KidsArtBench: Multi-Dimensional Children's Art Evaluation with Attribute-Aware MLLMs

Mingrui Ye[1], Chanjin Zheng[2], Zengyi Yu[2], Chenyu Xiang[3], Zhixue Zhao[3], Zheng Yuan[3], Helen Yannakoudakis[1]

[1]King's College London [2]East China Normal University [3]University of Sheffield

{mingrui.ye, helen.yannakoudakis}@kcl.ac.uk, chjzheng@dep.ecnu.edu.cn, 51284118014@stu.ecnu.edu.cn, {cxiang10, zhixue.zhao, zheng.yuan1}@sheffield.ac.uk

Abstract

Multimodal Large Language Models (MLLMs) show progress across many visual–language tasks; however, their capacity to evaluate artistic expression remains limited: aesthetic concepts are inherently abstract and open-ended, and multimodal artwork annotations are scarce. We introduce KidsArtBench, a new benchmark of over 1k children's artworks (ages 5-15) annotated by 12 expert educators across 9 rubric-aligned dimensions, together with expert comments for feedback. Unlike prior aesthetic datasets that provide single scalar scores on adult imagery, KidsArtBench targets children's artwork and pairs multi-dimensional annotations with comment supervision to enable both ordinal assessment and formative feedback. Building on this resource, we propose an attribute-specific multi-LoRA approach – where each attribute corresponds to a distinct evaluation dimension (e.g., Realism, Imagination) in the scoring rubric – with Regression-Aware Fine-Tuning (RAFT) to align predictions with ordinal scales. On Qwen2.5-VL-7B, our method increases correlation from 0.468 to 0.653, with the largest gains on perceptual dimensions and narrowed gaps on higher-order attributes. Our results show that educator-aligned supervision and attribute-aware training yield pedagogically meaningful evaluations and establish a rigorous testbed for sustained progress in educational AI. We release data and code with ethics documentation.[1]

Figure 1: Sample from the KidsArtBench dataset. Each artwork includes 9 rubric-based scores and (in a subset) educator comments.

1 Introduction

Multimodal Large Language Models (MLLMs) have shown impressive capabilities across visual-language tasks such as captioning, reasoning, and instruction following (Achiam et al., 2023; Team et al., 2024; Biswas and Talukdar, 2024). In education, MLLMs offer the potential to transform assessment and feedback workflows across modalities (Xing et al., 2024; Lee et al., 2024), enhancing accessibility, scalability, and personalization.

Among educational tasks, evaluating student artwork remains especially challenging. Artistic expression is inherently abstract and subjective (Zhu et al., 2021). This issue is further compounded by the scarcity of human-annotated multimodal artwork data, limiting MLLMs' ability to accurately perceive and assess artworks (Huang et al., 2024b). Furthermore, traditional evaluation methods often rely on labor-intensive, inconsistent human judgment (Salı et al., 2014; Meyer et al., 2024). While recent research has explored deep learning approaches to aesthetics (Jiang et al., 2024; She et al., 2021), these models tend to collapse creativity into a single scalar score, lack fine-grained interpretability, and alignment with educational or pedagogical goals (Huang et al., 2024b).

However, structured visual evaluation – particularly in art education – plays a critical role in fostering self-expression, technical skill, and creative development (Denac et al., 2014; Robson and

[1]https://github.com/bigrayss/KidsArtBench

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics Volume 1: Long Papers, pages 5702–5722

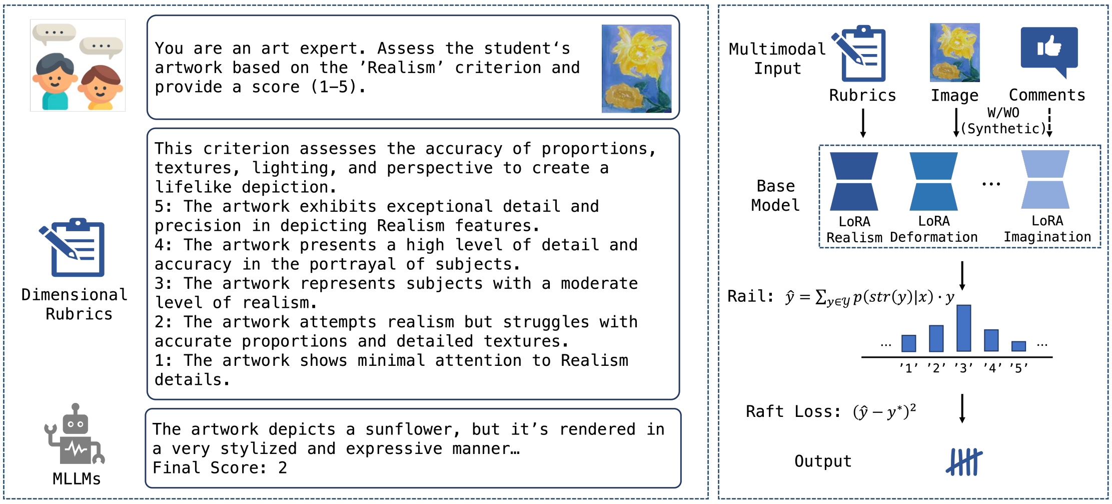

Figure 2: Overview of our framework. Left: prompting MLLMs with rubric-aligned instructions. Right: multi-LoRA architecture with RAFT/RAIL for score prediction.

Rowe, 2012; Seo et al., 2022; Zhao et al., 2024). Effective feedback not only assesses surface-level features (e.g., color use, line structure) but also supports higher-order reflection on originality, transformation, and composition. This calls for models capable of nuanced, dimension-specific, educator-aligned evaluation.

Our contributions. We present KidsArtBench, a new benchmark for multi-dimensional evaluation of children's artwork (ages 5–15), comprising 1,046 student submissions scored by 12 experienced art educators across 9 standardized rubric dimensions (e.g., Realism, Color Richness, Imagination). We refer to these dimensions as *attributes* throughout the paper, aligning with the terminology of multi-attribute modeling in vision–language learning. In contrast to prior datasets that provide only scalar ratings or aesthetics labels [e.g., AVA (Murray et al., 2012), AADB (Kong et al., 2016), ACPP (Jiang et al., 2024)], KidsArtBench captures a richer educational perspective, with both quantitative scores and qualitative feedback comments. Our work is the first to curate a dataset for art assessment that provides comprehensive point-wise evaluations across multiple dimensions together with pedagogically oriented feedback, as illustrated in Figure 1.

To evaluate and improve MLLMs on this task, we first prompt open-source MLLMs using rubric-aligned instructions and observe consistent underperformance, especially on abstract or compositional dimensions (e.g., Line Combination, Imagination). We then propose an attribute-aware fine-tuning framework that incorporates: a multi-branch LoRA architecture (Wang et al., 2023), where each module specializes in a distinct rubric dimension; Regression-Aware Fine-Tuning (RAFT) (Lukasik et al., 2025); and Regression-Aware Inference (RAIL) (Lukasik et al., 2024), aligning predictions with ordinal scales by performing logits-level learning and applying minimum Bayes risk decoding objective. Figure 2 illustrates our full pipeline, from prompt-based rubric alignment to attribute-specific adapters and score decoding.

On Qwen2.5-VL-7B, our approach improves average correlation from 0.468 (prompted baseline) to 0.653, with substantial gains on both perceptual dimensions (e.g., Realism, Color Richness) and abstract attributes (e.g., Transformation, Picture Organization). Nonetheless, certain dimensions (e.g., Line Texture) remain challenging, underscoring the complexity of structured artistic evaluation – and highlighting the value of KidsArtBench as a challenging testbed for advancing fine-grained visual understanding in MLLMs.

In summary, our contributions are fourfold: i) We introduce KidsArtBench, the first public benchmark for multi-dimensional evaluation of children's artwork with expert-annotated rubric scores and comments; ii) We propose an attribute-aware MLLM fine-tuning method using multi-LoRA with RAFT, and RAIL to align predictions with ordinal supervision; iii) We show that our method substantially improves MLLM performance on aesthetic assessment tasks, in particular in perceptual dimensions such as Realism and Transformation; iv) We release all data and code to support further research in educationally grounded multimodal AI.

2 Related Work

Multimodal Large Language Models (MLLMs) Recent commercial MLLMs such as GPT-series, Gemini, and Claude (Achiam et al., 2023; Team et al., 2024; Biswas and Talukdar, 2024), as well as open-source models like Qwen2.5-VL (Bai et al., 2025), Qwen3-VL (Team, 2025), Gemma3 (Team et al., 2025a), Kimi-VL (Team et al., 2025b), and Mimo-VL (Xiaomi, 2025) have achieved strong performance across visual-language tasks including captioning, visual question answering, and cross-modal retrieval (Zhang et al., 2024). Qwen2.5-VL, for example, incorporates native spatiotemporal processing and efficient windowed ViTs, while Gemma3 employs local-to-global attention for long-context reasoning. Despite these advances, MLLMs still struggle with fine-grained, multi-attribute evaluation tasks, especially in domains involving abstract or perceptual qualities (Li et al., 2025; Anis et al., 2025). They exhibit limited spatial reasoning, inconsistent attention to visual detail, and poor alignment with structured rubrics – posing major obstacles for use in art assessment and educational settings.

Aesthetic Evaluation and Understanding Aesthetic evaluation models such as AesCLIP (Sheng et al., 2023) and AesExpert (Huang et al., 2024a) leverage CLIP-like architectures or LLAVA-style prompting for zero-shot scoring or aesthetic question answering. While effective in general image aesthetics, these systems are optimized for scalar preference prediction or aesthetic Q&A rather than pedagogically grounded, dimension-specific feedback. Moreover, their training data often reflect domain biases (e.g., adult photography or generative art), limiting their applicability to children's work. SemArt (Garcia and Vogiatzis, 2018) and ACPP (Jiang et al., 2024) begin to address semantic and child-centered art understanding, respectively. However, they do not provide rubric-aligned, multi-dimensional annotations or comment-level feedback. Existing methods typically collapse artistic merit into a single score, limiting interpretability and educational relevance.

MLLMs in Educational Assessment LLM-based tools have begun to support automated assessment in text-based education (Bewersdorff et al., 2023; Latif and Zhai, 2024), but multimodal assessment remains underexplored. The ArtMentor framework (Zheng et al., 2025) represents an important step: it uses GPT-4o to generate formative feedback for children's artwork using teacher-in-the-loop evaluation. However, its reliance on proprietary models has raised concerns around transparency, cost, and replicability (Yan et al., 2024). In contrast, our work uses open-source MLLMs and contributes both a dataset and fine-tuning framework for rubric-based, pedagogically meaningful visual assessment.

Aesthetic and Educational Art Datasets Many aesthetics datasets exist; e.g., AVA (Murray et al., 2012), AADB (Kong et al., 2016), PCCD (Chang et al., 2017), OmniArt (Strezoski and Worring, 2018), and Art500k (Mao et al., 2017). However, most focus on adult art or photography and rely on aggregate preference ratings (e.g., votes or likes), without explicit rubrics. ACPP (Jiang et al., 2024) is one of the few datasets focused on children's art, containing scalar scores across eight attributes. However, it lacks rubric calibration, feedback comments, and modeling benchmarks. Our proposed dataset, uniquely provides expert-annotated, nine-dimensional scores with a rubric-aligned evaluation protocol and a modeling suite designed for both assessment and feedback generation in educational settings. To our knowledge, KidsArtBench is the first dataset to combine expert rubric supervision, multi-dimensional annotation, and open benchmarking protocols for MLLMs in educational art evaluation.

3 KidsArtBench Dataset

We collected 1,046 original artworks from students aged 5–15 through an online submission platform designed to encourage authentic creative expression. All data collection followed an approved IRB protocol with parental consent and child assent (see Ethics Statement and Bias section). The dataset primarily consists of artworks collected from children across multiple primary and middle schools in Eastern China. Educator feedback comments were originally written in Chinese; we translated them into English using Google Translate and then manually verified and corrected them with the help of proficient English speakers who are native Chinese speakers. We release both the original Chinese comments and their English translations to support a broad range of research communities.

Each artwork was independently scored by at least two trained art educators across nine standardized rubric dimensions (Table 1) on a scale from

Category	Dimension	Evaluation Criteria
Formative Creativity	Realism (Biswas, 2021) Deformation (Sfarra et al., 2014) Imagination (Searle and Shulha, 2016)	Accuracy in depicting subjects and objects Creative reinterpretation of reality Novelty and originality of concepts
Color Expressiveness	Color Richness (Lu et al., 2015; Pylypchuk et al., 2021) Color Contrast (Zhang et al., 2021)	Diversity and harmony of color palette Visual impact through hue interactions
Line Work Richness	Line Combination (Locher et al., 1999) Line Texture (Ding et al., 2020)	Structural arrangement of strokes Expressive tactile quality of linework
Conceptual Thinking	Picture Organization (Locher et al., 1999) Transformation (Du, 2020)	Balanced composition and spatial logic Effective rendering of abstract concepts

Table 1: KidsArtBench rubric dimensions. Full rubric text in Appendix A.1.

1 to 5. A senior expert then adjudicated discrepancies, consulting original raters when necessary to assign a final score for each dimension. This multi-stage process ensured high-quality, educator-aligned annotations. Our dataset also contains expert-written formative comments, providing qualitative guidance aligned with the rubric (Figure 1).

The dataset is split into 80% train, 10% validation, and 10% test to support model development. Figure 3 shows the score distributions across all dimensions: most scores cluster around 3-4, reflecting realistic assessment tendencies in educational contexts (in our dataset, most student artworks typically demonstrate intermediate to above-average performance), with relatively few extreme values. The rubric captures nuanced variation in artistic quality and provides a representative basis for training and evaluating MLLMs in art education.

4 Methodology

Given an artwork image (and optional comment text), the goal is to predict nine ordinal scores $y_m \in \{1, \ldots, 5\}$ for each rubric dimension $m \in \mathcal{D}$ where $\mathcal{D} = \{\text{realism}, ..., \text{transformation}\}$. This enables dimension-specific assessment rather than a single scalar aesthetic score.

4.1 Attribute-Aware Prompting with MLLMs

We explore open-source MLLMs for art evaluation using rubric-guided prompting. Each artwork is assessed using a structured prompt template shown in Figure 2 (left). Full rubric definitions and prompt examples are provided in Appendix A.1-A.2. Our attribute-aware prompting strategy enables dimension-specific queries, encouraging the model to reason about aspects such as Realism, or Color Richness. However, our results show that existing MLLMs perform inconsistently on KidsArtBench, especially on abstract and compositional dimensions (e.g., Imagination, Line Combination), motivating the need for fine-tuning with targeted supervision.

4.2 Attribute-Specific MultiLoRA

We propose an attribute-aware MLLM fine-tuning framework, in which each evaluation attribute – corresponding to a rubric dimension such as *Realism* or *Imagination* – is modeled by a dedicated LoRA adapter. This modular design decomposes the multi-dimensional assessment task into independent, specialized branches, mitigating inter-dimensional interference and enabling fine-grained learning across all criteria (Wang et al., 2023; Hu et al., 2022).

Let x denote the input, which includes the student artwork and optional textual description, and let $\mathbf{z} = f_0(x) \in \mathbb{R}^d$ be the corresponding embedding from a frozen backbone encoder f_0. We define the set of evaluation dimensions as $\mathcal{D}$. For each dimension $m \in \mathcal{D}$, we attach a LoRA adapter to generate a scalar prediction $\hat{y}_m$:

$$\hat{y}_m = (\mathbf{W}_0 + \mathbf{B}_m \mathbf{A}_m)\, \mathbf{z} + b_m, \tag{1}$$

where $\mathbf{W}_0 \in \mathbb{R}^{d \times d}$ is the frozen, shared projection matrix; $\mathbf{A}_m \in \mathbb{R}^{r \times d}$ and $\mathbf{B}_m \in \mathbb{R}^{d \times r}$ are trainable low-rank update matrices (LoRA rank $r \ll d$); $b_m \in \mathbb{R}$ is a dimension-specific learnable bias.

This formulation allows each adapter to specialize in one rubric criterion while leveraging the shared visual-linguistic representation $\mathbf{z}$. Importantly, the architecture supports flexible composition: any subset of dimensions $\mathcal{S} \subseteq \mathcal{D}$ can be selectively activated for context-specific evaluation or targeted feedback.

4.3 Regression-Aware Fine-Tuning (RAFT) and Inference (RAIL)

To produce well-calibrated ordinal scores, we adopt the Regression-Aware Inference for Language models (RAIL) framework (Lukasik et al., 2024), which seeks predictions that minimize expected error under a minimum Bayes-risk decoding objective. For

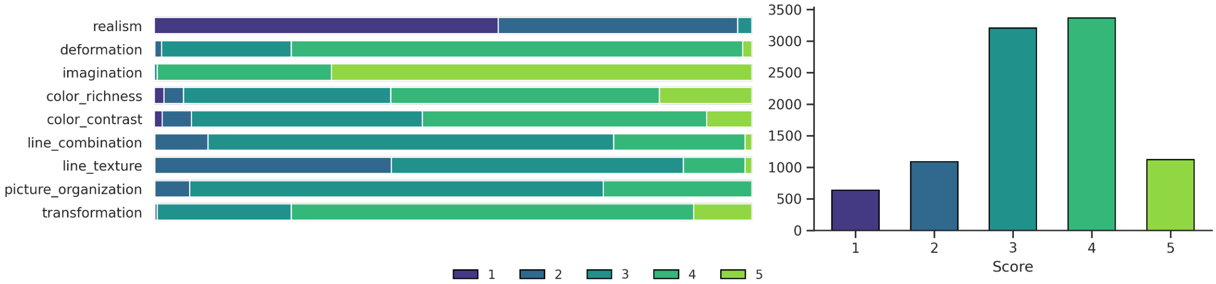

Figure 3: Score distribution across nine dimensions (left) and overall dataset (right) in KidsArtBench.

our discrete target space $\mathcal{Y} = \{1, 2, 3, 4, 5\}$, the expected value can be computed exactly by scoring all candidates and taking their probability-weighted average:

$$\hat{y}_m(x) = \sum_{y \in \mathcal{Y}} p(\text{str}(y) \mid x)\, y, \quad (2)$$

where $\text{str}(\cdot)$ denotes the string representation of each candidate extracted from the MLLM output logits. This decision rule yields smoother, more consistent ordinal predictions than simple argmax decoding.

To align training with this inference procedure, we integrate the decision rule directly into fine-tuning via Regression-Aware Fine-Tuning (RAFT) (Lukasik et al., 2025). Specifically, each dimension-specific adapter is optimized using mean squared error (MSE) between predicted and gold scores:

$$\mathcal{L}_m = \frac{1}{N} \sum_{j=1}^{N} \left(\hat{y}_m^{(j)} - y_m^{*(j)}\right)^2, \quad (3)$$

where $\{y_m^{*(j)}\}_{j=1}^{N}$ are the ground-truth scores for dimension m. This directly optimizes for regression performance while preserving the discrete nature of the targets. Combined with our multi-LoRA architecture, RAFT/RAIL produces a modular and interpretable system for fine-grained art assessment (Figure 2, right). By isolating updates within each LoRA adapter, we further reduce inter-dimensional interference and enhance specialization. Moreover, the architecture allows flexible composition: any subset of adapters $\mathcal{S} \subseteq \mathcal{D}$ can be selectively activated for context-specific evaluation or targeted feedback, mirroring how human educators emphasize different rubric criteria in different contexts.

5 Results

5.1 Prompting-Based Evaluation Results

We evaluate 8 open-source MLLMs on the KidsArtBench dataset using rubric-aligned prompting: Qwen3-VL-30B-A3B (Oct. 2025), Qwen2.5-VL-7B / 32B / 72B (Apr. 2025), Gemma3-12B / 27B (Apr. 2025), Mimo-VL-7B (June 2025), and Kimi-VL-A3B (June 2025). Performance is assessed across five metrics: Spearman's rank correlation (SC), Pearson's correlation (PC), exact match accuracy (ACC), mean squared error (MSE), and quadratic weighted kappa (QWK). These collectively capture both ordinal and regression quality, as well as model-human agreement beyond chance.

A comparative performance summary of attribute-aware prompting across all models is shown in Figure 4, with detailed results in Tables 12-21. Among all models, the Qwen-VL family demonstrates the strongest performance. Qwen2.5-VL-72B achieves the best overall prompting results (SC = 0.487, PC = 0.492, QWK = 0.376), particularly in dimensions such as Realism (ACC = 0.76) and Picture Organization (QWK = 0.510). Interestingly, the newer Qwen3-VL-30B-A3B, while stronger on general VQA benchmarks, underperforms Qwen2.5-VL-7B in our task (ΔSC = –0.105, ΔQWK = –0.078), possibly due to inconsistent expert routing in its Mixture-of-Experts (MoE) architecture. Gemma3-27B shows strong performance on abstract dimensions such as Imagination (QWK = 0.557) and Color Contrast (QWK = 0.686), but struggles with others such as Deformation (SC = 0.118). Mimo-VL-7B yields a balanced overall profile, close to Qwen2.5-VL-7B. In contrast, Kimi-VL-A3B lags across all metrics, with low scores on Realism (ACC = 0.02) and an overall SC of just 0.215.

Across dimensions, Color Richness emerges as the most consistently well-predicted attribute: Qwen2.5-VL-72B achieves SC = 0.651 and QWK

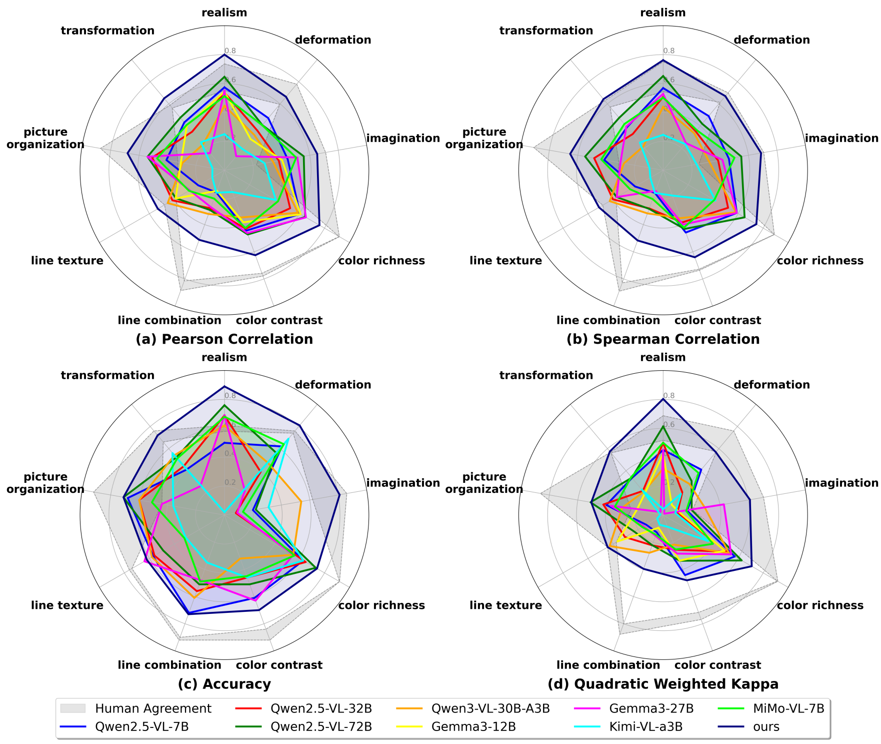

Figure 4: Radar plot showing per-dimension performance of MLLMs on KidsArtBench. We also compare them with our attribute-aware fine-tuned model, and human rater agreement. Enlarged version is provided in Appendix A.7.

= 0.628, while Gemma3-27B slightly surpasses this with SC = 0.655. Realism is also reliably captured (SC ≈ 0.635-0.652 for Qwen models). However, more complex and abstract dimensions – Imagination, Line Combination, Line Texture – remain persistently challenging: performance on Line Combination, for instance, rarely exceeds SC = 0.28 or QWK = 0.22. These results indicate a key limitation of prompt-based approaches: MLLMs tend to perform better on surface-level, perceptual attributes (e.g., color, realism) but struggle with evaluative dimensions that require abstraction, compositional reasoning, or creative interpretation.

Prompting Strategies We conduct ablation experiments comparing two alternative prompting setups: minimal prompting and few-shot prompting, with prompting examples and results shown in Appendix A.2 Tables 22, 23. In minimal prompting, models are asked to assess each artwork based on a given dimension using a short textual instruction, without including any rubric definitions. In contrast, the few-shot prompting setup provides two exemplar image–text pairs from the training set, one low-quality (score 1) and one high-quality (score 5), each accompanied by a brief dimension-specific description to guide the model's inference. Across all dimensions, both the minimal and few-shot strategies consistently underperformed relative to the atribute-aware prompting approach. This suggests that structured rubrics offer essential guidance for model reasoning in aesthetic evaluation, while exemplar-based few-shot learning provides only limited benefit. These findings further underscore that current MLLMs, even when aided with prompting strategies, remain insufficient to achieve human-level reliability, particularly for abstract or higher-order dimensions of artistic assessment.

Human Agreement To contextualize these results, two in-service art teachers (over 2 years experience) re-annotated the test set. The gray band in Figure 4 shows inter-rater agreement. Even the best model, Qwen2.5-VL-72B, falls short of human-level performance: ΔQWK = –0.188 on average, with the largest discrepancies in Line Combination (ΔQWK = –0.618) and Color Richness (ΔQWK = –0.287). Teachers show especially high agreement in color and line-based dimensions – areas where MLLMs fall short. Taken together, these results demonstrate that prompting alone is insufficient for high-fidelity, human-aligned art evaluation, particularly for abstract, higher-order rubric dimensions. This motivates the need for fine-tuning approaches that better encode pedagogically relevant criteria.

Metrics	Methods	Realism	Deformation	Imagination	Color Richness	Color Contrast	Line Combination	Line Texture	Picture Organization	Transformation	Average
SC ↑	Prompting	0.569	0.490	0.462	0.589	0.458	0.167	0.224	0.414	0.389	0.418
	LoRA	0.457	0.524	0.581	0.615	0.517	0.254	0.285	0.583	0.374	0.466
	Multi-LoRA	0.618	0.553	0.584	0.662	0.573	0.321	0.335	0.469	0.448	0.507
	LoRA+RAFT	0.545	0.465	0.567	0.700	0.650	0.367	0.381	0.466	0.523	0.518
	Multi-LoRA+RAFT	**0.762**	0.670	0.689	0.745	0.640	0.517	**0.513**	0.654	**0.643**	**0.648**
	Human Upbound	0.751	**0.700**	**0.707**	**0.889**	**0.736**	**0.899**	0.483	**0.910**	0.629	0.600
PC ↑	Prompting	0.574	0.472	0.445	0.604	0.444	0.163	0.209	0.408	0.433	0.417
	LoRA	0.459	0.518	0.573	0.617	0.519	0.245	0.278	0.612	0.377	0.466
	Multi-LoRA	0.618	0.514	0.577	0.661	0.600	0.323	0.344	0.486	0.456	0.508
	LoRA+RAFT	0.541	0.447	0.569	0.694	0.625	0.356	0.395	0.476	0.519	0.514
	Multi-LoRA+RAFT	**0.800**	0.664	0.652	0.760	0.625	0.512	**0.533**	0.682	**0.650**	**0.653**
	Human Upbound	0.737	**0.778**	**0.711**	**0.919**	**0.779**	**0.885**	0.479	**0.872**	0.587	0.575
ACC ↑	Prompting	0.500	0.620	0.200	0.630	0.610	0.720	0.570	0.680	0.410	0.549
	LoRA	0.700	0.770	0.790	0.660	0.690	0.650	0.500	0.770	0.630	0.684
	Multi-LoRA	0.760	**0.810**	0.790	0.670	0.710	0.720	0.540	0.710	0.640	0.706
	LoRA+RAFT	0.750	0.720	0.810	0.710	0.750	0.760	0.590	0.720	0.720	0.726
	Multi-LoRA+RAFT	**0.890**	**0.810**	0.810	0.740	0.700	0.730	0.620	0.710	0.720	0.748
	Human Upbound	0.620	0.760	**0.860**	**0.920**	**0.920**	**0.910**	**0.760**	**0.920**	**0.760**	**0.771**
MSE ↓	Prompting	0.560	0.410	1.520	0.400	0.390	0.280	0.430	0.320	0.710	0.558
	LoRA	0.330	0.230	0.210	0.370	0.340	0.350	0.530	0.230	0.370	0.329
	Multi-LoRA	0.240	0.220	0.210	0.330	0.290	0.280	0.490	0.290	0.360	0.301
	LoRA+RAFT	0.250	0.310	0.190	0.290	0.250	0.240	0.410	0.280	0.280	0.278
	Multi-LoRA+RAFT	**0.106**	**0.153**	0.170	0.201	0.227	0.199	**0.247**	0.178	**0.237**	**0.191**
	Human Upbound	0.360	0.240	**0.140**	**0.080**	**0.140**	**0.100**	0.260	**0.080**	0.360	0.272
QWK ↑	Prompting	0.451	0.411	0.169	0.573	0.442	0.128	0.134	0.399	0.335	0.338
	LoRA	0.450	0.511	0.610	0.615	0.519	0.238	0.270	0.610	0.359	0.461
	Multi-LoRA	0.613	0.505	0.574	0.654	0.600	0.310	0.332	0.481	0.430	0.499
	LoRA+RAFT	0.537	0.429	0.558	0.693	0.610	0.330	0.385	0.428	0.497	0.496
	Multi-LoRA+RAFT	**0.804**	0.568	0.573	0.708	0.480	0.396	**0.444**	0.505	**0.574**	**0.566**
	Human Upbound	0.688	**0.760**	**0.706**	**0.917**	**0.769**	**0.878**	**0.444**	**0.863**	0.567	0.564

Table 2: Performance of Qwen2.5-VL-7B across different training configurations. Best and second-best scores are highlighted in **bold** and underlined, respectively. Human Upper Bound reports the better score from two human annotators per dimension. The Average cell in that row reflects the mean performance across the two human raters.

5.2 Attribute-Aware Fine-Tuning

To balance performance and efficiency, we select Qwen2.5-VL-7B as the base model for fine-tuning, following the methodology described in Section 4.2. To assess the effectiveness of our proposed attribute-aware multi-LoRA architecture with RAFT, we conduct a series of comparative experiments, summarized in Table 2. Specifically, we evaluate the contributions of (i) the multi-LoRA design, by comparing it against a shared-LoRA configuration, and (ii) RAFT, by comparing it against a standard regression head trained with MSE loss.

The shared-LoRA + standard regression baseline achieves moderate overall performance (SC = 0.466, QWK = 0.461). While it performs relatively well on dimensions such as Imagination (SC = 0.581) and Picture Organization (SC = 0.583), it struggles on structural attributes such as Line Combination (SC = 0.254) and Line Texture (SC = 0.285), highlighting its limitations in modeling fine-grained compositional features.

Introducing multi-LoRA yields consistent improvements, increasing avg SC to 0.507 and QWK to 0.499, with notable gains in Realism and Color Contrast. Adding RAFT further enhances performance, especially on perceptual dimensions, where Color Richness reaches SC = 0.700 and QWK = 0.693, outperforming standard regression training. The full multi-LoRA + RAFT configuration yields the strongest results overall (SC = 0.648, QWK = 0.566, ACC = 0.748), outperforming all baselines across most dimensions. The largest gains appear in Realism (SC = 0.762, ACC = 0.890), as well as in Imagination and Deformation, indicating the model benefits not only from greater perceptual alignment but also from improved abstraction and compositional reasoning. These results demonstrate that our fine-tuning strategy substantially enhances MLLMs' ability to perform nuanced, multi-dimensional aesthetic assessment.

To contextualize model performance relative to humans, we compare our best model against expert annotations. Interestingly, the model surpasses human-level agreement in select dimensions, including Realism (ΔSC = +0.11, ΔQWK = +0.116) and Line Texture (ΔSC = +0.30). While overall model performance approaches human-level reliability (average ΔSC = +0.014, ΔQWK = +0.007), it still lags behind in stylistic dimensions – particularly those involving line and color – consistent with our earlier prompting analysis (Section 5.1). Full details on ablation settings and hyperparameter sensitivity are in Appendix A.4.

5.3 Comments and Data Augmentation

To examine how expert-written comments contribute to score assessment, we incorporate them as auxiliary linguistic feedback and report results utilising both a 100-sample training subset and the full training dataset. The subset offers a more tightly controlled, low-variance setting, while the full dataset reflects performance under broader, more diverse conditions. Consistent trends across both evaluations suggest the comments effect is robust, even though the aggregated improvement is attenuated on the full dataset. We retain the attribute-specific multi-LoRA model and modify only the input x, which now includes additional comment text alongside the existing rubric-based instructions (prompt example with comments shown in Appendix A.2).

Using the full training set, comment-aware training led to modest but consistent gains on several challenging dimensions, as shown in Table 32. Imagination exhibits substantial improvement, with PC increasing from 0.575 to 0.677. Also, Line Combination shows a comparable positive shift, with PC increasing from 0.325 to 0.445, indicating that comment-based supervision is particularly helpful for dimensions that require higher-level structural reasoning. In contrast, some predominantly perceptual dimensions exhibit small declines – Realism, for instance, shows a PC drop from 0.726 to 0.697 – suggesting that comments (which emphasize semantic or conceptual aspects) are more aligned with higher-level reasoning dimensions than with attributes driven primarily by low-level visual cues.

For the ablation experiments on the 100-sample subset, training without comment supervision yields weak performance (SC of 0.304 and QWK of 0.25). Adding comments as auxiliary input raises the average correlation to 0.355, with clear improvements in dimensions such as Color Contrast (ΔSC = +0.321), Realism (ΔSC = +0.066), and Picture Organization (ΔSC = +0.067). Augmenting the training subset by creating multiple copies of each comment-scored sample, each paired with a different instructional prompt (examples shown in Appendix A.2), leads to an average SC improvement of 0.174. In particular, Realism reaches SC = 0.654 with QWK = 0.485, while Imagination improves to SC = 0.531 (Tables 28–30).

We also test whether visual robustness can be improved through image augmentation, as shown in Table 31. Specifically, moderate color jitter (brightness/contrast/saturation = 0.2, hue = 0.05) yields small robustness gains, whereas stronger distortions reduce performance, suggesting that maintaining perceptual fidelity – alongside curated linguistic feedback and controlled augmentation – better supports the model's reasoning.

5.4 Qualitative Analysis

To better understand how the model internalizes rubric dimensions, we analyze our fine-tuned multi-LoRA architecture using subspace overlapping scores (Ilharco et al., 2023) (Figure 5, left), which quantify the independence of LoRA adapters – where lower overlap reflects greater specialization and higher values indicate shared representations. Notably, Deformation and Transformation show a moderate overlap of 0.328, suggesting they capture similar abstract or conceptual features. Likewise, Line Combination and Line Texture overlap at 0.404, and Color Contrast and Color Richness at 0.329. These pairs fall within shared higher-level aesthetic categories, such as line expressiveness or color dynamics, reflecting how the adapters naturally cluster around semantically related features. To compare this to real dataset distribution, we also compute the inter-dimensional correlation matrix over the annotated scores (Figure 5, right). The correlation patterns align closely with subspace overlap: dimensions that share visual or conceptual grounding (e.g., Deformation–Transformation, Color Contrast–Color Richness) show stronger correlations, while unrelated dimensions (e.g., Realism vs. Imagination) remain more independent.

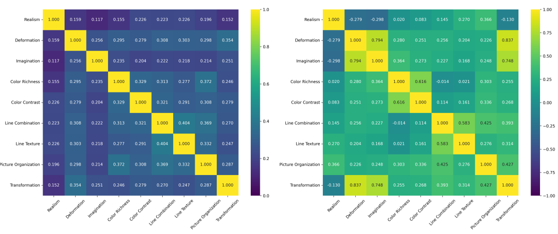

Figure 5: Subspace overlap of multi-LoRA adapters (left) and empirical dimension correlations (right). Enlarged version is provided in Appendix A.7.

Overall, these analyses suggest that the multi-LoRA model captures meaningful structure in the underlying rubric, with distinct yet interpretable groupings that align with educator-defined aesthetic dimensions. Additional qualitative results and more analyses are included in Appendix A.6.

5.5 Error Analysis

To better characterize the limitations of our models, we analyze failure cases by defining error cases as samples in which more than two predicted dimensions do not exactly match the ground-truth scores on the test set. Here, any deviation from the ground truth is treated as an error, although most mispredictions differ by only one score. Lower-grade students' artworks (grades 1–3, mostly ages 6–11) make up roughly 60% of the dataset yet account for 88% of these high-error cases, suggesting that early-grade drawings – often more abstract and structurally irregular – pose greater challenges. Error patterns vary across dimensions: Transformation and Line Texture show the highest error rates (both 61.1%), followed by Line Combination (52.7%), whereas perceptual attributes such as Realism exhibit much lower error rates (13.9%). Despite these discrepancies, prediction deviations typically remain within ±1 of the ground-truth score. Media-type distributions among high-error samples (47.2% marker, 16.7% oil pastel, 19.4% crayon/colored pencil, 11.1% watercolor) largely mirror their overall dataset frequencies, indicating no strong medium-specific effects. This further shows that failures commonly arise in drawings that are highly schematic, contain overlapping or fragmented objects, or employ unconventional spatial layouts.

6 Conclusions

We present KidsArtBench, the first public benchmark for multi-dimensional evaluation of children's artwork using rubric-guided annotations. To model this structured aesthetic assessment task, we introduce an attribute-aware fine-tuning framework combining Multi-LoRA (with one adapter per rubric dimension) and RAFT, aligned with regression-aware inference. Our method significantly outperforms prompting and shared-adapter baselines, achieving strong agreement with expert ratings. Analyses of adapter subspaces and attention patterns confirm that the model learns distinct, interpretable representations for each attribute. Further improvements from comment supervision and instruction-level augmentation highlight the value of linguistic and pedagogical cues. KidsArtBench bridges multimodal AI and education, demonstrating that properly aligned MLLMs can deliver nuanced, rubric-based assessment in creative domains. We release all data, code, and annotations to support future research in educationally grounded multimodal learning.

Limitations

This study has several limitations that suggest directions for future work. Our dataset only contains about 1k samples and focuses primarily on children's artworks from a specific region. Despite many efforts we made for generalization, we fully acknowledge the demographic and cultural limitations of the current dataset. Conducting large-scale cross-cultural data collection lies beyond the scope of the present work, but we view it as an important direction for future work, including consideration of potential extensions such as more systematic cross-regional sampling and explicit cultural comparisons. Additionally, comments are used solely as auxiliary training signals in this paper. Future work will investigate how expert-written comments can be more deeply integrated into learning and inference, for example as intermediate supervision signals in reinforcement learning; structured guidance for multi-step assessment; and as explicit targets for generating pedagogically grounded feedback. In this study, we focus on fine-tuning a 7B-scale model due to computational constraints. As modeling efficiency continues to improve, extending this framework to larger MLLMs remains a promising direction.

Ethics Statement and Bias

KidsArtBench currently contains children artwork (aged 5 to 15) collected from multiple primary and middle schools predominantly located in Eastern China, as well as several summer-camp programs whose participants come from diverse regions across the country. This dataset is under protocol WZU-2025-106 approved by the East China Normal University Institutional Review Board on 20 August 2025, in line with the Personal Information Protection Law of the People's Republic of China (PIPL) where applicable and established academic ethical policies. We obtained parent/guardian consent and child assent; forms specified collected data (artwork image, optional title, statement, age band), research purpose, security, withdrawal rights, and release conditions. We excluded items with identifying marks, identifiable photos, or missing consent. Prior to any access or release we removed direct identifiers, replaced age with bands, manually redacted signatures, assigned non-reversible IDs, and screened statements. The public package contains only de-identified images and labels. Annotations were produced by 12 art educators with more than 5 years' experience after an average 24-hour calibration. The dataset primarily consists of artworks collected from children across multiple primary schools and middle schools within eastern China, which may introduce regional and cultural biases inherent to the local educational context. The dataset may also reflect socioeconomic biases, as participating schools tend to represent regions with relatively higher access to art education resources. The dataset is released for research use only, and is designed to enable research on formative feedback for children's artwork, not for summative assessment, grading, or student placement. We release our code and data (https://github.com/bigrayss/KidsArtBench) under the ACL Code of Ethics and the MIT License with terms forbidding re-identification, high-stakes use, and unauthorised redistribution.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahmad Mustafa Anis, Hasnain Ali, and Saquib Sarfraz. 2025. On the limitations of vision-language models in understanding image transforms. *arXiv preprint arXiv:2503.09837*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Arne Bewersdorff, Kathrin Seßler, Armin Baur, Enkelejda Kasneci, and Claudia Nerdel. 2023. Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *Computers and Education: Artificial Intelligence*, 5:100177.

Anjanava Biswas and Wrick Talukdar. 2024. Robustness of structured data extraction from in-plane rotated documents using multi-modal large language models (llm). *Journal of Artificial Intelligence Research*.

Moinak Biswas. 2021. Realism. *BioScope: South Asian Screen Studies*, 12(1-2):158–161.

Kuang-Yu Chang, Kung-Hung Lu, and Chu-Song Chen. 2017. Aesthetic critiques generation for photos. In *Proceedings of the IEEE international conference on computer vision*, pages 3514–3523.

Olga Denac and 1 others. 2014. The significance and role of aesthetic education in schooling. *Creative education*, 5(19):1714.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. 2020. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581.

Yingbei Du. 2020. Research on the transformation and innovation of visual art design form based on digital fusion technology. *Applied Mathematics and Nonlinear Sciences*.

Noa Garcia and George Vogiatzis. 2018. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. 2024a. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5911–5920.

Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. 2024b. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Shiqi Jiang, Ning Li, Chen Shi, Liping Guo, Changbo Wang, and Chenhui Li. 2024. Aacp: Aesthetics assessment of children's paintings based on self-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2534–2542.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 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*, pages 662–679. Springer.

Ehsan Latif and Xiaoming Zhai. 2024. Fine-tuning chatgpt for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100210.

Unggi Lee, Minji Jeon, Yunseo Lee, Gyuri Byun, Yoorim Son, Jaeyoon Shin, Hongkyu Ko, and Hyeoncheol Kim. 2024. Llava-docent: Instruction tuning with multimodal large language model to support art appreciation education. *Computers and Education: Artificial Intelligence*, 7:100297.

Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. 2025. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *arXiv preprint arXiv:2501.02189*.

Paul J Locher, Pieter Jan Stappers, and Kees Overbeeke. 1999. An empirical evaluation of the visual rightness theory of pictorial composition. *Acta psychologica*, 103(3):261–280.

Peng Lu, Zhijie Kuang, Xujun Peng, and Ruifan Li. 2015. Discovering harmony: A hierarchical colour harmony model for aesthetics assessment. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part III 12*, pages 452–467. Springer.

Michal Lukasik, Zhao Meng, Harikrishna Narasimhan, Yin-Wen Chang, Aditya Krishna Menon, Felix Yu, and Sanjiv Kumar. 2025. Better autoregressive regression with llms via regression-aware fine-tuning. In *The Thirteenth International Conference on Learning Representations*.

Michal Lukasik, Harikrishna Narasimhan, Aditya Krishna Menon, Felix Yu, and Sanjiv Kumar. 2024. Regression aware inference with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13667–13678, Miami, Florida, USA. Association for Computational Linguistics.

Hui Mao, Ming Cheung, and James She. 2017. Deepart: Learning joint representations of visual arts. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1183–1191.

Louie Meyer, Johanne Engel Aaen, Anitamalina Regitse Tranberg, Peter Kun, Matthias Freiberger, Sebastian Risi, and Anders Sundnes Løvlie. 2024. Algorithmic ways of seeing: Using object detection to facilitate art exploration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE.

Oksana Pylypchuk, Andrii Polubok, Olga Krivenko, Olena Safronova, Danylo Kosenko, and Nataliia Avdieieva. 2021. Developing an approach to colour assessment of works of art on aim to creating a comfortable and harmonious interior. In *2021 International Conference on Social Sciences and Big Data Application (ICSSBDA 2021)*, pages 181–187. Atlantis Press.

Sue Robson and Victoria Rowe. 2012. Observing young children's creative thinking: engagement, involvement and persistence. *International Journal of Early Years Education*, 20(4):349–364.

Güneş Salı, Aysel Köksal Akyol, and Gülen Baran. 2014. An analysis of pre-school children's perception of schoolyard through their drawings. *Procedia-Social and Behavioral Sciences*, 116:2105–2114.

Michelle J Searle and Lyn M Shulha. 2016. Capturing the imagination: Arts-informed inquiry as a method in program evaluation. *Canadian Journal of Program Evaluation*, 31(1):34–60.

Woosuk Seo, Joonyoung Jun, Minki Chun, Hyeonhak Jeong, Sungmin Na, Woohyun Cho, Saeri Kim, and Hyunggu Jung. 2022. Toward an ai-assisted assessment tool to support online art therapy practices: A pilot study. In *Proceedings of 20th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET).

S Sfarra, C Ibarra-Castanedo, D Ambrosini, D Paoletti, A Bendada, and X Maldague. 2014. Discovering the defects in paintings using non-destructive testing (ndt) techniques and passing through measurements of deformation. *Journal of Nondestructive Evaluation*, 33:358–383.

Dongyu She, Yu-Kun Lai, Gaoxiong Yi, and Kun Xu. 2021. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8475–8484.

Xiangfei Sheng, Leida Li, Pengfei Chen, Jinjian Wu, Weisheng Dong, Yuzhe Yang, Liwu Xu, Yaqian Li, and Guangming Shi. 2023. Aesclip: Multi-attribute contrastive learning for image aesthetics assessment. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1117–1126.

Gjorgji Strezoski and Marcel Worring. 2018. Omniart: A large-scale artistic benchmark. *ACM Trans. Multimedia Comput. Commun. Appl.*, 14(4).

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. *URL https://arxiv. org/abs/2403.05530*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025a. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025b. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yiming Wang, Yu Lin, Xiaodong Zeng, and Guannan Zhang. 2023. Multilora: Democratizing lora for better multi-task learning. *arXiv preprint arXiv:2311.11501*.

LLM-Core-Team Xiaomi. 2025. Mimo-vl technical report. *Preprint*, arXiv:2506.03569.

Weicheng Xing, Tianqing Zhu, Jenny Wang, and Bo Liu. 2024. A survey on mllms in education: Application and future directions. *Future Internet*, 16(12):1–31.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. MM-LLMs: Recent advances in MultiModal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12401–12430, Bangkok, Thailand. Association for Computational Linguistics.

Jiajing Zhang, Yongwei Miao, and Jinhui Yu. 2021. A comprehensive survey on computational aesthetic evaluation of visual art images: Metrics and challenges. *IEEE Access*, 9:77164–77187.

Liang Zhao, Eslam Hussam, Jin-Taek Seong, Assem Elshenawy, Mustafa Kamal, and Etaf Alshawarbeh. 2024. Revolutionizing art education: Integrating ai and multimedia for enhanced appreciation teaching. *Alexandria Engineering Journal*, 93:33–43.

Chanjin Zheng, Zengyi Yu, Yilin Jiang, Mingzi Zhang, Xunuo Lu, Jing Jin, and Liteng Gao. 2025. Artmentor: Ai-assisted evaluation of artworks to explore multimodal large language models capabilities. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Hancheng Zhu, Yong Zhou, Leida Li, Yaqian Li, and Yandong Guo. 2021. Learning personalized image aesthetics from subjective and objective attributes. *IEEE Transactions on Multimedia*, 25:179–190.

A Appendix

A.1 Evaluation Rubrics for Prompt-based Score Assessment

Score	Description
5	The artwork exhibits exceptional detail and precision in depicting Realism features. Textures and lighting are used masterfully to mimic real-life appearances with accurate proportions and perspective. The representation is strikingly lifelike, demonstrating advanced skills in realism.
4	The artwork presents a high level of detail and accuracy in the portrayal of subjects. Proportions and textures are very well executed, and the lighting enhances the realism. Although highly Realism, minor discrepancies in perspective or detail might be noticeable.
3	The artwork represents subjects with a moderate level of realism. Basic proportions are correct, and some textures and lighting effects are used to enhance realism. However, the depiction may lack depth or detail in certain areas.
2	The artwork attempts realism but struggles with accurate proportions and detailed textures. Lighting and perspective may be inconsistently applied, resulting in a less convincing depiction.
1	The artwork shows minimal attention to Realism details. Proportions, textures, and lighting are poorly executed, making the depiction far from lifelike.

Table 3: Realism assessment rubrics for prompting. This criterion assesses the accuracy of proportions, textures, lighting, and perspective to create a lifelike depiction.

Score	Description
5	The artwork demonstrates masterful use of deformation to enhance the emotional or conceptual impact of the piece. The transformations are thoughtful and integral to the artwork's message, seamlessly blending with the composition to engage viewers profoundly.
4	The artwork effectively uses deformation to express artistic intentions. The modifications are well-integrated and contribute significantly to the viewer's understanding or emotional response. Minor elements of the deformation might detract from its overall effectiveness.
3	The artwork includes noticeable deformations that add to its artistic expression. While these elements generally support the artwork's theme, they may be somewhat disjointed from the composition, offering mixed impact on the viewer.
2	The artwork attempts to use deformation but does so with limited success. The deformations are present but feel forced or superficial, only marginally contributing to the artwork's expressive goals.
1	The artwork features minimal or ineffective deformation, with little to no enhancement of the artwork's message or emotional impact. The attempts at deformation seem disconnected from the artwork's overall intent.

Table 4: Deformation assessment rubrics for prompting. This criterion evaluates the artist's ability to creatively and intentionally deform reality to convey a message, emotion, or concept.

Score	Description
5	The artwork masterfully employs contrasting colors to create a striking and effective visual impact.
4	The artwork effectively uses contrasting colors to enhance visual interest, though the contrast may be less pronounced.
3	The artwork has some contrast in colors, but it is not used effectively to enhance the artwork's overall appeal.
2	The artwork makes minimal use of color contrast, resulting in a lackluster visual impact.
1	The artwork lacks effective color contrast, making the piece visually unengaging.

Table 5: Color contrast assessment Rubrics for prompting. This criterion evaluates the effective use of contrasting colors to enhance artistic expression.

Score	Description
5	The artwork uses a wide and harmonious range of colors, each contributing to a vivid and dynamic composition.
4	The artwork features a good variety of colors that are well-balanced, enhancing the visual appeal of the piece.
3	The artwork includes a moderate range of colors, but the palette may not fully enhance the subject matter.
2	The artwork has limited color variety, with a palette that does not significantly contribute to the piece's impact.
1	The artwork shows poor use of colors, with a very restricted range that detracts from the visual experience.

Table 6: Color richness assessment rubrics for prompting. This criterion assesses the use and range of colors to create a visually engaging experience.

Score	Description
5	The artwork demonstrates a wide variety of line textures, each skillfully executed to enhance the piece's aesthetic and thematic elements.
4	The artwork includes a good range of line textures, well executed but with some areas that may lack definition.
3	The artwork features moderate variety in line textures, with generally adequate execution but lacking in detail.
2	The artwork has limited line textures, with execution that does not significantly contribute to the artwork's quality.
1	The artwork lacks variety and sophistication in line textures, resulting in a visually dull piece.

Table 7: Line texture assessment rubrics for prompting. This criterion evaluates the variety and execution of line textures within the artwork.

Score	Description
5	The artwork displays a profound level of originality and creativity, introducing unique concepts or interpretations that are both surprising and thought-provoking.
4	The artwork presents creative ideas that are both original and nicely executed, though they may be similar to conventional themes.
3	The artwork shows some creative ideas, but they are somewhat predictable and do not stray far from traditional approaches.
2	The artwork has minimal creative elements, with ideas that are largely derivative and lack originality.
1	The artwork lacks imagination, with no discernible original ideas or creative concepts.

Table 8: Imagination assessment rubrics for prompting. This criterion evaluates the artist's ability to use their creativity to form unique and original ideas within their artwork.

Score	Description
5	The artwork exhibits exceptional integration of line combinations, creating a harmonious and engaging visual flow.
4	The artwork displays good use of line combinations that contribute to the overall composition, though some areas may lack cohesion.
3	The artwork shows average use of line combinations, with some effective sections but overall lacking in cohesiveness.
2	The artwork has minimal effective use of line combinations, with lines that often clash or do not contribute to a unified composition.
1	The artwork shows poor integration of lines, with combinations that disrupt the visual harmony of the piece.

Table 9: Line combination assessment Rubrics for prompting. This criterion assesses the integration and interaction of lines within the artwork.

Score	Description
5	The artwork is impeccably organized, with each element thoughtfully placed to create a balanced and compelling composition.
4	The artwork has a good organization, with a well-arranged composition that effectively guides the viewer's eye, though minor elements may disrupt the flow.
3	The artwork has an adequate organization, but the composition may feel somewhat unbalanced or disjointed.
2	The artwork shows poor organization, with a composition that lacks coherence and does not effectively engage the viewer.
1	The artwork is poorly organized, with a chaotic composition that detracts from the piece's overall impact.

Table 10: Picture organization assessment rubrics for prompting. This criterion evaluates the overall composition and spatial arrangement within the artwork.

Score	Description
5	The artwork is transformative, offering a fresh and innovative take on traditional elements, significantly enhancing the viewer's experience.
4	The artwork successfully transforms familiar elements, providing a new perspective, though the innovation may not be striking.
3	The artwork shows some transformation of familiar elements, but the changes are somewhat predictable and not highly innovative.
2	The artwork attempts transformation but achieves only minimal success, with changes that are either too subtle or not effectively executed.
1	The artwork lacks transformation, with traditional elements that are replicated without any significant innovation or creative reinterpretation.

Table 11: Transformation assessment rubrics for prompting. This criterion assesses the artist's ability to transform traditional or familiar elements into something new and unexpected.

A.2 Prompt Example

Simple Prompt Example

<image> Assess the student's artwork based on the 'Color contrast ' criterion and provide a score. This criterion evaluates the effective use of contrasting colors to enhance artistic expression. Output a score (1-5).

Rubric-based Prompt Example

<image> Assess the student's artwork based on the 'Color contrast ' criterion and provide a score.
This criterion evaluates the effective use of contrasting colors to enhance artistic expression.
5: The artwork masterfully employs contrasting colors to create a striking and effective visual impact.
4: The artwork effectively uses contrasting colors to enhance visual interest, though the contrast may be less pronounced.
3: The artwork has some contrast in colors, but it is not used effectively to enhance the artwork's overall appeal.
2: The artwork makes minimal use of color contrast, resulting in a lackluster visual impact.
1: The artwork lacks effective color contrast, making the piece visually unengaging.
Output a score (1-5).

Few-shot Prompt Example

<image> Assess the student's artwork based on the 'Color contrast' criterion and provide a score.
<image> This is an example of 'Color contrast' with a score of 1.
<image> This is an example of 'Color contrast' with a score of 5.
This criterion evaluates the effective use of contrasting colors to enhance artistic expression.
5: The artwork masterfully employs contrasting colors to create a striking and effective visual impact.
4: The artwork effectively uses contrasting colors to enhance visual interest, though the contrast may be less pronounced.
3: The artwork has some contrast in colors, but it is not used effectively to enhance the artwork's overall appeal.
2: The artwork makes minimal use of color contrast, resulting in a lackluster visual impact.
1: The artwork lacks effective color contrast, making the piece visually unengaging.
Output a score (1-5).

Prompt Example with Comments for Fine-tuning

<image> Assess the student's artwork based on the 'Color contrast' criterion and provide a score.
This criterion evaluates the effective use of contrasting colors to enhance artistic expression.
5: The artwork masterfully employs contrasting colors to create a striking and effective visual impact.
4: The artwork effectively uses contrasting colors to enhance visual interest, though the contrast may be less pronounced.
3: The artwork has some contrast in colors, but it is not used effectively to enhance the artwork's overall appeal.
2: The artwork makes minimal use of color contrast, resulting in a lackluster visual impact.
1: The artwork lacks effective color contrast, making the piece visually unengaging.
Reference/Expert/Teacher comment: [Insert the expert's qualitative feedback here, if available.]
Output a score (1-5).

A.3 Detailed Results

This section details results on our test dataset. To ensure a fair comparison across MLLMs, all images are uniformly resized to 448×448 during preprocessing, while MLLMs are configured with consistent parameters ($max\_new\_tokens = 128$, $temperature = 0.7$, $top_k = 50$, and $top_p = 1.0$). Table 12-19 summarize the results of our prompting experiments for attribute-aware evaluation using MLLMs. Table 20-21 present the scores from two practicing art teachers, which serve as the human agreements on the test dataset. We used up to 4 NVIDIA A100 GPUs for all experiments.

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.652	0.647	0.760	0.240	0.613
Deformation	0.420	0.408	0.560	0.470	0.307
Imagination	0.549	0.558	0.220	0.780	0.201
Color Richness	0.651	0.646	0.730	0.270	0.628
Color Contrast	0.432	0.473	0.510	0.490	0.336
Line Combination	0.284	0.292	0.510	0.490	0.222
Line Texture	0.371	0.388	0.490	0.510	0.230
Picture Organization	0.549	0.541	0.710	0.290	0.510
Transformation	0.474	0.476	0.520	0.540	0.338
Average	0.487	0.492	0.557	0.453	0.376

Table 12: Prompting results with Qwen2.5-VL-72B

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.580	0.584	0.53	0.53	0.47
Deformation	0.490	0.472	0.62	0.41	0.411
Imagination	0.476	0.459	0.22	1.47	0.183
Color Richness	0.594	0.608	0.64	0.39	0.579
Color Contrast	0.476	0.459	0.62	0.38	0.458
Line Combination	0.167	0.163	0.72	0.28	0.128
Line Texture	0.312	0.288	0.60	0.40	0.211
Picture Organization	0.430	0.422	0.69	0.31	0.413
Transformation	0.399	0.444	0.44	0.68	0.347
Overall	0.436	0.433	0.564	0.539	0.355

Table 13: Prompting results with Qwen2.5-VL-32B

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.635	0.637	0.570	0.460	0.511
Deformation	0.506	0.530	0.620	0.380	0.458
Imagination	0.495	0.481	0.200	1.220	0.194
Color Richness	0.589	0.604	0.630	0.400	0.547
Color Contrast	0.524	0.505	0.670	0.330	0.494
Line Combination	0.310	0.310	0.730	0.270	0.289
Line Texture	0.147	0.196	0.530	0.500	0.127
Picture Organization	0.579	0.565	0.750	0.250	0.550
Transformation	0.423	0.451	0.480	0.610	0.374
Average	0.468	0.475	0.576	0.491	0.394

Table 14: Prompting results with Qwen2.5-VL-7B

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.439	0.446	0.620	0.380	0.321
Deformation	0.335	0.320	0.470	0.560	0.282
Imagination	0.325	0.342	0.540	0.520	0.293
Color Richness	0.582	0.588	0.550	0.510	0.527
Color Contrast	0.352	0.347	0.320	1.090	0.218
Line Combination	0.317	0.322	0.610	0.420	0.278
Line Texture	0.427	0.454	0.580	0.420	0.428
Picture Organization	0.276	0.296	0.600	0.430	0.289
Transformation	0.214	0.232	0.540	0.610	0.210
Average	0.363	0.372	0.537	0.549	0.316

Table 15: Prompting results with Qwen3-VL-30B-A3B

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.529	0.544	0.69	0.31	0.426
Deformation	0.246	0.282	0.22	1.05	0.093
Imagination	0.417	0.405	0.09	1.72	0.113
Color Richness	0.589	0.609	0.55	0.45	0.492
Color Contrast	0.400	0.383	0.63	0.43	0.336
Line Combination	0.151	0.151	0.48	0.67	0.094
Line Texture	0.371	0.390	0.64	0.36	0.366
Picture Organization	0.300	0.306	0.44	0.68	0.187
Transformation	0.396	0.405	0.26	1.13	0.197
Average	0.378	0.386	0.444	0.756	0.256

Table 16: Prompting results with by Gemma3-12B

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.543	0.550	0.720	0.529	0.451
Deformation	0.118	0.127	0.210	1.127	0.012
Imagination	0.517	0.513	0.690	0.557	0.427
Color Richness	0.655	0.652	0.680	0.566	0.544
Color Contrast	0.447	0.462	0.560	0.686	0.284
Line Combination	0.282	0.272	0.750	0.500	0.154
Line Texture	0.215	0.277	0.600	0.678	0.156
Picture Organization	0.536	0.542	0.650	0.592	0.368
Transformation	0.147	0.155	0.310	0.949	0.029
Average	0.384	0.395	0.574	0.687	0.269

Table 17: Prompting results with by Gemma3-27B

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.503	0.513	0.680	0.320	0.500
Deformation	0.362	0.407	0.640	0.420	0.381
Imagination	0.502	0.504	0.130	1.140	0.162
Color Richness	0.409	0.428	0.560	0.440	0.398
Color Contrast	0.422	0.425	0.450	0.640	0.254
Line Combination	0.212	0.208	0.490	0.570	0.145
Line Texture	0.264	0.284	0.320	1.190	0.166
Picture Organization	0.435	0.474	0.510	0.490	0.339
Transformation	0.418	0.408	0.510	0.630	0.344
Average	0.392	0.406	0.477	0.649	0.299

Table 18: Prompting results with by Mimo-VL-7B

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.245	0.250	0.02	3.05	0.041
Deformation	0.237	0.193	0.69	0.40	0.192
Imagination	0.236	0.282	0.31	0.72	0.085
Color Richness	0.419	0.411	0.63	0.40	0.360
Color Contrast	0.186	0.160	0.46	0.68	0.086
Line Combination	0.167	0.165	0.35	0.74	0.072
Line Texture	0.102	0.097	0.31	1.14	0.048
Picture Organization	0.096	0.084	0.36	0.82	0.019
Transformation	0.251	0.254	0.56	0.47	0.232
Average	0.215	0.211	0.41	0.936	0.126

Table 19: Prompting results with by Kimi-a3b

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.544	0.551	0.620	0.56	0.517
Deformation	0.700	0.778	0.760	0.24	0.760
Imagination	0.707	0.711	0.860	0.14	0.706
Color Richness	0.889	0.919	0.920	0.08	0.917
Color Contrast	0.728	0.758	0.840	0.22	0.717
Line Combination	0.828	0.814	0.920	0.14	0.801
Line Texture	0.378	0.389	0.760	0.30	0.375
Picture Organization	0.310	0.327	0.440	0.68	0.226
Transformation	0.629	0.587	0.660	0.46	0.567
Average	0.629	0.587	0.753	0.313	0.567

Table 20: Teacher 1 performance

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.751	0.737	0.580	0.60	0.688
Deformation	0.614	0.596	0.740	0.32	0.573
Imagination	0.252	0.225	0.620	0.50	0.208
Color Richness	0.888	0.917	0.920	0.08	0.913
Color Contrast	0.736	0.779	0.920	0.14	0.769
Line Combination	0.899	0.885	0.900	0.10	0.878
Line Texture	0.483	0.479	0.740	0.26	0.444
Picture Organization	0.910	0.872	0.920	0.08	0.863
Transformation	0.571	0.563	0.760	0.36	0.561
Average	0.571	0.563	0.789	0.271	0.561

Table 21: Teacher 2 performance

A.4 Results and Ablation for Base Model

This appendix presents detailed ablation studies with base model (Qwen2.5-VL-7B). Table 22 reports the prompting ablation results, where models are evaluated without rubric guidance to examine the effect of structured rubrics prompting. Tables 24–27 summarize the results of multi-LoRA and RAFT-based ablation experiments, illustrating the contribution of attribute-specific adaptation and feature fusion. Tables 28–30 show experiments conducted on the comment-annotated subset, comparing models trained with and without comment signals as well as additional data augmentation through multi-instruction templates. For model training, we fix the image resolution to 224×224 and adopt a base training configuration without comment augmentation, which serves as the reference setting throughout the paper. This base configuration uses rank = 8, lora_alpha = 16, learning rate = 2e–5, and batch size = 16, and corresponds to the results reported in Table 2 and 33. Unless otherwise specified, all experiments and ablations are conducted by modifying one factor at a time relative to this base configuration. The corresponding parameter tuning ablations are summarized in Tables 34–39. Finally, Table 31 reports the comparative results of visual augmentation strategies.

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.640	0.680	0.780	0.220	0.621
Deformation	0.219	0.203	0.730	0.300	0.120
Imagination	0.258	0.245	0.270	0.730	0.053
Color Richness	0.404	0.425	0.630	0.370	0.327
Color Contrast	0.262	0.313	0.540	0.460	0.163
Line Combination	0.168	0.169	0.550	0.480	0.145
Line Texture	0.277	0.276	0.270	1.120	0.123
Picture Organization	0.446	0.469	0.600	0.400	0.405
Transformation	0.283	0.276	0.670	0.330	0.217
Average	0.328	0.340	0.560	0.490	0.241

Table 22: Simple prompting results

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.612	0.610	0.510	0.580	0.463
Deformation	0.337	0.393	0.330	0.670	0.223
Imagination	0.410	0.407	0.380	1.190	0.226
Color Richness	0.535	0.561	0.360	0.790	0.348
Color Contrast	0.523	0.539	0.280	0.930	0.298
Line Combination	0.270	0.287	0.510	0.790	0.218
Line Texture	0.236	0.229	0.350	1.720	0.118
Picture Organization	0.629	0.625	0.270	1.060	0.355
Transformation	0.329	0.324	0.440	0.770	0.294
Average	0.431	0.442	0.381	0.944	0.282

Table 23: Few-shot prompting results (two-shot).

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.457	0.459	0.700	0.330	0.450
Deformation	0.524	0.518	0.770	0.230	0.511
Imagination	0.581	0.573	0.790	0.210	0.573
Color Richness	0.615	0.617	0.660	0.370	0.615
Color Contrast	0.517	0.519	0.690	0.340	0.519
Line Combination	0.254	0.245	0.650	0.350	0.238
Line Texture	0.285	0.278	0.500	0.530	0.270
Picture Organization	0.583	0.612	0.770	0.230	0.610
Transformation	0.374	0.377	0.630	0.370	0.359
Average	0.466	0.466	0.684	0.329	0.461

Table 24: Results with LoRA and standard regression.

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.618	0.618	0.760	0.240	0.613
Deformation	0.553	0.514	0.810	0.220	0.505
Imagination	0.584	0.577	0.790	0.210	0.574
Color Richness	0.662	0.661	0.670	0.330	0.654
Color Contrast	0.573	0.600	0.710	0.290	0.600
Line Combination	0.321	0.323	0.720	0.280	0.310
Line Texture	0.335	0.344	0.540	0.490	0.332
Picture Organization	0.469	0.486	0.710	0.290	0.481
Transformation	0.448	0.456	0.640	0.360	0.430
Average	0.507	0.508	0.706	0.301	0.499

Table 25: Results with multi-LoRA and standard regression.

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.545	0.541	0.750	0.250	0.537
Deformation	0.465	0.447	0.720	0.310	0.429
Imagination	0.567	0.569	0.810	0.190	0.558
Color Richness	0.700	0.694	0.710	0.290	0.693
Color Contrast	0.650	0.625	0.750	0.250	0.610
Line Combination	0.367	0.356	0.760	0.240	0.330
Line Texture	0.381	0.395	0.590	0.410	0.385
Picture Organization	0.466	0.476	0.720	0.280	0.428
Transformation	0.523	0.519	0.720	0.280	0.497
Average	0.518	0.514	0.726	0.278	0.496

Table 26: Results with LoRA and RAFT.

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.762	0.800	0.890	0.106	0.804
Deformation	0.670	0.664	0.810	0.153	0.568
Imagination	0.689	0.652	0.810	0.170	0.610
Color Richness	0.745	0.760	0.740	0.201	0.708
Color Contrast	0.640	0.625	0.700	0.227	0.480
Line Combination	0.517	0.512	0.730	0.199	0.396
Line Texture	0.513	0.533	0.620	0.247	0.444
Picture Organization	0.654	0.682	0.710	0.178	0.505
Transformation	0.643	0.650	0.720	0.237	0.574
Average	0.648	0.653	0.748	0.191	0.566

Table 27: Results with multi-LoRA and RAFT.

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.411	0.433	0.604	0.423	0.312
Deformation	0.455	0.427	0.703	0.324	0.406
Imagination	0.202	0.198	0.441	0.559	0.120
Color Richness	0.683	0.667	0.730	0.297	0.656
Color Contrast	0.181	0.190	0.514	0.622	0.183
Line Combination	-0.065	-0.033	0.297	0.757	-0.019
Line Texture	0.172	0.164	0.369	0.847	0.108
Picture Organization	0.386	0.402	0.441	0.586	0.301
Transformation	0.316	0.326	0.396	0.712	0.181
Average	0.304	0.308	0.499	0.570	0.250

Table 28: Results on the 100-sample subset without utilising comments.

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.477	0.476	0.685	0.342	0.390
Deformation	0.322	0.334	0.459	0.541	0.227
Imagination	0.196	0.194	0.450	0.550	0.128
Color Richness	0.603	0.622	0.523	0.505	0.508
Color Contrast	0.502	0.564	0.360	0.775	0.355
Line Combination	0.067	0.078	0.631	0.450	0.051
Line Texture	0.219	0.193	0.468	0.802	0.147
Picture Organization	0.451	0.472	0.613	0.387	0.349
Transformation	0.359	0.340	0.450	0.739	0.245
Average	0.355	0.364	0.516	0.566	0.267

Table 29: Results on the 100-sample subset with comments.

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.654	0.640	0.739	0.212	0.485
Deformation	0.494	0.495	0.396	0.481	0.158
Imagination	0.531	0.359	0.450	0.496	0.132
Color Richness	0.704	0.720	0.396	0.599	0.424
Color Contrast	0.615	0.661	0.640	0.297	0.545
Line Combination	0.081	0.120	0.486	0.729	0.040
Line Texture	0.259	0.250	0.505	0.465	0.137
Picture Organization	0.541	0.589	0.622	0.326	0.160
Transformation	0.420	0.411	0.486	0.617	0.252
Average	0.478	0.472	0.525	0.469	0.259

Table 30: Results on the 100-sample subset with comments and data augmentation (Colorjitter+Multi-copies).

Dimension	SC	PC	ACC	MSE	QWK
Base	0.545	0.546	0.728	0.275	0.530
Rotation	0.385	0.385	0.727	0.280	0.353
Hflip	0.458	0.475	0.731	0.276	0.466
Blur	0.523	0.520	0.733	0.270	0.508
Colorjitter	0.582	0.585	0.743	0.263	0.558

Table 31: Results under different augmentation strategies.

Dimension	PC (w/o C)	RMSE (w/o C)	QWK (w/o C)	PC (w C)	RMSE (w C)	QWK (w C)
Realism	0.726	0.387	0.722	0.697	0.387	0.519
Deformation	0.570	0.469	0.557	0.599	0.434	0.512
Imagination	0.575	0.469	0.559	0.677	0.374	0.664
Color Richness	0.679	0.538	0.679	0.759	0.450	0.722
Color Contrast	0.591	0.529	0.586	0.660	0.468	0.683
Line Combination	0.325	0.490	0.290	0.445	0.455	0.332
Line Texture	0.486	0.600	0.468	0.442	0.555	0.380
Picture Organization	0.640	0.447	0.627	0.641	0.435	0.506
Transformation	0.603	0.500	0.599	0.587	0.481	0.537
Average	0.577	0.496	0.565	0.612	0.452	0.539

Table 32: Results on the full dataset w/wo comments.

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.604	0.599	0.780	0.220	0.593
Deformation	0.526	0.488	0.780	0.250	0.480
Imagination	0.618	0.611	0.810	0.190	0.604
Color Richness	0.688	0.700	0.670	0.330	0.690
Color Contrast	0.716	0.722	0.790	0.210	0.710
Line Combination	0.341	0.342	0.730	0.270	0.324
Line Texture	0.437	0.455	0.570	0.430	0.392
Picture Organization	0.414	0.434	0.690	0.310	0.426
Transformation	0.565	0.568	0.730	0.270	0.557
Average	0.546	0.546	0.728	0.276	0.531

Table 33: Results with the base configuration.

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.602	0.609	0.770	0.230	0.603
Deformation	0.482	0.447	0.770	0.260	0.442
Imagination	0.652	0.644	0.830	0.170	0.641
Color Richness	0.665	0.681	0.700	0.300	0.680
Color Contrast	0.667	0.672	0.760	0.240	0.671
Line Combination	0.320	0.310	0.730	0.270	0.298
Line Texture	0.408	0.412	0.520	0.480	0.325
Picture Organization	0.485	0.490	0.720	0.280	0.461
Transformation	0.507	0.507	0.720	0.280	0.496
Average	0.532	0.530	0.724	0.279	0.513

Table 34: Results with batch_size = 4.

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.563	0.558	0.760	0.240	0.556
Deformation	0.451	0.418	0.740	0.290	0.410
Imagination	0.634	0.627	0.820	0.180	0.622
Color Richness	0.695	0.710	0.730	0.270	0.706
Color Contrast	0.613	0.637	0.740	0.260	0.636
Line Combination	0.340	0.329	0.740	0.260	0.313
Line Texture	0.391	0.385	0.560	0.440	0.363
Picture Organization	0.542	0.552	0.750	0.250	0.545
Transformation	0.554	0.557	0.730	0.270	0.537
Average	0.532	0.530	0.730	0.273	0.521

Table 35: Results with batch_size = 24.

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.478	0.481	0.710	0.290	0.463
Deformation	0.474	0.440	0.780	0.250	0.431
Imagination	0.629	0.621	0.810	0.190	0.610
Color Richness	0.726	0.737	0.740	0.260	0.734
Color Contrast	0.691	0.697	0.780	0.220	0.693
Line Combination	0.338	0.326	0.730	0.270	0.311
Line Texture	0.407	0.402	0.560	0.440	0.363
Picture Organization	0.510	0.519	0.740	0.260	0.497
Transformation	0.563	0.569	0.750	0.250	0.556
Average	0.535	0.532	0.733	0.270	0.518

Table 36: Results with learning_rate = 1e-5.

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.642	0.633	0.800	0.200	0.630
Deformation	0.455	0.422	0.770	0.260	0.415
Imagination	0.654	0.649	0.840	0.160	0.649
Color Richness	0.713	0.727	0.740	0.260	0.719
Color Contrast	0.664	0.668	0.760	0.270	0.662
Line Combination	0.230	0.238	0.720	0.280	0.197
Line Texture	0.385	0.420	0.530	0.470	0.336
Picture Organization	0.494	0.498	0.710	0.290	0.492
Transformation	0.442	0.444	0.680	0.320	0.425
Average	0.520	0.522	0.728	0.279	0.503

Table 37: Results with learning_rate = 5e-5.

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.583	0.576	0.770	0.230	0.574
Deformation	0.331	0.308	0.750	0.280	0.268
Imagination	0.616	0.612	0.820	0.180	0.611
Color Richness	0.724	0.736	0.750	0.250	0.735
Color Contrast	0.645	0.620	0.750	0.250	0.608
Line Combination	0.240	0.237	0.730	0.270	0.216
Line Texture	0.296	0.286	0.550	0.480	0.283
Picture Organization	0.535	0.543	0.750	0.250	0.529
Transformation	0.567	0.568	0.740	0.260	0.549
Average	0.504	0.498	0.734	0.272	0.486

Table 38: Results with rank = 4 and lora_alpha = 16.

Dimension	SC	PC	ACC	MSE	QWK
Realism	0.528	0.527	0.740	0.260	0.519
Deformation	0.496	0.460	0.770	0.260	0.454
Imagination	0.649	0.645	0.840	0.160	0.643
Color Richness	0.732	0.743	0.750	0.250	0.741
Color Contrast	0.613	0.627	0.720	0.280	0.627
Line Combination	0.263	0.258	0.740	0.260	0.231
Line Texture	0.341	0.344	0.550	0.450	0.325
Picture Organization	0.524	0.523	0.740	0.260	0.513
Transformation	0.578	0.579	0.740	0.260	0.569
Average	0.525	0.523	0.732	0.271	0.513

Table 39: Results with rank = 16 and lora_alpha = 32.

A.5 Annotation Details

For the annotation process, evaluating children's artwork inherently involves conceptual overlap among dimensions and a degree of subjectivity, as many aesthetic and creative attributes are conceptually adjacent rather than strictly orthogonal. Our rubric is therefore designed to provide a practically usable and pedagogically meaningful decomposition of children's artistic expression, balancing fine-grained descriptive power with the need for consistent large-scale annotation.

To ensure reliability, we employed a multi-stage annotation protocol in which initial ratings were iteratively refined and verified, and final scores were consolidated by expert annotators. This process aimed to standardize rubric interpretation and reduce variability across annotators. The resulting framework reflects a deliberate design choice: dimensions are more specific than broad holistic categories yet not as atomized as highly detailed taxonomies, enabling nuanced assessment without sacrificing feasibility. Some degree of inter-dimension correlation is structurally expected given the conceptual nature of the attributes. In art-education practice, for instance, Imagination, Deformation, and Transformation often co-occur in children's creative processes, while remaining qualitatively distinct (e.g., imagination foregrounding novelty, transformation emphasizing systematic modification). Furthermore, analyses of model behavior and dataset structure (Section 5.4) support these distinctions.

To quantify annotation reliability, we computed inter-rater agreement prior to expert consolidation, as shown in Table 40. Concrete perceptual attributes such as color richness and color contrast show high agreement (α = 0.83–0.86), whereas more abstract, intent-based attributes – including imagination, deformation, and transformation – exhibit lower agreement. This pattern reflects the inherent interpretive difficulty of these constructs. Importantly, the labels used in our experiments are derived from the multi-stage consolidation workflow, which explicitly resolves disagreements and improves consistency relative to the initial independent ratings.

A.6 Qualitative results

Throughout this section, the base model refers to the model trained using the base configuration detailed in Appendix A.4, without comment augmentation and with the default hyperparameter setting (rank=8, lora_alpha= 6, learning_rate=2e-5, batch_size=16). Figure 6 illustrates the distribution of our base model's predictions across each artistic dimension and the overall distribution. When compared with the ground-truth score distributions shown in Figure 3, the predicted results exhibit a broadly consistent statistical pattern, indicating that the model captures the relative score tendencies rather than producing random or biased estimations. This similarity suggests that the model has learned to approximate human scoring behavior in a stable and interpretable manner.

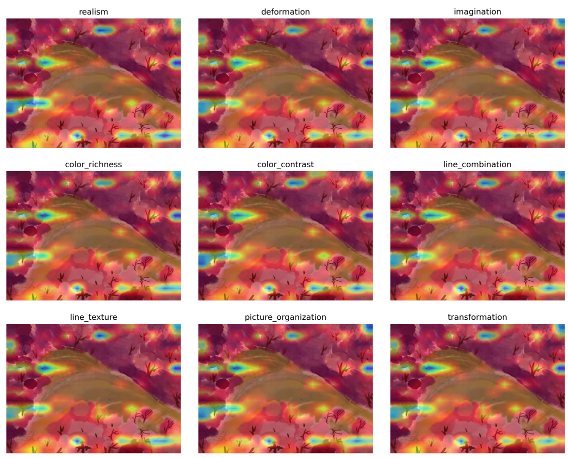

Figure 7: Spatial visual attention visualized by Gradient-CAM for an example image.

To further explore the model's perceptual reasoning, Figure 7 presents spatial visual attention maps generated via Gradient-CAM. We observe that models for different dimensions attend to distinct regions and details, while maintaining some shared focus on general visual structures. This indicates both specialization and commonality in attention mechanisms across aesthetic dimensions.

To investigate the representational structure underlying these attention patterns, we conduct an attention-based clustering analysis (Figure 8- 9). We conduct visualization experiments on the test dataset mentioned in Section 3, from which statisti-

Dimension	PC_gt_T1	PC_gt_T2	PC_T1_T2	ICC_gt_T1	ICC_gt_T2	ICC_T1_T2	α_gt_T1	α_gt_T2	α_T1_T2
Realism	0.7611	0.7433	0.5602	0.7383	0.7335	0.5581	0.7906	0.7592	0.6050
Deformation	0.4806	0.5521	0.2744	0.3935	0.4424	0.2744	0.4072	0.4484	0.2377
Imagination	0.5013	0.4874	0.2225	0.4269	0.4234	0.2224	0.4604	0.4482	0.2145
Color Richness	0.8655	0.8701	0.7606	0.8612	0.8647	0.7607	0.8567	0.8613	0.7509
Color Contrast	0.8357	0.8313	0.6971	0.8254	0.8199	0.6968	0.8274	0.8345	0.7091
Line Combination	0.5449	0.5584	0.2841	0.4737	0.4707	0.2837	0.4967	0.5122	0.2890
Line Texture	0.6217	0.6246	0.3612	0.5728	0.5652	0.3610	0.6047	0.5916	0.3554
Picture Organization	0.7425	0.7145	0.5191	0.7102	0.6831	0.5191	0.7614	0.7344	0.5672
Transformation	0.5311	0.5190	0.2666	0.4657	0.4589	0.2667	0.4928	0.4872	0.2530

Table 40: Inter-rater agreement statistics across dimensions, including Pearson correlation (PC), intraclass correlation coefficient (ICC), and Krippendorff's alpha (α).

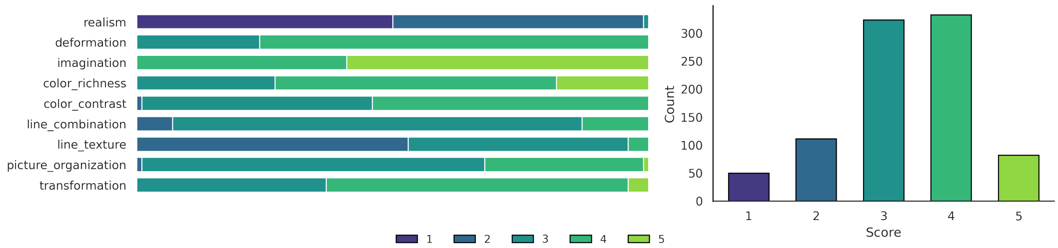

Figure 6: The distribution for our base model predictions with each dimension (Left) and overall results (Right).

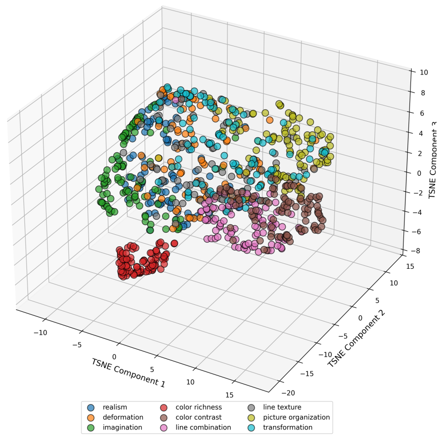

Figure 8: t-SNE projection of final-layer attention features across dimensions.

cal attention features were extracted from the final layer and projected into a shared embedding space using t-SNE and PCA projection. The clusters corresponding to different dimensions (e.g., Color Richness, Imagination, Picture Organization) exhibit clear separability, suggesting that the models capture distinct perceptual representations aligned with artistic attributes. However, the Line Texture cluster (in gray) appears notably dispersed, indicating weaker intra-dimensional consistency—an observation that aligns with its relatively lower quantitative performance (see Figure 2 and Table 2). This suggests that Line Texture remains a challenging aspect for the model to comprehend, likely due to its subtle and fine-grained visual characteristics.

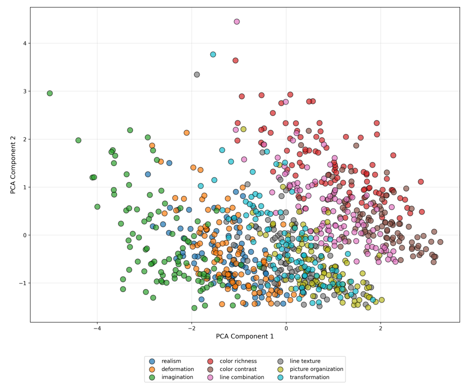

Figure 9: PCA projection of final-layer attention features across dimensions.

Overall, these results indicate that the learned feature representations corresponding to different dimensions are not overly abstract or heavily overlapping. Across data-level patterns, model performance trends, and feature-space structures, the results collectively support the necessity of explicitly modeling these dimensions. In the context of our task, this dimensional formulation enables more faithful modeling of the diverse perceptual and cognitive aspects involved in children's artwork assessment, thereby supporting more interpretable and fine-grained evaluation outcomes.

A.7 Enlarged Figures

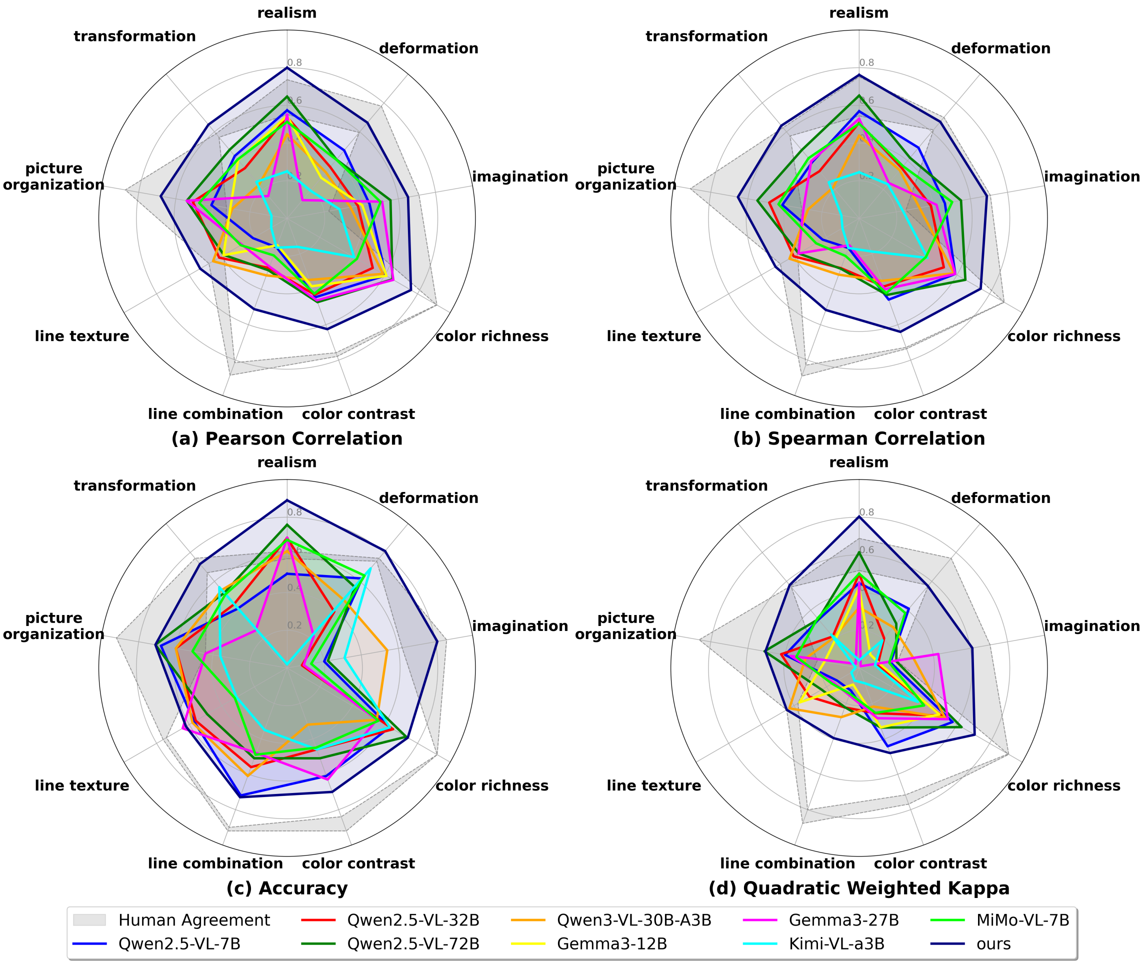

Figure 10: Enlarged version of Figure 4 in the main paper.

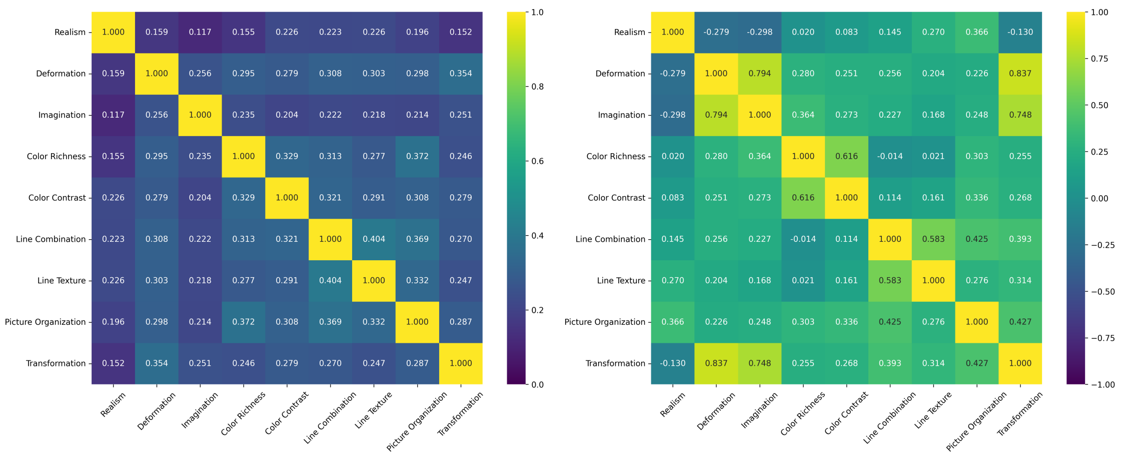

Figure 11: Enlarged version of Figure 5 in the main paper.