Regression Over Classification: Assessing Image Aesthetics via Multimodal Large Language Models

Xingyuan Ma*, Shuai He*†, Anlong Ming†, Haobin Zhong, Huadong Ma

School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China
{maxy, hs19951021, mal, zhonghaobin2023, mhd}@bupt.edu.cn

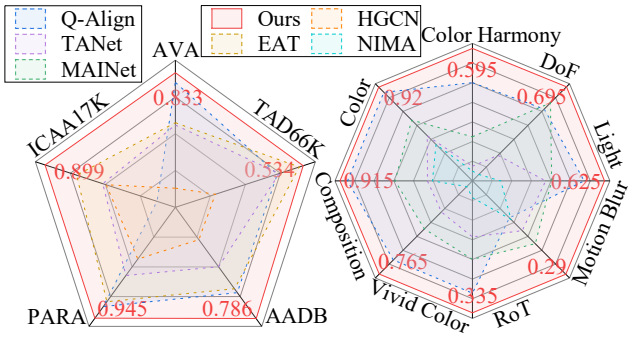
Abstract

Image Aesthetics Assessment (IAA) evaluates visual quality through user-centered perceptual analysis and can guide various applications. Recent advances in Multimodal Large Language Models (MLLMs) have sparked interest in adapting them for IAA. However, two critical limitations persist in applying MLLMs to IAA: 1) the tokenization strategy leads to insensitivity to scores, and 2) the classification-based decoding mechanisms introduce score quantization errors. Current MLLM-based IAA methods treat the task as coarse rating classification followed by probability-to-score mapping, which loses fine-grained information. To address these challenges, we propose ROC4MLLM, offering complementary solutions from two perspectives: ***1) Representation***: We separate scores from the word token space to **avoid tokenizing scores as text**. An independent position token bridges these spaces, improving the **sensitivity of the model to score positions** in text. ***2) Computation***: We apply distinct loss functions for text and score predictions to enhance the **sensitivity of the model to score gradients**. Decoupling scores from text ensures effective supervision while **preventing interference between scores and text** in the loss computation. Extensive experiments across five datasets demonstrate that ROC4MLLM achieves state-of-the-art performance without requiring additional training data. Additionally, its plug-and-play design ensures seamless integration with existing MLLMs, boosting their IAA performance.

Code —
https://github.com/woshidandan/Assessing-Image-Aesthetics-via-Multimodal-Large-Language-Models

Introduction

As digital photography expands rapidly, Image Aesthetics Assessment (IAA) has become one of the most important criteria to automatically assess whether the image meets users' aesthetic preferences. It is also an essential step in imaging measurements among manufacturers to evaluate the performance of smartphones and cameras. The



(a) Holistic IAA Tasks (b) Multi-Attribute IAA Tasks

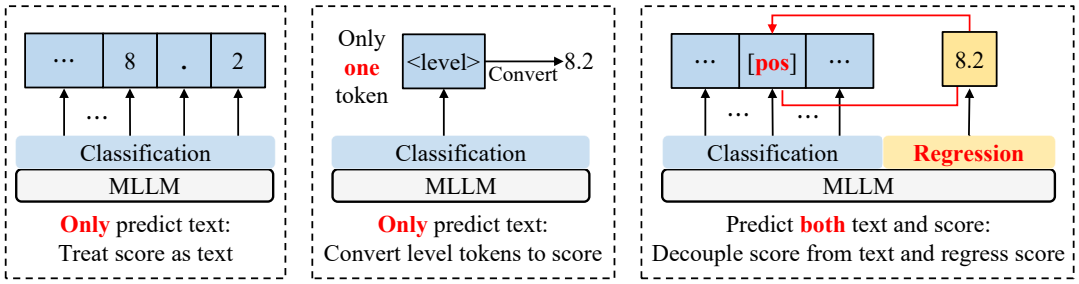
Figure 1: Comparison of ROC4MLLM and prevailing IAA methods via (SRCC + PLCC)/2. (a) Holistic IAA on five datasets (axes: dataset names). (b) Multi-attribute IAA on two datasets (axes: attributes).

complexity of IAA tasks stems from integrating qualitative and quantitative visual analysis, with the subjective nature of human preferences. Compared to conventional CNN-based or Transformer-based methods, Multimodal Large Language Models (MLLMs) excel at aligning high-level visual content with human preferences. As a result, applying MLLMs to specific IAA tasks has yielded promising results (Huang et al. 2024a; Wu et al. 2024c; Ke et al. 2023; Zhou et al. 2024). However, predicting human-aligned aesthetic scores—a core quantitative task in IAA—remains significant challenges for current MLLMs (Wu et al. 2024a). Specifically, there are three key challenges.

Insensitivity to Scores. In MLLM, the tokenizer converts words into token sequences for processing. In IAA, this process treats scores as text, splitting them into arbitrary tokens. This disrupts their numerical structure. As a result, MLLMs predict scores digit by digit and struggle to recognize numerical place value (e.g., thousands versus hundreds), until the entire score is processed (Schwartz et al. 2024). This leads to a loss of critical magnitude information (Testolin 2024). Such errors undermine the accuracy of IAA tasks, where precise numerical interpretation is essential.

Unlike mathematical tasks that rely on logical reasoning from symbolic inputs to derive precise answers, IAA quantifies subjective aesthetics by mapping visual features

*These authors contributed equally.
†Corresponding author



(a) Direct scoring (b) Level-based scoring (c) Our regression-based scoring

Figure 2: Comparison of MLLM-based scoring methods: (a) Direct scoring, (b) Level-based scoring, and (c) Our regression-based scoring, which is capable of: (*i*) **avoiding tokenization of scores as text**, (*ii*) **ensuring correct positioning of scores within the text**, and (*iii*) **preventing interference between scores and text during simultaneous constraints on both**.

to scores. Training MLLMs on mathematical datasets focuses on symbol manipulation, which does not enhance their ability to map visual features to numerical scores. Besides, IAA datasets are small and lack unified scoring standards. Consequently, unlike in mathematical tasks—where large-scale training data can enhance MLLMs' sensitivity to number —the application of MLLMs to IAA faces inherent challenges in achieving comparable score sensitivity due to data scarcity and subjectivity. ***A detailed discussion are provided in the Appendix.***

Score Quantization Errors. MLLMs generate coherent text through an autoregressive process, predicting the next token's probability distribution step by step and selecting the most likely token. However, each token prediction resembles a multi-class classification task, disrupting the continuity of scores. In classification, outputs fall into discrete categories—either correct or incorrect—with no measure of distance between them. This disrupts the inter-class relationships among tokens representing scores, resulting in score quantization errors (Talebi and Milanfar 2018).

Precision Loss. To address the imprecision of MLLMs in predicting scores for IAA tasks, Q-Align (Wu et al. 2024c) adopts a level-based scoring method, converting scores into five coarse-grained categories: bad, poor, fair, good, and excellent. However, this approach struggles to differentiate images with varying aesthetic qualities within the same category. For example, in the AVA dataset (Murray, Marchesotti, and Perronnin 2012), images with scores of 1.8 and 3.16 are both categorized as "bad" during training, despite their noticeable differences. As a result, the conversion between scores and levels leads to precision loss. Furthermore, Q-Align treats the task as a coarse rating classification, predicting the image's aesthetic level and calculating the final scores using a weighted sum of level probabilities, which still leads to score quantization errors.

To address above issues, we propose Regression Optimized Components for MLLMs (ROC4MLLM). Unlike traditional classification-based scoring methods, ROC4MLLM uses regression for more precise and continuous scoring, as shown in Fig. 2(c). First, we decouple score representation from text representation, along with a score position token to indicate the position of the scores in the text, which eliminates the need for text-based tokenization of scores. Second, we transform classification into regression for score prediction and decouple the loss computation between scores and text, eliminating score quantization errors, while preventing interference between scores and text in the loss computation. Contributions are concluded as follows:

- We reveal the limitations of MLLMs in IAA: insensitivity to scores, score quantization errors caused by classification-based decoder mechanism, and precision loss caused by level-based methods. By highlighting these issues, we emphasize the need for more robust, score-sensitive methods in IAA.
- The proposed ROC4MLLM decouples score and text representations and shifts from classification to regression-based loss computation. This method addresses the identified limitations of MLLMs and provides a practical solution for the community, to improve MLLMs' performance in score-sensitive tasks like IAA.
- ROC4MLLM achieves state-of-the-art (SOTA) results on five IAA datasets, as shown in Fig. 1. Additionally, it can be seamlessly integrated with existing MLLMs, significantly boosting their performance in IAA tasks.

Related Work

Generally, IAA involves both quantitative and qualitative tasks: 1) aesthetic scoring, which includes aesthetic binary classification (Datta et al. 2006; Luo and Tang 2008), aesthetic score regression (Ma, Liu, and Chen 2017; He et al. 2023a), and score distribution prediction (Chen et al. 2020; She et al. 2021); and 2) aesthetic commenting (Ke et al. 2023; Liu et al. 2024c), which entails generating comments about the aesthetic attributes of an image.

Traditional CNN-based and Transformer-based methods demonstrate strong performance in aesthetic scoring tasks (Zhu et al. 2020; Ke et al. 2021; She et al. 2021; He et al. 2022; Tu et al. 2022; He et al. 2023b). However, their ability to address the intrinsic subjectivity of IAA remains limited. These methods primarily focus on extracting low-level features and lack the capacity for understanding and reasoning about high-level information. Moreover, provid-

Dataset: **PARA**

Q: Rate the composition, DoF, light and content of the image.
A: The composition rate of the image is 7.57, the DoF rate of the image is 7.26, the light rate of the image is 7.29, the content rate of the image is 6.99.

Dataset: **AVA**

Q: Please rate the aesthetics of the image and provide some comments.
A: The aesthetic rate of the image is 6.93. I'm uncertain about the tree in the center of the image, I think I might prefer the picture without it.

Figure 3: Examples of our method's output across five datasets, with all scores are normalized to a range of 1 to 10.

ing only a score limits the explainability, making it difficult to fully understand the aesthetics of an image.

MLLMs, such as LLaVA (Liu et al. 2024b,a), mPLUG-Owl (Ye et al. 2023, 2024), and the QwenVL (Bai et al. 2023; Wang et al. 2024) series have shown remarkable performance in image captioning and demonstrated potential in understanding high-level visual content (Liu et al. 2025). Recent studies also show their ability to perceive low-level visual attributes (Zhang et al. 2023). Building on these strengths, recent efforts (Zhou et al. 2024; Huang et al. 2024a; Wu et al. 2024b; Huang et al. 2024b) have applied these models to IAA, adapting them for aesthetic evaluation tasks. However, as highlighted in the Introduction section, MLLMs face difficulty in accurately predicting scores while simultaneously generating text. Most MLLMs treat scores as text and tokenize a complete score into multiple individual tokens (Fig. 2(a)), which makes them insensitive to scores and introduces score quantization errors.

Despite a few 7B-scale models achieving over 50% accuracy on MATH (Hendrycks et al. 2021), these models benefit from extensive supervised fine-tuning with large mathematical datasets, such as Qwen2.5-Math with over 1 trillion tokens. However, limited high-quality data for IAA hinders data-driven approaches, and our method improves MLLM scoring in IAA without extra data. Besides, strong mathematical performance does not ensure success in scoring tasks; for example, Qwen2 significantly outperforms LLaMA2 on MATH, yet in experiments, we found Qwen2-VL underperforms mPLUG-Owl2 in aesthetic scoring.

To enhance the accuracy of MLLMs in aesthetic scoring, Q-Align (Wu et al. 2024c) converts aesthetic scores into levels during training and maps these levels back to scores during inference, which has been adopted by subsequent methods (Zhou et al. 2024; You et al. 2025). However, as shown in Fig. 2(b), level-based methods require MLLMs to classify the aesthetics of an image without directly utilizing ground truth scores during training, which leads to precision loss. Furthermore, they do not account for the generation of subsequent text and the interference between scores and text.

To overcome these limitations in applying MLLMs to IAA, our proposed ROC4MLLM decouples scores from text and then regresses the score, alongside regular text generation, as illustrated in Fig. 2(c). This method enhances the robustness and precision of MLLMs in both quantitative and qualitative tasks, enabling them to generate continuous text while accurately predicting scores.

Method

We start by analyzing existing approaches, identifying two key issues: tokenization-induced score sensitivity (Problem I) and inadequacy of classification loss for continuous scores (Problem II). Subsequently, the interference between score and text loss (Problem III) is revealed, assuming that an additional regression loss for the score is used in a straightforward manner. To address these problems, we then propose an architecture presented in Fig. 4. Our ROC4MLLM optimizes score representation and the computation of the loss function, enabling precise numerical predictions without compromising contextual coherence.

Problem Analysis

The text generation operates through autoregressive token prediction. Specifically, at each step, the model computes the probability distribution $\hat{p}$ based on input I (which typically includes an image and a prompt) and preceding tokens $\hat{y}_0, \ldots, \hat{y}_{t-1}$. The token with the highest probability in the MLLMs' vocabulary V is selected as the next token $\hat{y}_t$:

$$\hat{y}_t = \arg\max_{y \in V} P(y | I, \hat{y}_0, \ldots, \hat{y}_{t-1}). \quad (1)$$

However, applying this process directly to aesthetic scoring exposes three problems.

Problem I: Tokenization-Induced Score Sensitivity. Tokenization strategies break scores into multiple discrete tokens. For example, "7.25" becomes ["7", ".", "2", "5"], forming a subsequence $S_{pr} = [\hat{y}_k, \ldots, \hat{y}_{k+j}]$. Unlike free-form text, where semantic consistency allows error correction through later tokens, numerical scores in IAA tasks require absolute precision. ***Even a slight error in one token can disrupt all following predictions, which causes MLLMs to struggle with score sensitivity, making direct scoring difficult.***

To overcome the aforementioned problem, level-based scoring methods simplify aesthetic score prediction. They classify images into predefined text rating levels, bypassing the need to split scores into tokens. However, these approaches encounter three challenges: 1) they do not directly use the ground truth scores, resulting in a loss of fine-grained information; 2) the reliance on common rating levels may lead to semantic ambiguity with existing words; 3) varying tokenization methods among different MLLMs can cause a single level to be split into a different number of tokens, thereby impeding their application to other MLLMs.

Problem II: Inadequacy of Classification Loss for Continuous Scores. The model treats token generation as a classification task, with the loss (simplified) computed as cross-entropy over the target sequence $y_0, \ldots, y_n$:

$$L_{text} = -\frac{1}{n} \cdot \sum_{i=0}^{n} \log(P(y = y_i | I, y_0, \ldots, y_{i-1})). \quad (2)$$

However, this ***classification-based loss fails to measure differences between continuous numerical values.*** For instance, predicting "0" or "6" for a ground truth score of "7" may yield the same loss, despite their different magnitudes. Similarly, predictions of "6.25" and "7.24" for a ground truth score of "7.25" may incur equal penalties. The loss focuses only on token accuracy, ignoring the numerical relationships between predicted and target values.

Level-based scoring methods retain the original model architecture, which means the training still relies on the loss function in Eq. 2. Consequently, both direct and level-based scoring in MLLMs lead to score quantization errors.

Problem III: Interference between Score and Text Loss. In level-based scoring methods, a single word represents the score. This seems like a simple way to shift from score classification to regression, avoiding issues with classification-based decoding in MLLMs. *But is this really the case?* In practice, this approach introduces interference between score and text loss. The loss function is expressed as:

$$\begin{aligned} L =& L_{text} + L_{score} \\ =& -\frac{1}{n} \cdot \sum_{i=0, i\neq t}^{n} \log(P(y = y_i | I, y_0, \ldots, y_{i-1})) \\ & - \frac{1}{n} \cdot \log(P(y = T_{level} | I, y_0, \ldots, y_{t-1})) \\ & + L_{score}(S_{pr}, S_{gt}), \end{aligned} \tag{3}$$

where L_{score} is a regression loss to predict scores, and T_{level} is one of the rating levels V', selected from the MLLMs' vocabulary V. The predicted score S_{pr} is computed based on predefined weight w_i using following formula:

$$S_{pr} = \sum_{y' \in V'} w_i \cdot \frac{\exp(P(y'|I, y_0, \ldots, y_{t-1}))}{\sum_{y \in V'} \exp(P(y|I, y_0, \ldots, y_{t-1}))}. \tag{4}$$

This approach, however, leads to interference between the text and score losses. The score loss L_{score} depends on S_{pr}, which is computed from the probabilities of tokens in V'. Since T_{level} is an element of V', the computation of S_{pr} is directly influenced by the predicted probability $P(y = T_{level}|I, y_0, \ldots, y_{t-1})$. However, the text loss seeks to maximize the probability of T_{level}, while the score loss adjusts the weighted sum of probabilities across all tokens in V' to match S_{pr}. These inconsistent goals can cause issues during optimization, the text loss gradient increases $P(y = T_{level}|I, y_0, \ldots, y_{t-1})$, while the score loss gradient may tweak it to optimize S_{pr}. This interference can negatively impact model convergence and performance.

Our Proposed ROC4MLLM

Decoupling Score Representation from Text Representation. We introduce a score space within the word space of MLLMs, consisting of a score position token and a set of score representation tokens. This setup allows for the independent representation of scores. Specifically, we propose a position token, $T_p = \{[pos]\}$, to indicate whether the current output is a score, We also add a set of representation tokens, $T_s = \{[\mathrm{S}_1], [\mathrm{S}_2], \ldots, [\mathrm{S}_N]\}$ to the MLLMs' vocabulary V,

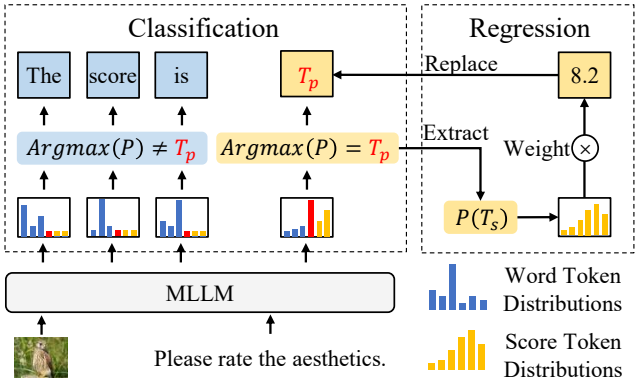


Figure 4: Architecture of ROC4MLLM: We decouple score representation from text by introducing score representation tokens T_s and a score position token T_p in the MLLMs' word space. A classification loss constrains text generation and score position prediction, while a regression loss constrains score prediction. T_p prevents interference between score and text losses, and during inference, the predicted score is placed at the T_p position.

expanding the vocabulary to $V_s = [V \quad T_p \quad T_s]$. Here, N denotes the number of new tokens.

When the probability distribution $\hat{p}$ predicted by the MLLMs shows T_p with the highest probability, it indicates that the current output is a score. At this point, MLLMs generate T_p and ***extract the probability of each token in*** T_s ***from*** $\hat{p}$ ***for the subsequent score computation.*** To predict scores independently, we map the score tokens to the final score S_{pr} using this formula:

$$S_{pr} = \sum_{y' \in T_s} s_i \cdot \frac{\exp(P(y'|I, y_0, \ldots, y_{t-1}))}{\sum_{y \in T_s} \exp(P(y|I, , y_0, \ldots, y_{t-1}))}. \tag{5}$$

Here, s_i represents the score corresponding to each token in T_s. Next, we replace T_p in the generated text with S_{pr}. ***This approach offers three main advantages.*** First, it allows a score to be represented as a single token with arbitrary precision, separating score prediction from textual content. Second, incorporating custom tokens avoids semantic ambiguity with existing word tokens. Third, representing a score with a single custom token, unlike level-based methods, prevents a word from being split into multiple tokens.

Decoupling Score Loss From Text Loss. We introduce a score-specific loss function, L_{score}, which shifts score prediction from classification to regression. By incorporating T_p, we update the loss function from Eq. 3 to:

$$\begin{aligned} L = & -\frac{1}{n} \cdot \sum_{i=0, i\neq t}^{n} \log(P(y = y_i | I, T_p, y_0, \ldots, y_{i-1})) \\ & - \frac{1}{n} \cdot \log(P(y = T_p | I, T_p, y_0, \ldots, y_{t-1})) \\ & + L_{score}(S_{pr}, S_{gt}). \end{aligned} \tag{6}$$

Unlike the loss function in *Problem III*, we do not use T_s to represent the score position in the text as T_{level} does. In-

Metric (AVA)	CNN-based				Transformer-based			Mamba-based
	NIMA (TIP 2018)	BIAA (TCYB 2020)	HGCN (CVPR 2021)	TANet (IJCAI 2022)	MUSIQ (ICCV 2021)	MaxVit (ECCV 2022)	EAT (ACMMM 2023)	AesMamba (ACMMM 2024)
$\mathcal{S}\uparrow$	0.612	0.651	0.665	0.758	0.726	0.708	0.759	0.751
$\mathcal{P}\uparrow$	0.636	0.668	0.687	0.765	0.738	0.745	0.770	0.760
$\mathcal{M}\downarrow$	0.454	0.473	0.460	0.414	0.438	0.416	0.372	-

Metric (AVA)	MLLM-based							
	LLaVA 1.5 (CVPR 2024)	Q-Instruct (CVPR 2024)	AesExpert (ACMMM 2024)	mPLUG-Owl2 (CVPR 2024)	Qwen2-VL (arXiv 2024)	Q-Align (ICML 2024)	CALM (AAAI 2025)	ROC4MLLM (Ours)
$\mathcal{S}\uparrow$	0.781	0.782	0.784	0.788	0.778	0.822	0.815	**0.833**
$\mathcal{P}\uparrow$	0.777	0.778	0.779	0.786	0.780	0.817	0.829	**0.832**
$\mathcal{M}\downarrow$	0.371	0.369	0.369	0.361	0.429	0.934	-	**0.319**

Table 1: Comparison of 16 methods on the AVA dataset, with the bold numbers indicating the best results.

Method	TAD66K		ICAA17K	
	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$
NIMA (TIP 2018)	0.390	0.405	0.809	0.815
BIAA (TCYB 2020)	0.417	0.431	0.820	0.829
MUSIQ (ICCV 2021)	0.489	0.517	0.835	0.841
HGCN (CVPR 2021)	0.486	0.493	0.826	0.831
TANet (IJCAI 2022)	0.513	0.531	0.829	0.836
MaxVit (ECCV 2022)	0.484	0.513	0.855	0.874
EAT (ACMMM 2023)	0.517	0.546	-	-
ICAA (ICCV 2023)	-	-	0.873	0.890
AesMamba (ACMMM 2024)	0.475	0.503	-	-
QAligns (ICML 2024)	0.506	0.536	0.761	0.781
ROC4MLLM (Ours)	**0.518**	**0.550**	**0.894**	**0.903**

Table 2: Comparison on TAD66K and ICAA17K.

Method	PARA		AADB	
	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$
NIMA (TIP 2018)	0.877	0.862	0.700	0.711
BIAA (TCYB 2020)	0.858	0.886	0.710	0.733
MUSIQ (ICCV 2021)	0.899	0.918	0.751	0.761
HGCN (CVPR 2021)	0.865	0.881	0.716	0.734
TANet (IJCAI 2022)	0.887	0.899	0.749	0.742
MaxVit (ECCV 2022)	0.902	0.936	0.742	0.748
EAT (ACMMM 2023)	0.909	0.940	0.759	0.767
AesMamba (ACMMM 2024)	0.902	0.936	0.768	0.774
QAligns (ICML 2024)	0.923	0.940	0.762	0.770
ROC4MLLM (Ours)	**0.934**	**0.956**	**0.783**	**0.789**

Table 3: The holistic results on AADB and PARA.

stead, we replace T_s with T_p within the text to eliminate interference between score and text loss. With this adjustment, the computation of S_{pr} excludes tokens associated with the text loss. As a result, the text and score losses function independently, each affecting only its gradient.

In the prediction process, ROC4MLLM generates tokens sequentially until it predicts T_p. At that point, it extracts the probabilities of each token in T_s from $\hat{p}$. By applying Eq. 5, the model computes the score and inserts it into the correct position. Then, it resumes normal token prediction.

Experiments

Experimental Settings

We employ three popular evaluation metrics to assess the models' aesthetic scoring capabilities: Spearman's rank correlation coefficient (SRCC, $\mathcal{S}$), Pearson's linear correlation coefficient (PLCC, $\mathcal{P}$), and mean absolute error (MAE, $\mathcal{M}$). To maintain consistency, we normalize the predicted results to match the value range of each IAA dataset. We applied ROC4MLLM to several popular MLLMs and full fine-tuned them for IAA tasks. Among them, we found that our method performs best with mPLUG-Owl2 (Ye et al. 2024). Therefore, we fine-tuned the model using pre-trained weights from mPLUG-Owl2, with a batch size of 16 across all datasets and a fixed learning rate of $2e-5$. For the scoring loss function, we adopted the smooth L1 loss across all datasets. When training on the AVA dataset, we additionally included cross-entropy loss (*cf.* Appendix for details). To align with the score distribution range provided by the AVA dataset, we initialized the token number N to 10 in all experiments. The impact of N is discussed in the Appendix.

Datasets

We selected three representative IAA datasets to evaluate the ***aesthetic scoring*** capabilities of our method. These datasets include the well-known and general AVA dataset (Murray, Marchesotti, and Perronnin 2012), the theme-oriented TAD66K dataset (He et al. 2022), and the color-oriented ICAA17K dataset (He et al. 2023a). We also validated our model's multi-task learning capability using two ***multi-attribute*** datasets: AADB and PARA. To verify our model's ability to ***predict both text and scores***, we tested it on the AVA-Captions dataset (Ghosal, Rana, and Smolic 2019). We adhered to the official data splits for AVA and AVA-Captions. We selected one comment per image. Our evaluation was conducted on the official AVA-Captions and AVA test set. More dataset details are provided in Appendix.

Method		Attributes in AADB						Attributes in PARA			
		Color Harmony	DoF	Light	Motion Blur	RoT	Vivid Color	Compos -ition	Color	DoF	Light
Kong *et al.* (ECCV 2016)	$\mathcal{P}\uparrow$	0.48	0.46	0.42	0.10	0.21	0.64	0.66	0.69	0.69	0.68
	$\mathcal{S}\uparrow$	0.47	0.48	0.44	0.10	0.23	0.65	0.67	0.67	0.71	0.69
Malu *et al.* (arXiv 2017)	$\mathcal{P}\uparrow$	0.50	0.47	0.48	0.12	0.21	0.62	0.77	0.77	0.82	0.78
	$\mathcal{S}\uparrow$	0.48	0.50	0.48	0.14	0.22	0.64	0.77	0.80	0.80	0.80
MP_{ada} (ACMMM 2018)	$\mathcal{P}\uparrow$	0.48	0.50	0.36	0.16	0.17	0.64	0.65	0.67	0.71	0.72
	$\mathcal{S}\uparrow$	0.48	0.50	0.40	0.13	0.18	0.68	0.67	0.66	0.67	0.72
NIMA (TIP 2018)	$\mathcal{P}\uparrow$	0.48	0.55	0.39	0.12	0.14	0.62	0.72	0.76	0.74	0.77
	$\mathcal{S}\uparrow$	0.46	0.29	0.35	0.12	0.12	0.60	0.74	0.75	0.74	0.74
MUSIQ (ICCV 2021)	$\mathcal{P}\uparrow$	0.43	0.27	0.32	0.03	0.06	0.61	0.77	0.77	0.79	0.76
	$\mathcal{S}\uparrow$	0.44	0.22	0.33	0.04	0.07	0.60	0.76	0.78	0.80	0.77
TANet (IJCAI 2022)	$\mathcal{P}\uparrow$	0.47	0.48	0.48	0.17	0.18	0.68	0.74	0.79	0.77	0.76
	$\mathcal{S}\uparrow$	0.48	0.48	0.48	0.14	0.22	0.63	0.74	0.77	0.78	0.76
MAINet (ACMMM 2024)	$\mathcal{P}\uparrow$	0.51	0.64	0.48	0.19	0.25	0.67	0.83	0.84	0.86	0.85
	$\mathcal{S}\uparrow$	0.50	**0.65**	0.51	0.20	0.23	0.70	0.78	0.81	0.82	0.81
Q-Align (ICML 2024)	$\mathcal{P}\uparrow$	0.55	0.75	0.54	0.10	0.27	0.73	0.91	0.91	0.91	0.91
	$\mathcal{S}\uparrow$	0.55	0.53	0.51	0.10	0.27	0.73	0.89	0.90	0.89	0.89
ROC4MLLM (Ours)	$\mathcal{P}\uparrow$	**0.57**	**0.77**	**0.64**	**0.31**	**0.31**	**0.74**	**0.93**	**0.93**	**0.93**	**0.93**
	$\mathcal{S}\uparrow$	**0.57**	0.61	**0.57**	**0.22**	**0.31**	**0.75**	**0.90**	**0.91**	**0.90**	**0.90**

Table 4: The multi-attribute performance comparison on the AADB and PARA datasets.

Benchmark Models

We compare our ROC4MLLM with nine prevailing non-MLLM-based IAA models (Talebi and Milanfar 2018; Zhu et al. 2020; Ke et al. 2021; She et al. 2021; He et al. 2022; Tu et al. 2022; He et al. 2023b,a; Gao et al. 2024) on holistic IAA datasets, as well as with seven prevailing non-MLLM-based IAA models (Kong et al. 2016; Malu, Bapi, and Indurkhya 2017; Sheng et al. 2018; Talebi and Milanfar 2018; Ke et al. 2021; He et al. 2022; Xie et al. 2024) on multi-attribute IAA datasets. On the AVA dataset, we also compare our method with five popular MLLMs (Liu et al. 2024a; Wu et al. 2024b; Huang et al. 2024a; Ye et al. 2024; Wang et al. 2024). Furthermore, we select an MLLM-based method, Q-Align (Wu et al. 2024c), which represents a SOTA approach for image quality assessment (IQA) and visual question answering (VQA) tasks. To ensure consistency, we retrained Q-Align using the official code across all datasets. We also examine CALM (Liu et al. 2024c), another MLLM-based method for IAA tasks. For the aesthetic comments datasets, we select three popular MLLMs (Liu et al. 2024a; Wang et al. 2024; Ye et al. 2024) as baseline models.

Performance Comparison

Holistic IAA Tasks. Tab. 1 presents performance comparisons with prevailing methods on the AVA dataset. ***Our method achieves the best performance across all metrics.*** Compared with the general MLLMs (not specifically designed for aesthetic scoring), our ROC4MLLM achieves the best performance in terms of the SRCC, PLCC and MAE, and it surpasses the previous best results of its MLLM-based counterparts by +4.5% in the SRCC, +4.6% in the PLCC and -4.2% in the MAE. Notably, it outperforms methods requiring additional low-level visual training data, such as Q-instruct, or aesthetic-related information, like AesExpert.

By contrast, our method does not rely on additional information, showing that ROC4MLLM effectively boosts the accuracy of aesthetic scoring predictions in MLLMs. Despite Q-Align performs well in terms of SRCC and PLCC, its predictions exhibit a higher MAE. In contrast, ROC4MLLM achieves the best results in MAE as well. As shown in Tab. 2, our model also outperforms others on the TAD66K and ICAA17K datasets. Notably, ROC4MLLM outperforms Q-Align by 13.3% in SRCC and 12.2% in PLCC on ICAA17K.

Multi-attribute IAA Tasks. As shown in Tab. 3 and Tab. 4, ROC4MLLM achieves superior performance across all metrics on the AADB and PARA datasets. Notably, the compared methods typically train on one labeled attribute to predict its score and then retrain on another attribute, ***thus requiring multiple steps of training***. In contrast, our method introduces a novel score position token T_p, which allows it to locate the position of multiple scores within the text. ***This allows our model to predict scores for multiple attributes in a single sentence, removing the need for separate training for each attribute.*** After training, our model can predict scores for all attributes simultaneously, offering an elegant solution to the multi-attribute IAA tasks.

Aesthetic Commenting Tasks. Optimizing the performance of MLLMs on aesthetic commenting tasks requires careful attention to how text generation interacts with score prediction. We conducted experiments that incorporate aesthetic comments into the score prediction process, using the AVA-Captions dataset. We employed two metrics: Consensus-

Task	Method	Score Prediction			Comments Generation	
		$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{M}\downarrow$	CIDEr ↑	SPICE ↑
Score Prediction	mPLUG-Owl2	0.761	0.762	0.379	-	-
	Q-Align	0.821	0.815	0.938	-	-
	mPLUG-Owl2 + ROC4MLLM	0.831	0.830	0.322	-	-
Score Prediction & Comments Generation	mPLUG-Owl2	0.758 (-0.3%)	0.758 (-0.4%)	0.382 (+0.3%)	0.082	**0.067**
	Q-Align	0.819 (-0.2%)	0.818 (+0.3%)	0.895 (-4.3%)	0.069	0.061
	mPLUG-Owl2 + ROC4MLLM	**0.835** (+0.4%)	**0.834** (+0.4%)	**0.317** (-0.5%)	**0.085**	0.066
Score Prediction	LLaVA 1.5	0.775	0.773	0.372	-	-
	LLaVA 1.5 + ROC4MLLM	0.793	0.794	0.354	-	-
Score Prediction & Comments Generation	LLaVA 1.5	0.767 (-0.8%)	0.767 (-0.6%)	0.377 (+0.5%)	0.077	**0.066**
	LLaVA 1.5 + ROC4MLLM	**0.794** (+0.1%)	**0.795** (+0.1%)	**0.353** (-0.1%)	**0.078**	0.064
Score Prediction	Qwen2-VL	0.784	0.785	0.405	-	-
	Qwen2-VL + ROC4MLLM	0.831	0.829	0.327	-	-
Score Prediction & Comments Generation	Qwen2-VL	0.780 (-0.4%)	0.781 (-0.4%)	0.415 (+1.0%)	0.089	**0.069**
	Qwen2-VL + ROC4MLLM	**0.831** (+0.0%)	**0.830** (+0.1%)	**0.325** (-0.2%)	**0.090**	0.068

Table 5: Results of representative MLLMs for predicting scores alone and for simultaneously predicting scores and generating text. Most methods, except ours, show reduced score prediction performance due to interference between scores and text.

Method	AVA		ICAA17K	
	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$	$\mathcal{S}\uparrow$	$\mathcal{P}\uparrow$
Baseline	0.788	0.786	0.688	0.707
Regression w/o Loss Decoupling	0.825	0.824	0.873	0.880
Regression w/ Loss Decoupling	**0.833**	**0.832**	**0.894**	**0.903**

Table 6: Ablation studies on AVA and ICAA17K.

based Image Description Evaluation (CIDEr) (Vedantam, Lawrence Zitnick, and Parikh 2015) and Semantic Propositional Image Caption Evaluation (SPICE) (Anderson et al. 2016; Chang, Lu, and Chen 2017). Detailed experimental settings are provided in Appendix.

As shown in Tab. 5, incorporating comment data degrades the scoring performance of all baseline models. Even Q-Align shows decreased SRCC. This result highlights that existing classification-based (direct scoring or level-based scoring) MLLMs are unable to prevent interference between scores and text. In contrast, our method remains stable and achieves marginal improvements by effectively leveraging additional textual information. For comment generation, Q-Align's use of rating levels for aesthetic classification may impair the generation of aesthetically relevant words in subsequent comments, ultimately leading to reductions in CIDEr and SPICE scores. However, integrating our method into these baseline models maintains the quality of comment generation. Fig. 3 provides examples of ROC4MLLM's output on the datasets corresponding to these tasks.

Ablation Study. Tab. 6 presents the results of ablation experiments on AVA and ICAA17K. We first separated score representation from text representation and switched from classification to regression for score prediction. This change boosted the baseline model's performance by 3.7% in SRCC and 3.8% in PLCC on the AVA dataset, and by 18.5% in SRCC and 17.3% in PLCC on the ICAA17K dataset. Then, we decoupled the loss computation for score and text, eliminating interference between them. This step further increased performance by 0.8% in SRCC and PLCC on the AVA dataset, and by 2.1% in SRCC and 2.3% in PLCC on the ICAA17K dataset. These step-by-step decoupling efforts lead to consistent performance gains, confirming the value of decoupling both representation and loss computation.

Enhancement for other MLLMs

ROC4MLLM provides complementary solutions that can be seamlessly integrated into existing MLLMs. Specifically, as described in the Method section, we add score representation and position tokens to the MLLMs' vocabulary and replace the original loss function with Eq. 6. We regress the scores based on the predicted probability distribution of the position token's location, as described in Eq. 5.

We apply ROC4MLLM to three MLLMs: LLaVA 1.5, Qwen2-VL, and mPLUG-Owl2. As shown in Tab. 5, ROC4MLLM consistently enhances the SRCC, PLCC, and MAE metrics. This demonstrates that our method alleviates the issues of insensitivity to scores and score quantization errors that are commonly observed in existing MLLMs.

Conclusions

This paper examines the limitations of current MLLM-based methods in IAA. We reveal that classification-based decoding mechanisms in MLLMs limit their effectiveness in aesthetic score prediction tasks. To overcome these challenges, we propose ROC4MLLM, which achieves SOTA performance on five IAA datasets, enabling accurate score prediction without interference between scores and text.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 62502040 and the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation under Grant GZC20251056.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 382–398. Springer.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Chang, K.-Y.; Lu, K.-H.; and Chen, C.-S. 2017. Aesthetic critiques generation for photos. In *ICCV*, 3514–3523.

Chen, Q.; Zhang, W.; Zhou, N.; Lei, P.; Xu, Y.; Zheng, Y.; and Fan, J. 2020. Adaptive Fractional Dilated Convolution Network for Image Aesthetics Assessment. In *CVPR*, 14114–14123.

Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 288–301. Springer.

Gao, F.; Lin, Y.; Shi, J.; Qiao, M.; and Wang, N. 2024. AesMamba: Universal Image Aesthetic Assessment with State Space Models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7444–7453.

Ghosal, K.; Rana, A.; and Smolic, A. 2019. Aesthetic image captioning from weakly-labelled photographs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

He, S.; Ming, A.; Li, Y.; Sun, J.; Zheng, S.; and Ma, H. 2023a. Thinking image color aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21838–21847.

He, S.; Ming, A.; Zheng, S.; Zhong, H.; and Ma, H. 2023b. Eat: An enhancer for aesthetics-oriented transformers. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1023–1032.

He, S.; Zhang, Y.; Xie, R.; Jiang, D.; and Ming, A. 2022. Rethinking Image Aesthetics Assessment: Models, Datasets and Benchmarks. *IJCAI*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. *NeurIPS*.

Huang, Y.; Sheng, X.; Yang, Z.; Yuan, Q.; Duan, Z.; Chen, P.; Li, L.; Lin, W.; and Shi, G. 2024a. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5911–5920.

Huang, Y.; Yuan, Q.; Sheng, X.; Yang, Z.; Wu, H.; Chen, P.; Yang, Y.; Li, L.; and Lin, W. 2024b. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. MUSIQ: Multi-scale Image Quality Transformer. In *ICCV*, 5148–5157.

Ke, J.; Ye, K.; Yu, J.; Wu, Y.; Milanfar, P.; and Yang, F. 2023. Vila: Learning image aesthetics from user comments with vision-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10041–10051.

Kong, S.; Shen, X.; Lin, Z.; Mech, R.; and Fowlkes, C. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 662–679. Springer.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 216–233. Springer.

Liu, Y.; Liu, S.; Gao, J.; Jiang, P.; Zhang, H.; Chen, J.; and Li, B. 2024c. Advancing Comprehensive Aesthetic Insight with Multi-Scale Text-Guided Self-Supervised Learning. *arXiv preprint arXiv:2412.11952*.

Luo, Y.; and Tang, X. 2008. Photo and video quality evaluation: Focusing on the subject. In *Eur. Conf. Comput. Vis.*, 386–399.

Ma, S.; Liu, J.; and Chen, C. W. 2017. A-Lamp: Adaptive Layout-Aware Multi-patch Deep Convolutional Neural Network for Photo Aesthetic Assessment. In *CVPR*, 722–731.

Malu, G.; Bapi, R. S.; and Indurkhya, B. 2017. Learning Photography Aesthetics with Deep CNNs. CoRR abs/1707.03981 (2017). *arXiv preprint arXiv:1707.03981*.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*. IEEE.

Schwartz, E.; Choshen, L.; Shtok, J.; Doveh, S.; Karlinsky, L.; and Arbelle, A. 2024. NumeroLogic: Number Encoding for Enhanced LLMs' Numerical Reasoning. In *Conference on Empirical Methods in Natural Language Processing*.

She, D.; Lai, Y.-K.; Yi, G.; and Xu, K. 2021. Hierarchical Layout-Aware Graph Convolutional Network for Unified Aesthetics Assessment. In *CVPR*, 8475–8484.

Sheng, K.; Dong, W.; Ma, C.; Mei, X.; Huang, F.; and Hu, B.-G. 2018. Attention-based multi-patch aggregation for image aesthetic assessment. In *ACMMM*.

Talebi, H.; and Milanfar, P. 2018. NIMA: Neural Image Assessment. *TIP*, 27(8): 3998–4011.

Testolin, A. 2024. Can neural networks do arithmetic? a survey on the elementary numerical skills of state-of-the-art deep learning models. *Applied Sciences*, 14(2): 744.

Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; and Li, Y. 2022. MaxViT: Multi-Axis Vision Transformer. *ECCV*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Li, C.; Sun, W.; Yan, Q.; Zhai, G.; et al. 2024a. Q-Bench: A Benchmark for General-Purpose Foundation Models on Low-level Vision. In *ICLR*.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Xu, K.; Li, C.; Hou, J.; Zhai, G.; et al. 2024b. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25490–25500.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2024c. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. In *International Conference on Machine Learning*, 54015–54029. PMLR.

Xie, R.; Ming, A.; He, S.; Xiao, Y.; and Ma, H. 2024. " Special Relativity" of Image Aesthetics Assessment: a Preliminary Empirical Perspective. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2554–2563.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13040–13051.

You, Z.; Cai, X.; Gu, J.; Xue, T.; and Dong, C. 2025. Teaching Large Language Models to Regress Accurate Image Quality Scores using Score Distribution. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, P.; Dong, X.; Wang, B.; Cao, Y.; Xu, C.; Ouyang, L.; Zhao, Z.; Duan, H.; Zhang, S.; Ding, S.; et al. 2023. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

Zhou, Z.; Wang, Q.; Lin, B.; Su, Y.; Chen, R.; Tao, X.; Zheng, A.; Yuan, L.; Wan, P.; and Zhang, D. 2024. UNIAA: A Unified Multi-modal Image Aesthetic Assessment Baseline and Benchmark. *arXiv preprint arXiv:2404.09619*.

Zhu, H.; Li, L.; Wu, J.; Zhao, S.; Ding, G.; and Shi, G. 2020. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. *TCYB*.