912 0.877 / 0.861	0.884 / 0.882 0.824 / 0.839	0.649 / 0.700 0.632 / 0.645	0.873 / 0.850 0.761 / 0.767	0.810 / 0.763 0.774 / 0.696	0.683 / 0.636 0.585 / 0.557	0.804 / 0.791 0.742 / 0.725	≈7.0	6.40 / 3.52
Autoencoder-like ($\alpha = 1, \beta = 0$)	0.919 / 0.907 0.868 / 0.838	0.885 / 0.883 0.838 / 0.840	0.618 / 0.695 0.588 / 0.630	0.879 / 0.849 0.803 / 0.777	0.804 / 0.766 0.764 / 0.680	0.691 / 0.631 0.616 / 0.548	0.799 / 0.787 0.746 / 0.719	≈3.5	5.86 / 3.42
Autoencoder-like ($\alpha = 1, \beta = 1$)	0.925 / 0.908 0.876 / 0.853	0.887 / 0.885 0.827 / 0.837	0.648 / 0.701 0.614 / 0.631	0.876 / 0.851 0.767 / 0.787	0.812 / 0.763 0.774 / 0.689	0.695 / 0.645 0.601 / 0.587	0.799 / 0.792 0.743 / 0.731	≈5.5	6.06 / 3.54

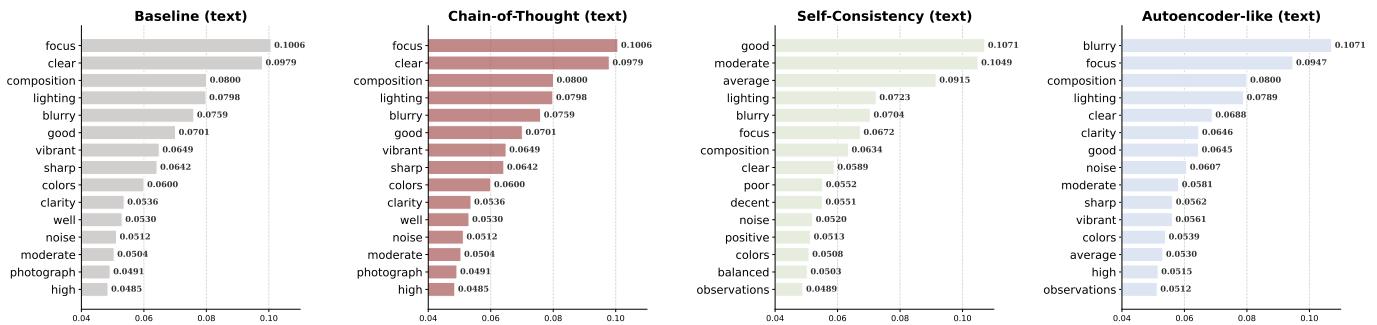


Fig. 3: **Comprehensive Analysis of Attention Distributions.** Please zoom in to check the details. We analyze model behavior on the KonIQ [23] test set (3,015 images) by examining softmax-normalized attention values to estimate token contributions to score prediction. This shows the text-conditioned results. Different learning paradigms lead to distinct shifts in the tokens emphasized by the model. Image-conditioned results are shown in Section 3.2 in supplemental materials.

the image-conditioned setting (details in supplemental materials), although the softmax-normalized attention weights of non-image tokens are around 0.06, their contributions remain negligible because the unscaled attention values of image tokens are substantially larger than those of other tokens. As a result, score prediction is dominated almost entirely by image tokens. In contrast, under the text-conditioned setting, the absence of dominant image-token activations allows other tokens to contribute more effectively to the prediction.

Compared with the baseline, the Chain-of-Thought model shows a negligible change in token usage. The Self-Consistency model, however, focuses on score-related tokens such as “good,” “moderate,” and “average,” which potentially explains its strong text-conditioned performance. By contrast, the Autoencoder-like model focuses on more natural quality cues, including “blurry,” “focus,” and “composition.” This behavior enables it to improve both image-conditioned and text-conditioned performance. In the “Score-related Words Removed” setting of Table I, models remain capable of producing reasonable predictions even after removing terms such as “good,” “moderate,” “average,” “poor,” and “decent.” This indicates that the model’s reasoning ability has improved

and that it remains robust without relying on these superficial cues.

VI. CONCLUSION

In this work, we investigated the information flow among image, text, and score in Blind Image Quality Assessment. By systematically comparing three learning paradigms—Chain-of-Thought, Self-Consistency, and an Autoencoder paradigm—we analyzed how textual reasoning contributes to quality prediction and how the image-text performance gap can be reduced. Our results show that naive Chain-of-Thought reasoning has a limited impact, while Self-Consistency and Autoencoder-like paradigms improve text-conditioned BIQA through distinct mechanisms. In particular, Self-Consistency effectively narrows the image–text gap, whereas the Autoencoder-like paradigm promotes more natural, quality-related textual explanations. Through token-level analysis, we further revealed how different training strategies shape the model’s reasoning focus. Overall, this study provides insights into the role of textual reasoning in BIQA and offers a principled basis for developing more interpretable quality assessment systems. We hope these insights inspire future work on integrating perceptual cues with textual explanations.

REFERENCES

- [1] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, “Making a “completely blind” image quality analyzer,” *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012. 1
- [2] Hossein Talebi and Peyman Milanfar, “Nima: Neural image assessment,” *IEEE transactions on image processing*, vol. 27, no. 8, pp. 3998–4011, 2018. 1, 4
- [3] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang, “Musiq: Multi-scale image quality transformer,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5148–5157. 1, 4
- [4] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin, “Topiq: A top-down approach from semantics to distortions for image quality assessment,” *IEEE Transactions on Image Processing*, vol. 33, pp. 2404–2418, 2024. 1
- [5] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang, “Blind image quality assessment using a deep bilinear convolutional neural network,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020. 1, 4
- [6] Sidi Yang, Tianhe Wu, Shuaiwei Shi, Shanshan Lao, Yuan Gong, Ming-deng Cao, Jiahao Wang, and Yujiu Yang, “Maniq: Multi-dimension attention network for no-reference image quality assessment,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1191–1200. 1, 4
- [7] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al., “Flamingo: a visual language model for few-shot learning,” *Advances in neural information processing systems*, vol. 35, pp. 23716–23736, 2022. 1
- [8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee, “Visual instruction tuning,” *Advances in neural information processing systems*, vol. 36, pp. 34892–34916, 2023. 1
- [9] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang, “mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration,” in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2024, pp. 13040–13051. 1
- [10] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al., “Qwen2.5-v1 technical report,” *arXiv preprint arXiv:2502.13923*, 2025. 1, 3
- [11] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma, “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*, 2023, pp. 14071–14081. 1
- [12] Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong, “Depicting beyond scores: Advancing image quality assessment through multi-modal language models,” in *European Conference on Computer Vision*. Springer, 2024, pp. 259–276. 1, 2
- [13] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al., “Q-instruct: Improving low-level visual abilities for multimodality foundation models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 25490–25500. 1, 2
- [14] Chaofeng Chen, Sensen Yang, Haoning Wu, Liang Liao, Zicheng Zhang, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin, “Q-ground: Image quality grounding with large multi-modality models,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 486–495. 1, 2
- [15] Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang, “Q-insight: Understanding image quality via visual reinforcement learning,” *arXiv preprint arXiv:2503.22679*, 2025. 1, 2, 4, 5
- [16] Zhuoxuan Cai, Jian Zhang, Xinbin Yuan, Peng-Tao Jiang, Wenxiang Chen, Bowen Tang, Lujian Yao, Qiyuan Wang, Jinwen Chen, and Bo Li, “Q-ponder: A unified training pipeline for reasoning-based visual quality assessment,” *arXiv preprint arXiv:2506.05384*, 2025. 1, 2
- [17] Yuan Li, Zitang Sun, Yen-ju Chen, and Shin’ya Nishida, “Building reasonable inference for vision-language models in blind image quality assessment,” in *International Conference on Neural Information Processing*. Springer, 2025, pp. 283–295. 1
- [18] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al., “Q-align: Teaching llms for visual scoring via discrete text-defined levels,” *arXiv preprint arXiv:2312.17090*, 2023. 1, 4
- [19] Yuan Li, Yahan Yu, Youyuan Lin, Yong-Hao Yang, Chenhui Chu, and Shin’ya Nishida, “Guiding perception-reasoning closer to human in blind image quality assessment,” 2025. 2
- [20] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola, “Multimodal chain-of-thought reasoning in language models,” *arXiv preprint arXiv:2302.00923*, 2023. 2
- [21] Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, et al., “Self-rewarding vision-language model via reasoning decomposition,” *arXiv preprint arXiv:2508.19652*, 2025. 2
- [22] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al., “Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning,” *arXiv preprint arXiv:2501.12948*, 2025. 3
- [23] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe, “Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment,” *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020. 3, 4, 5
- [24] Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong, “Teaching large language models to regress accurate image quality scores using score distribution,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14483–14494. 3, 4
- [25] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang, “Perceptual quality assessment of smartphone photography,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3677–3686. 3, 4
- [26] Deepthi Ghadiyaram and Alan C Bovik, “Live in the wild image quality challenge database,” *Online: http://live. ece. utexas. edu/research/ChallengeDB/index. html [Mar, 2017]*, 2015. 3, 4
- [27] Hanhe Lin, Vlad Hosu, and Dietmar Saupe, “Kadid-10k: A large-scale artificially distorted iqas database,” in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019, pp. 1–3. 3, 4
- [28] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin, “Agica-3k: An open database for ai-generated image quality assessment,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 6833–6846, 2023. 3, 4
- [29] Eric C Larson and Damon M Chandler, “Most apparent distortion: full-reference image quality assessment and the role of strategy,” *Journal of electronic imaging*, vol. 19, no. 1, pp. 011006–011006, 2010. 3, 4
- [30] Ilya Loshchilov and Frank Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017. 4
- [31] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy, “Exploring clip for assessing the look and feel of images,” in *Proceedings of the AAAI conference on artificial intelligence*, 2023, vol. 37, pp. 2555–2563. 4
- [32] Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang, “Adaptive image quality assessment via teaching large multimodal model to compare,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 32611–32629, 2024. 4