Bridging Cognitive Gap: Hierarchical Description Learning for Artistic Image Aesthetics Assessment

Henglin Liu[1, 2*], **Nisha Huang**[1, 3*], **Chang Liu**[1†], **Jiangpeng Yan**[1, 5], **Huijuan Huang**[2], **Jixuan Ying**[1], **Tong-Yee Lee**[4], **Pengfei Wan**[2], **Xiangyang Ji**[1†]

[1]Tsinghua University
[2]Kling Team, Kuaishou Technology
[3]Pengcheng Laboratory
[4]National Cheng Kung University
[5]E Fund Management Co., Ltd.
{liu-hl24,hns24,yingjx23}@mails.tsinghua.edu.cn, yanjp13@tsinghua.org.cn, {liuchang2022,xyji}@tsinghua.edu.cn, huanghuijuan.thu@gmail.com, tonylee@ncku.edu.tw, wanpengfei@kuaishou.com

Abstract

The aesthetic quality assessment task is crucial for developing a human-aligned quantitative evaluation system for AIGC. However, its inherently complex nature—spanning visual perception, cognition, and emotion—poses fundamental challenges. Although aesthetic descriptions offer a viable representation of this complexity, two critical challenges persist: (1) data scarcity and imbalance: existing dataset overly focuses on visual perception and neglects deeper dimensions due to the expensive manual annotation; and (2) model fragmentation: current visual networks isolate aesthetic attributes with multi-branch encoder, while multimodal methods represented by contrastive learning struggle to effectively process long-form textual descriptions. To resolve challenge (1), we first present the Refined Aesthetic Description (RAD) dataset, a large-scale (70k), multi-dimensional structured dataset, generated via an iterative pipeline without heavy annotation costs and easy to scale. To address challenge (2), we propose ArtQuant, an aesthetics assessment framework for artistic image which not only couple isolated aesthetic dimensions through joint description generation, but also better model long-text semantics with the help of LLM decoders. Besides, theoretical analysis confirms this symbiosis: RAD's semantic adequacy (data) and generation paradigm (model) collectively minimize prediction entropy, providing mathematical grounding for the framework. Our approach achieves state-of-the-art performance on several datasets while requiring only 33% of conventional training epochs, narrowing the cognitive gap between artistic image and aesthetic judgment. We will release both code and dataset to support future research.

Introduction

"Painting is a mental thing, a thought."

— Leonardo Da Vinci

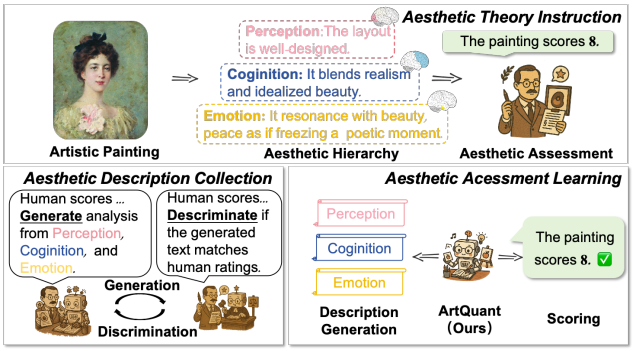


Figure 1: Top: Hierarchical cognitive mechanism of human artistic aesthetics, serving as our theoretical foundation. Bottom left: we employ an iterative data collection framework to optimize the training data distribution, ensuring that the generated aesthetic analysis aligns with human ratings. Bottom right: Dual-task learning (description generation + scoring prediction) enhances model's alignment with human judgement.

With the rapid development of computer vision, remarkable progress has been made in image generation and personalization, making artistic creation (Huang et al. 2025a,b) more accessible to the public. However, the development of quantitative evaluation of artwork aesthetics relatively lags. Among these, aesthetic quality assessment plays a pivotal role, as it bridges the gap between machine-generated outputs and human judgments.

This task is fundamentally challenging due to its inherently multidimensional nature. Prior work has typically addressed it through specialized architectures: Amirshahi et al. (Amirshahi and Denzler 2017) developed color-based features with an SVM classifier for artistic quality assessment. Chen et al. (Chen et al. 2024a) proposed AKA-Net, employing multi-branch networks to separately model color, composition, and content dimensions, while PVAFE (Tang et al. 2025) decomposed aesthetics into content, vivid color,

*These authors contributed equally.
†Corresponding author.

and color harmony components. Similarly, ArtCLIP (Jin et al. 2024b) introduced an assessment framework using contrastive learning on image-comment pairs, but required separate training for each of 10 aesthetic attributes.

However, current AIAA (Artistic Image Aesthetics Assessment) methods are not well-aligned with the human aesthetic process. Studies (Leder and Nadal 2014; Leder et al. 2004; Chatterjee 2003; Huang et al. 2024a, 2025c) indicate that human aesthetic experience is the result of multiple processes, including perceptual, cognitive, and emotional processes. Hierarchical aesthetic descriptions effectively capture the progression, combining comprehensive feature with analytical flexibility to provide a promising solution to address the existing misalignment. However, this direction remains unexplored in prior work, due to substantial challenges in both data and methodological design. (1) **Data**: the collection of human-annotated aesthetic descriptions remains prohibitively expensive, and existing dataset, APDD (Zhang et al. 2024), is overly generalized with insufficient visual evidence and limited to perception technical analysis while neglecting deeper aesthetic dimensions, which is essential for artistic image aesthetics assessment (as shown in Fig. 2). (2) **Method**: the existing method, such as contrastive learning architectures (Jin et al. 2024b) struggle to effectively process long-form textual descriptions. This raises three core research questions. **(1) Can hierarchical descriptive texts serve as effective carriers of aesthetic prior knowledge? (2) What are the specific contributions of different levels of textual information to performance improvement? (3) Beyond the sufficiency of textual information itself, are there other mechanisms that regulate the aesthetic capabilities of MLLM?** An in-depth exploration of these questions will contribute to a deeper understanding of the role of textual information for AIAA.

To explore these questions, we propose ArtQuant, the first MLLM tailored for the AIAA task, specializing in aesthetic score prediction. We first present an aesthetic description generation pipeline to alleviate the scarcity of aesthetic description data. We propose a multi-level description generation framework inspired by the data annotation approach (Tan et al. 2024). Specifically, we get distribution on certain dataset for better human-preference-aligned and create an aesthetic template for systematic and comprehensive generation. To ensure the quality of the generated content, we adopt the alignment score with human aesthetic ratings judged by LLM as an evaluation metric (Li et al. 2024). The framework not only preserves the statistical patterns of the original data but also enhances the rationality and diversity of the generated comments. Using this framework, we constructed the **R**efined **A**esthetic **D**escriptions (RAD) dataset, which contains a total of 70k comments data.

Building upon the RAD, we introduce a novel auxiliary description task to effectively learn aesthetic representations. This is based on a key hypothesis: generating aesthetic descriptions enables more comprehensive understanding of artistic elements. By training the model to produce textual descriptions of art images, it is forced to explicitly decompose aesthetic elements such as brushstroke, lighting, and mood. This decomposition yields interpretable features that serve as aesthetic prior knowledge for score prediction. Notably, more complete description levels lead to better score prediction performance.

APDD (Human annotation): The composition is rigorous and complete, the shape is exquisite, accurate and realistic, the character landscape is delicately portrayed, and the details are diverse. *(1. Overly generalized with insufficient visual evidence. 2. Limited to technical analysis.)*

Ours: The painting presents a harmonious blend of classicism and landscape, showcasing a serene scene that invites contemplation... The placement of the castle and the figures by the water creates a sense of depth, drawing the viewer into the scene *(Specifically analyzed how the positions of the castle and the characters create depth of field)* ... The warm tones of the sunset contrast beautifully with the cooler hues of the landscape, creating a harmonious balance *(It describes in detail the contrast effect of warm and cool colors)* ... The theme conveys a connection between humanity and nature, resonating well within the classical framework. *(Point out the theme of "Man and Nature" clearly)* ... The painting captures a moment in time, inviting viewers to ponder the beauty and quietness of nature. *(Deeply interpreted the emotional connotation of the painting and the viewer's experience)* ...

Figure 2: Grey font indicates limitations in the existing dataset APDD (Jin et al. 2024b). The green, blue, and orange fonts represent Perception, Cognition and Emotion respectively, demonstrating the fine-grained and multi-level nature of our methodology.

We formalize this intuition via information theory, proving that learning auxiliary descriptions inherently bounds scoring error. Specifically, the prediction uncertainty $H(Y|Z)$ is constrained by two key factors: (1) description sufficiency ($H(Y|D)$): the relevance of generated text to scoring, guaranteed by our dataset's curated descriptions and (2) representation quality ($H(D|Z)$): the model's ability to reconstruct accurate descriptions from images, ensured by our pretraining objective.

Besides, a main challenge is predicting continuous scores with MLLMs, which are built for discrete token outputs. Previous approaches (Wu et al. 2024; You et al. 2025) struggle to fit real labels accurately and handle datasets lacking variance information (common in art datasets like APDD (Jin et al. 2024b) and BAID (Yi et al. 2023)). Inspired by mapping, we propose Score-Based Distribution Estimation method. It models numerical outputs as expected values of discrete token probabilities by minimizing the error between the target distribution's expectation and the ground truth. This approach inherently reduce label errors while eliminate the dependence on variance label.

To validate the effectiveness of our method, we test it on three common AIAA datasets. Beyond achieving state-of-the-art performance, our model converges with fewer training epochs. Our main contributions are as follows:

- To overcome the rigid and isolated nature of traditional aesthetic modeling approaches, we propose a dual-level solution: (1) At the data level, we develop a scalable framework featuring multi-level generation with human-preference-aligned aesthetic templates and LLM-validated quality control (§3.1); (2) At the methodology level, we introduce an auxiliary description generation task that explicitly decomposes and learns fine-grained aesthetic features as interpretable intermediate represen-

tations for scoring and the Score-Based Distribution Estimation method to better model scoring (§3.2).

- We formalize this intuition with information-theoretic proof that auxiliary description learning bounds scoring error via description sufficiency ($H(Y|D)$) and representation quality ($H(D|Z)$) (§4).
- Extensive experiments demonstrate our method's superiority over baseline approaches on three datasets, within just 33% of the training epochs required by conventional approaches (§5).

Related Work

Artistic Image Aesthetics Assessment

Recent advances in artistic image aesthetics assessment (AIAA) have evolved from manual feature extraction to data-driven approaches. Initial work relied on manually designed features to quantify aesthetic properties. Li et al. (Li and Chen 2009) proposed a segmentation-based model to capture global composition and local attractiveness in paintings separately. The advent of large-scale datasets and self-supervision marked a significant shift. Yi et al. (Yi et al. 2023) contributed to the BAID dataset (60,000+ artworks) and the SAAN architecture, which jointly learns style and aesthetic features without explicit labels. Further refinements incorporated multi-attribute supervision: Chen et al. (Chen et al. 2024a) proposed AKA-Net to amalgamate attribute-specific knowledge. Jin et al. (Jin et al. 2024b) introduce ArtCLIP, a style-specific art assessment framework that employs contrastive learning on paired artistic images and short aesthetic comments. While recent work like ArtCLIP uses aesthetic comments, existing methods underutilize rich, semantically meaningful descriptions. To bridge this gap, we propose auxiliary description learning for multimodal foundation models, which leverages fine-grained textual annotations to enhance aesthetic understanding.

Multimodal Large Language Models

Recent advances in Multimodal Large Language Models (MLLMs) (Liu et al. 2023; Bai et al. 2023, 2025) have demonstrated remarkable capabilities in bridging vision and language modalities. In the domain of artistic image assessment, systems like AesExpert (Huang et al. 2024b) leverage rich aesthetic critique databases to fine-tune multimodal foundation models, enabling aesthetic-related question answering. Similarly, GalleryGPT (Bin et al. 2024) harnesses large multimodal models' perceptual and generative strengths to produce comprehensive art analyses. However, none of these existing approaches provide quantifiable aesthetic scoring, significantly restricting their practical utility in real-world applications.

In contrast to the aforementioned works, our approach leverages large language models to generate multi-level aesthetic descriptions, better capturing the hierarchical nature of human aesthetic perception and significantly enhancing quantitative scoring capabilities for artworks.

Method

Our approach comprises two core parts: (1) Aesthetic description collection and (2) Aesthetic Acessment Learning.

Aesthetic Description Collection

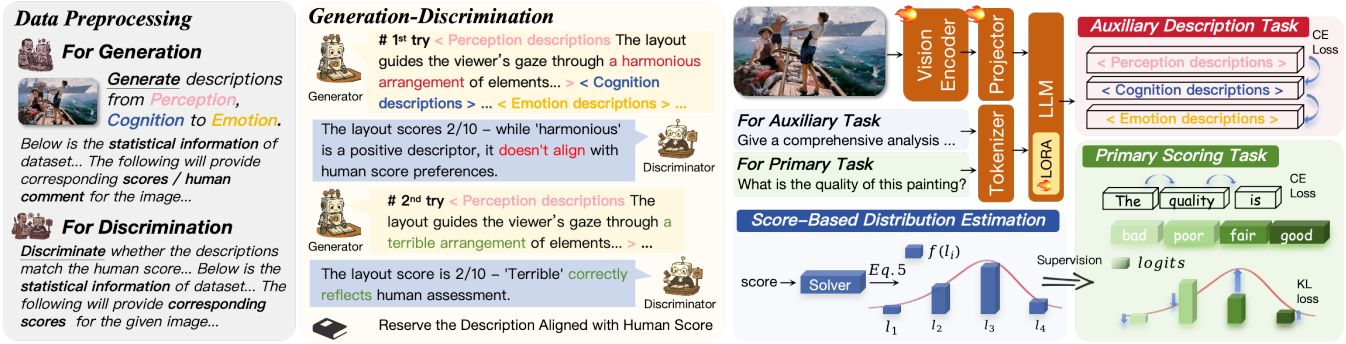
To produce meaningful aesthetic descriptions for score training, our hierarchical generation pipeline comprises three key components: (1) aesthetic data preprocessing that eliminates systematic bias using dataset statistics and emulates human judgment hierarchy, (2) structured description generation that synthesizes multi-level aesthetic analysis, and (3) discriminative quality control that ensures score-description consistency through iterative refinement.

Aesthetic data preprocessing. Instead of directly using artistic images as visual inputs and relying solely on artistic scores as human aesthetic guidance, we introduce two additional preprocessing steps. Due to differences in the annotating population, different datasets often have different data distributions. Samples with high absolute scores but low relative scores can mislead large models to generate descriptions with wrong preferences. Therefore, we need to combine the scores with the dataset distribution for correction. We incorporate dataset-level statistics (mean, median, and variance) as additional conditioning inputs to the MLLM. To emulate the hierarchical nature of human artistic aesthetic judgment (Leder and Nadal 2014), we propose a three-level generative framework that systematically progresses from perception to cognitive processing and emotional evaluation, mirroring the staged refinement of aesthetic appreciation in human experience.

Description generation. After preprocessing, we obtain a quadruple (x_v, s, D, T), where x_v denotes the visual output, $s \in [0, M]$ its quality score, $D \sim \mathcal{N}(\mu, \sigma^2)$ the dataset's score distribution, and T a set of aesthetic templates. These inputs are processed by $\mathcal{G}_{\text{gen}}$ (Eq. 1), where P structure these aesthetic information.

$$y = \mathcal{G}_{\text{gen}}(x_v, P(s, \mu, \sigma, T)) \tag{1}$$

Description discrimination. Inspired by *LLM-as-Judge (Li et al. 2024)*, we propose an iterative framework to ensure the generated analyses are consistent with aesthetic scores (as shown in Fig. 3a). The discriminator $\mathcal{G}_{\text{dis}}$ evaluates the consistency between the aesthetic scores and the generated analyses, thereby regulating output quality. Specifically, our generator $\mathcal{G}_{\text{gen}}$ is implemented using GPT-4o (Yang et al. 2023), whose multimodal reasoning synthesizes analyses can effectively integrate visual feature analysis with aesthetic information; the discriminator $\mathcal{G}_{\text{dis}}$ employs DeepSeek-chat (Bi et al. 2024) for alignment verification, ensuring that scores and visual content remain relatively independent. Although the current large language models inevitably have the problem of hallucination, we verified that our solution is still superior to human manual annotation in the experiment part. This is because human annotation is overly simplistic and generalized, whereas our method is more detailed and structured in a way that better aligns with the human aesthetic process.



(a) Aesthetic Description Collection (b) Multi-Task Aesthetic Training (MAT)

Figure 3: (a) We leverages a scalable, iterative framework to generate hierarchical aesthetic descriptions, ensuring the descriptions align with human scoring. (b) By employing Multi-Task Aesthetic Training to decompose hierarchical aesthetic elements and integrating a high-precision Score-Based Distribution Estimation method, ArtQuant achieves superior alignment with human artistic aesthetic scoring.

Aesthetic Acessment Learning

Auxiliary description task. To fully utilize the hierarchical analytical insights embedded in aesthetic descriptions, inspired by the inherent capabilities of MLLMs in generating textual outputs, we introduce a novel auxiliary task, aesthetic analysis generation, to systematically extract and integrate these insights into our framework. Specifically, we supervise our model to generate detailed aesthetic descriptions for images through supervised fine-tuning (SFT) on the RAD dataset's descriptions. Following standard practices in LLM training, we employ cross-entrophy loss (Eq. 2) that predicts subsequent tokens for aesthetic descriptions generation.

$$\mathcal{L}_{CE} = -\sum_{i=1}^{T} \log P(t_i|t_1, t_2, \ldots, t_{i-1}), \tag{2}$$

where t_i is the i-th token and T is the sequence length.

Primary scoring task. For accurate score estimation using MLLMs, we convert the scores into token representations that align with the model's output space (Wu et al. 2024; You et al. 2025). Let i denote a discrete *level* (e.g., excellent, good), l_i be the *score* associated with level i, and p_i be the model's predicted probability for level i. The score x is the expectation of the level scores, computed as follows:

$$x = \sum_i p_i l_i. \tag{3}$$

To construct probability p_i for ground truth x_{gt}, the probability distribution can be modeled as $f = N(\mu, \sigma)$ with μ as the Mean Opinion Score (MOS) and σ as the variance of human annotations, where d is the width of each level region (You et al. 2025):

$$L(x) = \int_{l_i - \frac{d}{2}}^{l_i + \frac{d}{2}} f(x)dx. \tag{4}$$

Unlike existing methods (You et al. 2025; Wu et al. 2024) that rely on dataset variance information and risk introducing errors before training, we propose a **Score-Based Distribution Estimation** method. Our solution stems from a mapping perspective: during inference, the final score is computed as the expectation of discrete levels (Eq. 3), while training aims to align predicted level probabilities $p(i_{pred})$ with ground-truth distributions $p(i_{gt})$. Therefore, we can minimize the error between the expectation and the labels by optimizing the distribution parameters. However, due to the property that the sum of probabilities is 1, corresponding constraints need to be added during the solving process. The learning objective is formalized as:

$$\begin{aligned} \mu^*, \sigma^* = \arg\min \left\| \sum f(l_i) l_i - x \right\|_2, \\ \text{s.t.} \sum f(l_i) = 1. \end{aligned} \tag{5}$$

Here, we present the formulation where f denotes the probability density function of the normal distribution: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. We employ an automatic derivation of optimal mean and variance parameters through a numerical solver. This formulation enables our MLLM framework to model scores more effectively, improving prediction accuracy without sacrificing computational performance.

For training, we adopt a hybrid loss combining cross-entropy (CE) and Kullback-Leibler (KL) divergence terms following DeQA (You et al. 2025). The CE loss $\mathcal{L}_{ce}$ supervises the prefix token sequence '*The quality of the painting is*' to establish score-related text representations. In contrast, the KL loss (Eq. 6) minimizes the divergence between label constructed by Score-Based Distribution Estimation and predicted score distributions. We combine these two loss terms into a unified objective, which we denote as the aesthetic score loss L_{ASL} (Eq. 7).

$$\mathcal{L}_{kl} = \sum_i p_i \log(\frac{p_i^{pred}}{p_i}), \tag{6}$$

$$\mathcal{L}_{ASL} = \mathcal{L}_{ce} + k\mathcal{L}_{kl}. \tag{7}$$

Training pipeline. In the description task, the vocabulary exhibits high diversity, far exceeding the predefined categories used for score regression. This may confuse the model when predicting the level tokens (You et al. 2025). To address this issue, we propose a progressive training strategy that effectively leverages the auxiliary description task to improve score prediction performance. Specifically, before training directly on the primary task, we design a multi-task aesthetic training (MAT) stage to capture the aesthetic characteristics of the artistic image. As formulated in Eq. 8, this stage jointly optimizes both the auxiliary and primary objectives. By incorporating detailed aesthetic analysis, MAT enables the model to learn detailed artistic features more effectively, thereby providing a stronger initialization for artistic image aesthetics assessment.

$$\mathcal{L}_{MAT} = \mathcal{L}_{CE} + k_{ASL}\mathcal{L}_{ASL}. \tag{8}$$

To fully unlock the model's scoring potential, in the second stage, we only train with $\mathcal{L}_{ASL}$. This stage ensures the good representations learned in the MAT stage are stimulated and converge to the final score prediction task.

Theoretical Analysis

To explore why generating hierarchical descriptive text can improve the model's aesthetics assessment performance, we present the theoretical analysis underlying our auxiliary description learning framework. The analysis establishes formal guarantees for our approach while revealing fundamental relationships between visual representations, textual descriptions, and aesthetic scores.

Formal Problem Setup

Let us consider the artistic image assessment task through an information-theoretic lens. The input space $\mathcal{X}$ consists of visual artworks with their inherent features x_v, while the description space $\mathcal{D}$ contains the textual aesthetic analyses generated through our RAD pipeline. The target space $\mathcal{Y}$ represents discrete quality scores discretized into K levels $\{l_1, ..., l_K\}$. The latent space $\mathcal{Z}$ captures the multimodal representations learned by ArtQuant.

Theoretical Guarantees

Our first fundamental result establishes an upper bound on the conditional entropy of score prediction:

Theorem 1 (Description-Score Dependency Bound)**.** *For any joint distribution $P(\mathcal{D}, \mathcal{Y}, \mathcal{Z})$ in our framework, the following inequality holds:*

$$H(\mathcal{Y}|\mathcal{Z}) \leq H(\mathcal{D}|\mathcal{Z}) + H(\mathcal{Y}|\mathcal{D}, \mathcal{Z}). \tag{9}$$

This bound reveals that the uncertainty in predicting aesthetic scores ($H(\mathcal{Y}|\mathcal{Z})$) depends critically on two factors: the quality of description encoding ($H(\mathcal{D}|\mathcal{Z})$) and the conditional uncertainty of scores given descriptions ($H(\mathcal{Y}|\mathcal{D}, \mathcal{Z})$). The ideal case occurs when textual descriptions $\mathcal{D}$ fully represent aesthetic scores $\mathcal{Y}$, leading to a simplified bound:

Theorem 2 (Conditional Independence Bound)**.** *When $\mathcal{Y} \perp \mathcal{Z}|\mathcal{D}$ (conditional independence), we obtain:*

$$H(\mathcal{Y}|\mathcal{Z}) \leq H(\mathcal{D}|\mathcal{Z}) + H(\mathcal{Y}|\mathcal{D}). \tag{10}$$

In practical scenarios, perfect conditional independence may not hold. We therefore introduce an ϵ-approximate independence condition (Theorem 3) that relaxes the strict requirement while maintaining theoretical guarantees. This leads to our main error propagation result:

Theorem 3 (Error Propagation Bound)**.** *For $\epsilon = \epsilon_{ver} + \epsilon_{gen}$ is small enough, the prediction error satisfies:*

$$H(\mathcal{Y}|\mathcal{Z}) \leq \underbrace{H(\mathcal{Y}|\mathcal{D}) + \epsilon \log |\mathcal{Y}| + H_2(\epsilon)}_{\textit{description sufficiency}} + \underbrace{H(\mathcal{D}|\mathcal{Z})}_{\textit{description generation ability}}. \tag{11}$$

The implementation aspects that control these error terms correspond directly to key components of our framework. The generation error ϵ_{gen} can be ensured by the powerful GPT-4o for generation, while the verification error ϵ_{ver} is managed through score-comment alignment evaluation.

Experimental Validation

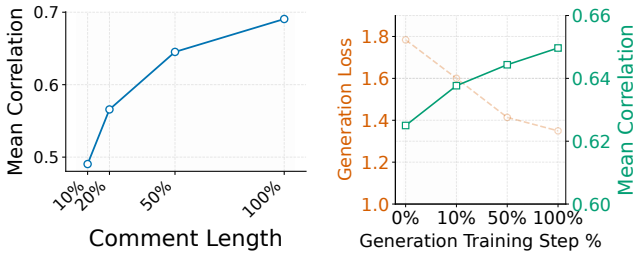


Figure 4: Assessment performance with different description sufficiency and generation capability, where more descriptions and generation capability make the model's assessment more consistent with human.

Our theoretical framework establishes that the model's artistic assessment capability (quantified by $H(\mathcal{Y}|\mathcal{Z})$) is fundamentally bounded by two factors: (1) the quality of generated descriptions ($H(\mathcal{D}|\mathcal{Z})$) and (2) the sufficiency of descriptions for assessment ($H(\mathcal{Y}|\mathcal{D})$). To validate these relationships, we conduct a series of controlled experiments. For the artistic image assessment task, model performance is quantified using the average of PLCC and SROCC metrics. The descriptive capability of the model is evaluated through the generation loss (L_{CE}), while description sufficiency is assessed by analyzing outputs of varying lengths. In the experiment on the relationship between the description ability and performance, we remove L_{ASL} in the MAT stage to prevent interference caused by the participation of the score task training. All the results in the figure are the mean values of the performance on the APDD, BAID, and VAPS datasets.

Model	APDD		BAID		VAPS		Epochs	Modality
	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑		
Generalist MLLMs								
Qwen-VL-Plus (Bai et al. 2023)	0.472	0.482	0.122	0.108	–	–	–	I+T+V
Qwen-VL-Plus (+CoT)	0.472	0.493	0.067	0.062	0.323	0.328	–	I+T+V
GPT-4o (Hurst et al. 2024)	0.531	0.558	0.122	0.108	0.170	0.197	–	I+T+A+V
GPT-4o (+CoT)	0.559	0.603	0.067	0.062	0.217	0.242	–	I+T+A+V
Specialist Models								
SAAN (Yi et al. 2023)	0.780	0.610	0.473	0.467	0.541	0.594	200	I
AANSPS (Jin et al. 2024a)	0.760	0.790	–	–	–	–	–	I
ArtCLIP (Jin et al. 2024b)	0.810	0.840	–	–	–	–	20+	I+T
LITA (Sunada et al. 2024)	–	–	0.490	0.573	–	–	15	I+T
EAMB-Net (Chen et al. 2024b)	–	–	0.496	0.487	0.567	0.628	100	I
AKA-Net (Chen et al. 2024a)	–	–	0.518	0.529	0.579	0.638	40	I
PVAFE (Tang et al. 2025)	–	–	0.533	0.583	–	–	200	I
Our Method								
ArtQuant	**0.871**	**0.894**	**0.543**	**0.589**	**0.625**	**0.681**	4/2/8	I+T

Table 1: Score regression results across datasets (SROCC/PLCC metrics). Modality: I=Image, T=Text, A=Audio, V=Video.

Fig. 4 (left) demonstrates the critical role of description sufficiency. As comment length increases from 10% to 100%, we observe a near-linear improvement in mean correlation scores (from 0.49 to 0.69, $r = 0.92$). This strong positive relationship confirms our theoretical prediction that richer descriptions reduce $H(\mathcal{Y}|\mathcal{D})$, thereby reducing the upper bound on assessment accuracy.

Fig. 4 (right) reveals the dual-axis relationship between description generation ability and model performance. As step size increases from 0 to 100%: Generation loss (orange dashed line) decreases and performance (green solid line) increases from 0.62 to 0.65. The inverse correlation ($r = -0.99$) between generation loss and assessment performance empirically validates $H(\mathcal{D}|\mathcal{Z})$ as an upper bound for $H(\mathcal{Y}|\mathcal{Z})$. This demonstrates that improved description generation directly enables better artistic assessment.

Experiments

Implementation Details

Dataset. To demonstrate the robustness of our method in predicting aesthetic scores, we employ three distinct datasets with significant variations. These datasets span professional expert annotations, crowd-sourced user preferences, and historical artwork assessments by non-specialized raters, collectively representing diverse aesthetic evaluation scenarios across temporal, cultural, and methodological dimensions.

Evaluation metrics. We evaluate AIAA performance using two metrics: Pearson's correlation coefficient (PLCC) and Spearman's rank correlation coefficient (SROCC) to measure prediction-ground truth alignment.

Training setting. For the final model, in the MAT phase, the APDD, BAID, and VAPS are 3, 1, and 1 epoch(s) respectively, while in the SOT phase, the APDD, BAID, and VAPS are 1, 1, and 7 epoch(s) respectively. On all datasets, both k and k_{ASL} are 1. All experiments are conducted on four NVIDIA RTX 4090 GPUs, with total training times of 1 hour (APDD), 1.8 hour (BAID), and 20 minutes (VAPS). Notably, ArtQuant completes training in just 7.2 GPU hours on RTX 4090 for BAID, representing a substantial improvement over SAAN's (Yi et al. 2023) 48-hour training time on RTX 3090 for the same dataset.

Quantitative Experiment

Art image quality assessment. In this section, we compare our model with two generalist MLLMs (GPT-4o (Hurst et al. 2024), Qwen-VL-Plus (Bai et al. 2023)) and several specialist methods across multiple datasets. As shown in Table 1, our model demonstrates superior performance. These experimental results demonstrate that ArtQuant, through pre-training on rich aesthetic comments, has learned fine-grained aesthetic representations, thereby enabling more accurate prediction of aesthetic scores. In terms of convergence efficiency, our method requires few training epochs to converge, significantly reducing computational costs compared to classic methods. This efficiency likely stems from the rich prior knowledge of LLM, particularly advantageous for smaller datasets like VAPS. While GPT-4o (Hurst et al. 2024) and Qwen-VL-Plus (Bai et al. 2023) show that scale and inference strategies (e.g., chain-of-thought) contribute to performance gains on APDD and VAPS, their overall results remain weaker than both our method and classic approaches. This suggests that general MLLMs currently lack specialized aesthetic assessment capabilities. The prediction limitations of general MLLMs on BAID could potentially be attributed to the narrow score distribution of the dataset.

Qualitative Experiment

Our visual analysis (Fig. 5) shows that ArtQuant achieves accurate predictions across varying aesthetic quality levels. Besides, generalist MLLMs tend to overestimate low-scoring artworks (evident in their 4.08 deviation for the

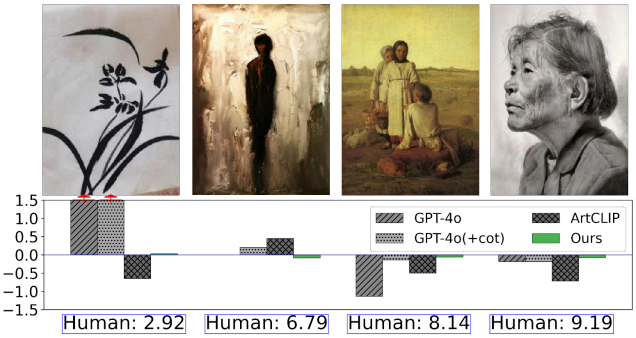


Figure 5: Qualitative results on the APDD test set. The vertical axis represents the error between model predictions and human scores, while the horizontal axis shows artistic images of varying aesthetic quality. Our method achieves better alignment with human judgments than other models across across paintings of different qualities.

	SROCC		PLCC	
Components	Value	Δ_{human}	Value	Δ_{human}
Human	0.837	–	0.864	–
Perception	0.865	+3.35%	0.888	+2.78%
+ Cognition	0.866	+3.46%	0.890	+3.01%
++ Emotion	**0.871**	**+4.06%**	**0.894**	**+3.47%**

Table 2: Progressive hierarchical ablation results with Δ_{human} indicating percentage improvements relative to human baseline performance on APDD dataset.

lowest-scored painting), which may be related to the lack of fine-grained artistic discrimination. The divergence between methods diminishes for high-quality artworks, implying that exceptional aesthetic merit is more universally recognizable across different models.

Ablation Experiments

Annotation method. To validate the benefits of LLM-assisted data generation for score prediction, we compared APDD manual annotations with the APDD subset of the RAD dataset under identical settings. As shown in Table 2, our method achieves significantly higher alignment with human judgments compared to simple and ambiguous manual annotations, with a 4.06% improvement in SROCC (0.871 vs. 0.837) and 3.47% gain in PLCC (0.894 vs. 0.864). The progressive performance improvement across perception (SROCC: +3.35%), cognition (SROCC: +3.46%), and emotion components (SROCC: +4.06%) confirms that structured, multi-level annotations better capture the nuances of image aesthetics. The results suggest that LLM-generated descriptions provide superior supervision for learning aesthetic representations.

Progressive learning. To analyze the role of MAT, we conduct comprehensive evaluations across three distinct artwork. As shown in Table 3, the inclusion of the MAT training stage consistently improves performance across all

Dataset	w/o MAT	w/ MAT	Gain (%)
APDD	0.863/0.889	0.871/0.894	+0.9%/+0.5%
BAID	0.499/0.580	0.543/0.589	+8.82%/+1.55%
VAPS	0.545/0.634	0.625/0.681	+14.68%/+7.41%

Table 3: Performance comparison and percentage improvements of whether train with MAT stage.

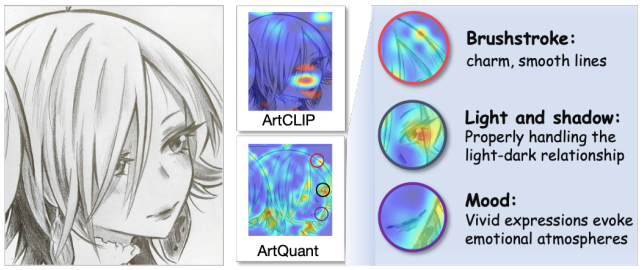


Figure 6: Comparative visualization of vision encoder heatmaps: original painting (left), ArtCLIP (top-center), and our method (bottom-center), with corresponding painting descriptions (right). Our method, enhanced by auxiliary description learning, achieves better semantic alignment in the heatmap representation.

datasets. Such enhancements indicate that establishing initial hierarchical, semantic aesthetic grounding via descriptions effectively bolsters the subsequent capability of score regression.

Feature Analysis

Fig. 6 presents a comparative analysis of vision encoder attention heatmap between ArtCLIP (Jin et al. 2024b) and our proposed model. Our attention map shows significant improvements over ArtCLIP by precisely highlighting contour lines (e.g., hair strands) that match the 'charm, smooth lines' description, demonstrating stroke-aware focus. The attention peaks align with actual shadow regions (nose bridge and neck), validating proper light-dark handling through physically-grounded lighting attention. Moreover, our model exhibits stronger emotional saliency with focused activation on expressive features (eyes and mouth), effectively capturing the annotated emotional atmospheres.

Conclusions

We designs a scalable, iterative framework to generate hierarchical aesthetic descriptions that maintain strong alignment with human scoring. Based on the dataset, we utilizes Multi-Task Aesthetic Training to systematically decompose aesthetic elements across different levels, while incorporating high-precision Score-Based Distribution Estimation to ensure superior correlation with human artistic assessment. Extensive experiments demonstrate that ArtQuant achieves state-of-the-art performance across multiple benchmarks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant U24B6012, 62406167, the Shenzhen Key Laboratory of Next Generation Interactive Media Innovative Technology, China (No. ZDSYS20210623092001004), and the National Science and Technology Council, Taiwan (No. 114-2221-E-006-114-MY3).

References

Amirshahi, S. A.; and Denzler, J. 2017. Judging Aesthetic Quality in Paintings Based on Artistic Inspired Color Features. *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 1–8.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; Du, Q.; Fu, Z.; et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Bin, Y.; Shi, W.; Ding, Y.; Hu, Z.; Wang, Z.; Yang, Y.; Ng, S.-K.; and Shen, H. T. 2024. Gallerygpt: Analyzing paintings with large multimodal models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7734–7743.

Chatterjee, A. 2003. Prospects for a cognitive neuroscience of visual aesthetics.

Chen, H.; Shao, F.; Chai, X.; Mu, B.; and Jiang, Q. 2024a. Art Comes from Life: Artistic Image Aesthetics Assessment via Attribute Knowledge Amalgamation. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Chen, H.; Shao, F.; Mu, B.; and Jiang, Q. 2024b. Image Aesthetics Assessment With Emotion-Aware Multibranch Network. *IEEE Transactions on Instrumentation and Measurement*, 73: 1–15.

Huang, N.; Dong, W.; Zhang, Y.; Tang, F.; Li, R.; Ma, C.; Li, X.; Lee, T.-Y.; and Xu, C. 2025a. Creativesynth: Cross-art-attention for artistic image synthesis with multimodal diffusion. *IEEE Transactions on Visualization and Computer Graphics*.

Huang, N.; Huang, K.; Pu, Y.; Wang, J.; Guo, J.; Yan, Y.; Li, X.; and Lee, T.-Y. 2025b. Artcrafter: Text-image aligning style transfer via embedding reframing. *arXiv preprint arXiv:2501.02064*.

Huang, N.; Liu, H.; Lin, Y.; Huang, K.; Chen, C.; Guo, J.; Lee, T.-y.; and Li, X. 2025c. MaTe: Images Are All You Need for Material Transfer via Diffusion Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15117–15126.

Huang, N.; Zhang, Y.; Tang, F.; Ma, C.; Huang, H.; Dong, W.; and Xu, C. 2024a. Diffstyler: Controllable dual diffusion for text-driven image stylization. *IEEE Transactions on Neural Networks and Learning Systems*.

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

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jin, X.; Qiao, Q.; Lu, Y.; Wang, H.; Gao, S.; Huang, H.; and Li, G. 2024a. Paintings and drawings aesthetics assessment with rich attributes for various artistic categories. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 7672–7680.

Jin, X.; Qiao, Q.; Lu, Y.; Wang, H.; Huang, H.; Gao, S.; Liu, J.; and Li, R. 2024b. APDDv2: Aesthetics of Paintings and Drawings Dataset with Artist Labeled Scores and Comments. *Advances in Neural Information Processing Systems*, 37: 103064–103075.

Leder, H.; Belke, B.; Oeberst, A.; and Augustin, D. 2004. A model of aesthetic appreciation and aesthetic judgments. *British journal of psychology*, 95(4): 489–508.

Leder, H.; and Nadal, M. 2014. Ten years of a model of aesthetic appreciation and aesthetic judgments : The aesthetic episode – Developments and challenges in empirical aesthetics. *British Journal of Psychology*, 105.

Li, C.; and Chen, T. 2009. Aesthetic visual quality assessment of paintings. *IEEE Journal of selected topics in Signal Processing*, 3(2): 236–252.

Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; et al. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Sunada, T.; Shiohara, K.; Xiao, L.; and Yamasaki, T. 2024. LITA: LMM-Guided Image-Text Alignment for Art Assessment. In *International Conference on Multimedia Modeling*, 268–281. Springer.

Tan, Z.; Li, D.; Wang, S.; Beigi, A.; Jiang, B.; Bhattacharjee, A.; Karami, M.; Li, J.; Cheng, L.; and Liu, H. 2024. Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*.

Tang, H.; Chen, Y.; Liang, X.; Chen, L.; Huang, P.; and Tang, Z. 2025. Artistic Image Aesthetics Assessment Assisted by Photographic Visual Attributes. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2024. Q-Align: Teaching LMMs for Visual Scoring via Discrete

Text-Defined Levels. In *International Conference on Machine Learning*, 54015–54029. PMLR.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1): 1.

Yi, R.; Tian, H.; Gu, Z.; Lai, Y.-K.; and Rosin, P. L. 2023. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22388–22397.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14483–14494.

Zhang, B.; Zhang, P.; Dong, X.; Zang, Y.; and Wang, J. 2024. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, 310–325. Springer.