AEM45K: A Large-Scale Multimodal Dataset and Method for Aesthetic Assessment of Student Artwork in Middle-School Education

Shuhao Hu[1], Yanhao Chen[1], Yan Liu[1], Ruoxuan Liang[1], Junjie Jiao[1],
Jing Jin[2,3], Wei Liu[4], Yin Zhang[4], Lei Wang[4,*], Qingqiang Wu[1,5,*]
[1]School of Film, Xiamen University, China
[5]School of Informatics, Xiamen University, China
Email: {horacehsh, cyhao, liuyany, liangruoxuan, jiaojj1204}@stu.xmu.edu.cn, wuqq@xmu.edu.cn
[2]Zhejiang Normal University, China
[3]Hangzhou Tianchang Guanchao Primary School, China
Email: 383230730@qq.com
[4]Academy of Military Medical Sciences, Academy of Military Sciences, China
Email: liuweilcf@163.com, aring2010@163.com, wangleienjoy@163.com

***Abstract*—In middle-school art education, teachers must grade large volumes of student drawings under strict time constraints, while ensuring rubric consistency and multi-dimensional feedback. To support scalable educational assessment, we present AEM45K, a dataset of 44,573 exam-condition drawings annotated with five rubric-aligned attributes and a final grade. A verified subset of 7,360 samples further includes student-written descriptions that clarify intent and theme, treated as an *optional* modality. We propose MADENet, a unified framework for rubric-aligned educational aesthetics assessment with or without text. MADENet retrieves attribute-specific evidence via a dimension query bank, grounds queries to image/text tokens through query-to-token cross-attention, and applies dimension-adaptive gated fusion to balance modalities under varying availability. Experiments under image-only and multimodal protocols show consistent improvements over strong baselines, with the largest gains on semantics-heavy dimensions when descriptions are available.**

***Index Terms*—image aesthetics assessment, multimodal learning, middle-school art education, rubric-aligned evaluation**

I. Introduction

In middle-school art education, aesthetic assessment is routine yet high-stakes. Large volumes of student drawings from homework, tests, and standardized exams must be graded under strict time constraints. Manual grading is often affected by rater inconsistency, fatigue, and rubric misalignment across classes or schools, which can compromise fairness and feedback quality. These challenges motivate **Image Aesthetics Assessment in Education (IAAE)**, where models should predict both an overall grade and *rubric-aligned multi-dimensional scores* for scalable and consistent assessment.

Despite strong progress in photo aesthetics assessment [1]–[3], educational drawings remain challenging and under-resourced. Compared with photographs, exam drawings exhibit sparse textures, varying abstraction, and large stylistic variance. More importantly, rubrics emphasize semantics-heavy dimensions that depend on student intent rather than low-level visual pleasantness, making direct transfer potentially shortcut-driven. Meanwhile, existing datasets are misaligned with rubric-driven educational grading: photo-centric benchmarks do not match educational criteria, and art-oriented datasets provide only partial coverage—JenAesthetics [4] and VAPS [5] are limited in scale, APDD [6] broadens styles but is not designed for rubric-based education, and AACP [7] focuses on children's drawings with a narrow age range and includes a non-trivial portion of generated images. As a result, current benchmarks rarely provide large-scale *exam-condition* student drawings with *multi-attribute rubric labels* and *intent-related text*, which are essential for educational assessment.

To bridge this gap, we build **AEM45K**, a dataset of **44,573** middle-school exam drawings collected across multiple schools and semesters. Each artwork is annotated by experienced art teachers with five rubric-aligned attributes (SHAPE, CREATIVITY, THEME, COLOR, INTERPRETATION) and an overall FINAL grade. A verified subset of **7,360** samples further includes short student-authored descriptions clarifying intent and theme, which we treat as an *optional* modality. Accordingly, we define two practical evaluation protocols: *image-only* (44,573) and *multimodal* (7,360).

Based on AEM45K, we propose **MADENet**, a unified framework for rubric-aligned educational assessment with or without text. MADENet retrieves attribute-specific evidence via a *dimension query bank*, grounds queries to image/text tokens through *query-to-token cross-attention*, and applies *dimension-adaptive gated fusion* to balance modalities per attribute. Experiments under both protocols demonstrate consistent improvements over strong image-only and multimodal baselines. This highlights the practical value of rubric-aware and modality-flexible assessment for real-world middle-school art education.

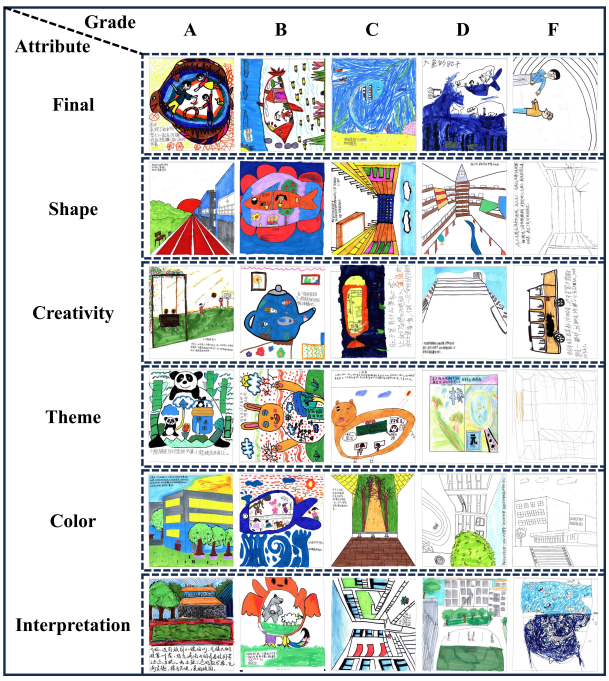

Fig. 1: Examples of images and grading levels in AEM45K. The aesthetics of paintings in middle school education are evaluated based on grade levels assigned to five attributes, along with a final grade.

II. Related Works

A. IAA Datasets

Large-scale datasets have driven rapid progress in **Image Aesthetics Assessment (IAA)**, especially for photography. Benchmarks such as AVA [1] enable large-scale learning, while AADB [8] provides additional interpretable attributes; later datasets (e.g., TAD66K [11], ICAA17K [12]) further expand scale and supervisory signals. However, these photo-centric resources are not designed for rubric-based educational grading and rarely capture the creator's intent or textual explanations.

For artistic imagery, datasets such as JenAesthetics [4] and VAPS [5] provide expert-labeled paintings at modest scale, and APDD [6] broadens stylistic diversity. AACP [7] focuses on children's drawings, but covers a narrow age range. As summarized in Table I, existing art-related datasets are typically limited in scale and seldom include **rubric-aligned multi-attribute labels** or **text/intent annotations**, which are crucial for educational assessment where semantics-heavy dimensions often require intent disambiguation.

B. Models

Early IAA relied on hand-crafted features and photographic priors. With deep learning, RAPID [15] addressed binary aesthetics via multi-view CNNs, and NIMA [3] introduced distributional supervision (EMD). Subsequent work explored multi-branch and multi-task designs [2], [16]—and mitigated resizing artifacts via composition-preserving pooling or patch aggregation [17].

Recent *multimodal* approaches incorporate vision and language to model subjective aesthetic judgments. CLIP-based pipelines [18] provide strong pretrained representations, and methods such as MSCAN [19], CDCM-Net [20], and Aes-Mamba [21] further explore cross-modal encoders and fusion strategies. IAACLIP [22] demonstrates that textual supervision can benefit aesthetics assessment. However, most existing methods rely on *generic or synthetic* language signals and adopt *uniform* fusion mechanisms, which may introduce stylistic bias and dilute dimension-specific evidence—especially for rubric-driven educational assessment. These limitations motivate our *dimension-aware* design that performs per-attribute evidence retrieval and adaptive modality fusion using authentic student intent when available.

III. Dataset

We construct **AEM45K** for Image Aesthetics Assessment in Education (IAAE), targeting rubric-driven assessment of middle-school exam drawings, as illustrated in Fig. 1.

A. Data Collection and Rubric

Collection setting. We collect drawings from middle-school final exams and targeted drawing tests across multiple schools and semesters, where prompts, time limits, and scoring rubrics are standardized. This setting reduces confounding factors and aligns with real-world educational grading workflows.

Rubric and labels. Guided by art-education experts, we adopt a compact five-dimension rubric—**Shape**, **Creativity**, **Theme**, **Color**, **Interpretation**—aligned with mainstream practice and usable under exam conditions. It is discriminative and reliable, and the *Final* score is computed as a normalized weighted sum (Shape 0.35, Color 0.25, Theme 0.15, Creativity 0.15, Interpretation 0.10), reflecting "technical control first, conceptual expression also valued."

Raters and aggregation. Each work is independently scored by qualified art teachers (more than 5 years of teaching experience). We assign at least two raters per work and aggregate scores by weighted averaging to reduce subjectivity. All identifying metadata (names, IDs, school-specific tags) are removed during preprocessing.

B. Data Analysis and Processing

Scores concentrate in the mid-to-high range, as shown in Figure 2a. This bias is common in educational assessment, where teachers tend to avoid assigning extremely low scores and exam prompts limit variance in foundational skills. To stabilize supervision while preserving educational meaning, we discretize raw scores in [0,100] into five fixed grades—F [0,60), D [60,70), C [70,80), B [80,90), A [90,100]—and map them to ordinal labels $\{1, \ldots, 5\}$ in ascending order. We retain both raw scores and discretized grades to support regression as well as ordinal or ranking-based formulations. As shown in Fig. 2b, the resulting grade distribution is naturally imbalanced, with tail categories such as F being underrepresented. We explicitly report this distribution and mitigate majority-class bias during training through class-aware sampling and class-balanced ordinal learning, as detailed in Sec. V.

Domain	Dataset	Number of Images	Text / Intent Annotations	Raters
Photo	AVA [1]	255,530	×	Crowdsourced
	AADB [8]	10,000	×	Invited raters
	PCCD [9]	4,235	×	Crowdsourced
	SPAQ [10]	11,125	×	Crowdsourced
	TAD66K [11]	66,327	×	Crowdsourced
	ICAA17K [12]	17,726	×	Crowdsourced
General Art	MART [13]	500	×	Invited raters
	JenAesthetics [4]	1,268	×	Invited raters
	VAPS [5]	999	×	Invited raters
	BAID [14]	60,337	×	Crowdsourced
	APDD [6]	4,985	×	Art experts
Child Art	AACP [7]	1,200	×	Art experts
Middle-School Art	**AEM45K (Ours)**	**44,573**	✓ (7,360) subset	**Art experts & experienced art teachers**

TABLE I: Comparison of AEM45K with representative photo and art aesthetics datasets. AEM45K targets middle-school drawings with multi-attribute grades, and includes a **verified subset of 7,360 samples** with student-written descriptions.

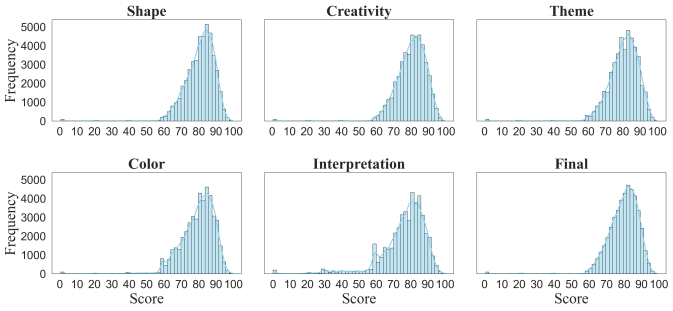

Fig. 2: AEM45K score statistics. Subfigure (a) shows the raw score distributions, and subfigure (b) shows the grade distributions after discretization.

C. Text Descriptions and Verified Multimodal Subset

In middle-school exams, some students provide short notes describing intent, story, or theme. These descriptions are particularly informative for semantics-heavy rubric dimensions (e.g., *Theme* and *Interpretation*), where purely visual evidence can be ambiguous. We therefore treat text as an **optional yet valuable modality**.

Descriptions are often embedded directly in the drawing image, as shown in Fig. 3. We extract text using PaddleOCR [23] and apply lightweight LLM-based correction to fix obvious

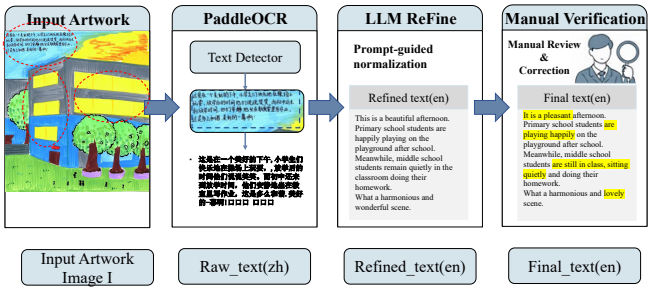

Fig. 3: Pipeline for extracting and verifying student-written descriptions. Text is first extracted via OCR,then normalized using a lightweight LLM, and finally verified by human reviewers to preserve fidelity to the original intent.

OCR noise without altering, rewriting, or adding content. We then conduct human verification to ensure fidelity to the original student writing. This pipeline yields a verified multimodal subset of **7,360** samples with aligned (image, text, label) triplets.

D. Dataset Usage Protocols

We define two practical evaluation protocols to clarify data scale under different modality assumptions:

- **Image-only:** all **44,573** drawings with rubric-aligned labels, evaluating methods that only use images.
- **Multimodal:** the verified-description subset of **7,360** samples, evaluating methods that leverage both image and text.

IV. Method

We propose the **Multimodal Aesthetic Dimension-aware Evaluation Network (MADENet)** for rubric-aligned, multi-attribute aesthetics assessment in middle-school art education (Fig. 4).

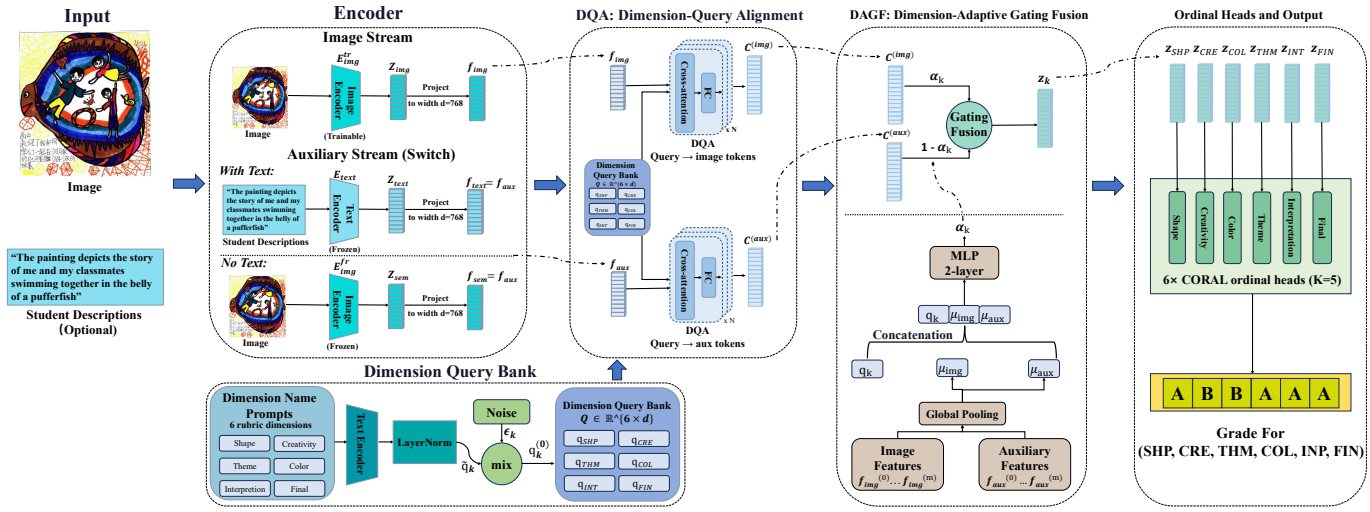

Fig. 4: Overall architecture of **MADENet**. A trainable visual encoder extracts image features, while an optional auxiliary stream provides text or semantic guidance depending on description availability. DQA retrieves dimension-specific evidence, DAGF adaptively fuses the two streams, and six CORAL ordinal heads predict rubric-aligned grades.

A. Encoders and Dimension Query Bank

MADENet adopts CLIP-style token encoders with two workflows. When a student description is available, the auxiliary stream uses the *frozen* CLIP text encoder; otherwise, it falls back to a *frozen semantic vision stream* extracted from a frozen CLIP ViT visual encoder to provide semantic-anchored tokens.

Given an input drawing I, a *trainable* visual encoder E_{img}^{tr} (initialized from CLIP ViT-L/14) and a *frozen* counterpart E_{img}^{fr} (same architecture) produce token sequences

$$\begin{aligned} Z_{img} &= E_{img}^{tr}(\mathrm{Embed}(I) + \mathrm{Pos}), \\ Z_{sem} &= E_{img}^{fr}(\mathrm{Embed}(I) + \mathrm{Pos}), \end{aligned} \tag{1}$$

where each $Z \in \mathbb{R}^{(M+1)\times d_{img}}$ contains a class token followed by M patch tokens. For text, we use the paired CLIP text encoder E_{text}:

$$Z_{text} = E_{text}(input_{ids}, attention_{mask}) \in \mathbb{R}^{(N+1)\times d_{text}}. \tag{2}$$

We project patch tokens to a shared width d=768 and discard the class token:

$$f_x = \mathrm{Proj}\big(Z_x[1{:}]\big), \quad x \in \{\mathrm{img}, \mathrm{sem}, \mathrm{text}\}, \tag{3}$$

with $f_{\mathrm{img}}, f_{\mathrm{sem}} \in \mathbb{R}^{M\times d}$ and $f_{\mathrm{text}} \in \mathbb{R}^{N\times d}$. The auxiliary stream is

$$f_{\mathrm{aux}} = \begin{cases} f_{\mathrm{text}}, & \text{description available,} \\ f_{\mathrm{sem}}, & \text{otherwise.} \end{cases} \tag{4}$$

For *dimension-aware* evidence gathering, MADENet uses six learnable queries, one per rubric dimension (SHAPE, CREATIVITY, THEME, COLOR, INTERPRETATION, FINAL), to retrieve attribute-specific evidence from token sequences.

We initialize each query from a dimension name prompt encoded by the CLIP text tower and add small isotropic noise to encourage exploration. Let $\mathcal{S} = \{$"Shape", "Creativity", "Theme", "Color", "Interpretation", "Final"$\}$. For each $s_k \in \mathcal{S}$,

$$\tilde{q}_k = \mathrm{LN}\Big(Z_{text}(s_k)_{[\mathrm{CLS}]}\, W_t\Big), \qquad q_k^{(0)} = (1-\beta)\tilde{q}_k + \beta\,\epsilon_k, \tag{5}$$

where $W_t \in \mathbb{R}^{d_t\times d}$ is learnable, $\epsilon_k \sim \mathcal{N}(\mathbf{0}, \frac{1}{d}\mathbf{I})$, and β is a small coefficient. During training, $\mathbf{Q}$ and W_t are jointly optimized to support downstream dimension-wise cross-attention and fusion.

B. Dimension-Query Alignment (DQA) and Dimension-Adaptive Gated Fusion (DAGF)

To achieve *attribute-level* alignment, MADENet introduce two components: *Dimension-Query Alignment (DQA)* and *Dimension-Adaptive Gated Fusion (DAGF)*.

DQA treats each learnable query q_k as a *per-attribute evidence retriever* that gathers cues from token sequences via query-to-token cross-attention. Given the projected features f_{img}, f_{aux} and the query bank $\mathbf{Q}$, we apply single-layer multi-head cross-attention to the two streams in parallel:

$$C^{(m)} = \mathrm{MHA}(\mathbf{Q},\, f_m), \quad m \in \{\mathrm{img}, \mathrm{aux}\}. \tag{6}$$

We compute masked global means μ_{img} and μ_{aux} as length-invariant summaries by averaging token embeddings.

DAGF adaptively balances the two evidence sources per dimension using the query and global context. For each k,

$$\begin{aligned} \alpha_k &= \sigma\Big(\mathrm{MLP}\big([\,q_k \,\|\, \mu_{img} \,\|\, \mu_{aux}\,]\big)\Big) \in [0,1], \\ z_k &= \alpha_k\, C_k^{(img)} + (1-\alpha_k)\, C_k^{(aux)} \;\in\; \mathbb{R}^d, \end{aligned} \tag{7}$$

where MLP is a lightweight two-layer gate and σ is the sigmoid.

DQA yields dimension-specific cues, while DAGF fuses them into a single, interpretable representation z_k per attribute.

C. Ordinal Heads and Objective

Each dimension uses a CORAL-style ordinal head producing $K-1$ logits (with $K=5$). Let $t_{k,r} = \mathbb{1}[y^{(k)} > r]$. The ordinal loss is

$$\mathcal{L}_k^{\text{ord}} = \sum_{r=1}^{K-1} \text{BCE}(\sigma(s_{k,r}),\, t_{k,r})\,, \qquad K = 5. \tag{8}$$

We obtain the ordinal probabilities $q_{k,r} = \sigma(s_{k,r})$ and follow the standard CORAL decoding to recover a 5-level class distribution $P_k(y)$, from which we compute the discrete prediction $\hat{y}^{(k)}$ and the expectation $\hat{m}_k$ used by the rubric-consistency term.

Although the rubric defines the final score as a weighted sum of attribute scores, real grading may include holistic adjustments (e.g., overall coherence, completeness, and narrative unity) that are not strictly captured by a linear combination. We therefore keep an explicit FINAL head and interpret its prediction as learning a non-linear *residual* over the rubric-induced baseline.

Let $\hat{m}_k$ denote the expectation implied by the CORAL probabilities for dimension k. We add a *Final–attribute consistency* term with stop-gradient on $\hat{m}_{\text{FINAL}}$:

$$\mathcal{L}^{\text{con}} = \Big\| \text{sg}(\hat{m}_{\text{FINAL}}) - \sum_{j=1}^{5} w_j\, \hat{m}_j \Big\|_2^2. \tag{9}$$

The overall training objective is

$$\mathcal{L} = \sum_{k \in \{\text{SHP,CRE,THM,COL,INT,FIN}\}} \mathcal{L}_k^{\text{ord}} \;+\; \lambda_{\text{con}}\, \mathcal{L}^{\text{con}}, \tag{10}$$

where λ_{con} controls the strength of rubric consistency.

V. EXPERIMENTS

A. Implementation Details

We compare MADENet with open-source SOTA aesthetics models. To match our rubric-aligned multi-attribute setting, we replace each baseline classifier with a 5-way ordinal head and instantiate six heads (five attributes + FINAL), while keeping all other components and training recipes unchanged.

We evaluate under two realistic protocols with the same overall setup. **Protocol-A (Image-only):** text is disabled; MADENet uses a fine-tunable E_{img}^{tr} and a *frozen* semantic visual auxiliary stream E_{img}^{fr}. **Protocol-B (Multimodal):** training/evaluation is restricted to the verified text subset; the auxiliary stream switches to the *frozen* text encoder E_{text}.

Due to the natural class imbalance after discretization (especially the tail grade F), we adopt ordinal learning and class-aware sampling to reduce majority-class bias. The model is trained for 50 epochs using Adam with a batch size of 64, and early stopping is applied with a patience of 10 based on FINAL accuracy. Learning rates are set to 10^{-6} for E_{img} and 10^{-5} for projection/DQA/DAGF/heads, cosine decay to 10^{-7} with a 1-epoch warmup. We set λ_{con}=0.1 and keep E_{text} and E_{img}^{fr} frozen. We report per-dimension Top-1 accuracy. All experiments run on a single RTX 4090.

Method	SHP	CRE	THM	COL	INP	FIN
AADB [8]	55.1	54.1	54.3	52.8	45.9	57.3
A-Lamp [2]	54.1	53.8	54.0	52.3	44.2	56.4
NIMA [3]	55.0	54.0	53.4	53.9	45.4	57.2
TANet [11]	55.2	54.0	55.3	51.7	44.7	57.2
EAT [24]	55.7	55.1	55.1	54.3	47.2	58.5
AesClip [18]	57.6	57.7	57.1	57.0	51.6	60.8
AesMamba-V [21]	58.6	58.3	57.0	57.5	51.7	61.7
MADENet (ours)	**59.5**	**59.3**	**58.4**	**58.1**	**53.2**	**63.1**

(a) Image-only protocol.

Method	SHP	CRE	THM	COL	INP	FIN
MSCAN [19]	60.4	59.8	59.4	58.5	55.1	62.5
CDCM-Net [20]	60.3	60.1	59.8	58.7	55.2	62.8
AesMamba-M [21]	61.8	61.3	60.9	60.0	56.7	64.9
MADENet (ours)	**63.6**	**62.7**	**62.2**	**62.4**	**58.8**	**67.3**

(b) Multimodal protocol.

TABLE II: Results on **AEM45k** under two protocols.

Variant	FIN
MADENet (full; DQA & DAGF)†	**67.3**
MADENet (no text; DQA & DAGF)‡	63.1
w/o DQA (no query-to-token)†	64.6
no DAGF (concat / fixed late fuse)†	65.1
shared query (1Q for all dimensions)†	65.9

TABLE III: Ablations on AEM45k, *Final* head only. †With text. ‡Without text.

B. Main Results

Tables IIa and IIb summarize results on **AEM45k**.

Image-only. MADENet attains the best accuracy on all six heads, outperforming the strongest baseline (AesMamba-V) by **+0.9/+1.0/+1.3/+0.6/+1.5/+1.4** pp on SHP/CRE/THM/COL/INP/FIN, respectively. The gains are not merely from a stronger backbone: our dimension-wise evidence retrieval (*DQA*) and adaptive fusion (*DAGF*) effectively balance the fine-tunable E_{img} with the *frozen* semantic auxiliary stream E_{img}^{fr}, yielding larger improvements on semantics-heavy dimensions (THM/INP) where purely low-level cues are insufficient.

Multimodal. On the text-available subset, MADENet likewise leads all baselines, surpassing AesMamba-M by **+1.8/+1.4/+1.3/+2.4/+2.1/+2.4** pp on SHP/CRE/THM/COL/INP/FIN. Notably, the largest gains occur on **Color**, **Interpretation**, and **Final**, indicating that intent-aware language cues provide complementary evidence and that MADENet performs more effective *dimension-specific* cross-modal grounding than generic early/late/attention fusion. Finally, directly predicting FINAL is empirically beneficial under both protocols, supporting our residual modeling of holistic grading beyond a fixed weighted sum of attributes.

C. Ablation Studies

Table III isolates the contribution of key components on the FINAL head. Overall, the improvements of MADENet do

not stem from generic multimodal fusion: removing either **dimension-wise retrieval** (**DQA**) or **adaptive gating** (**DAGF**) leads to clear degradation, validating the necessity of our design.

With text, the full model reaches **67.3**%, outperforming the no-text variant by **+4.2** pp, which quantifies the added value of student descriptions. Among components, removing **DQA** yields the largest drop (**–2.7** pp), indicating that attribute-specific query-to-token alignment is crucial for gathering discriminative evidence. Removing **DAGF** further hurts (**–2.2** pp), showing that per-dimension modality weighting is important under varying modality availability and uncertainty. Using a **shared query** also degrades performance (**–1.4** pp), confirming that different rubric dimensions require distinct evidence patterns.

VI. Conclusion and Further Works

We introduced **AEM45K**, a dataset of 44,573 middle-school exam drawings annotated along five rubric dimensions, including a verified image–text subset of 7,360 samples. We further proposed **MADENet**, a dimension-aware framework that integrates fine-tunable CLIP encoders with learnable dimension queries via query-to-token cross-attention. MADENet surpasses strong baselines, especially on semantics-heavy attributes and the overall grade, establishing a reproducible benchmark for educational aesthetic assessment. Current limitations include the exam-centric nature of the data, score discretization, partial text coverage with OCR noise, and potential cultural specificity. Future work will broaden coverage, enrich supervision, strengthen models, and deliver rubric-aligned, explainable feedback with fairness and calibration metrics.

VII. Acknowledgment

This work is supported by the Solfeggio ear training intelligent robot and cloud platform research and development project for music education (No.2024CXY0102), the public technology service platform project of Xiamen City (No.3502Z20231043) and Fujian Provincial Science and Technology Major Project (No.2024HZ022003), the 3D visualization digital twin integrated control system (No.2023CXY0111)

References

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

[2] Shuang Ma, Jing Liu, and Chang Wen Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4535–4544.

[3] Hossein Talebi and Peyman Milanfar, "Nima: Neural image assessment," *IEEE transactions on image processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[4] Seyed Ali Amirshahi, Gregor Uwe Hayn-Leichsenring, Joachim Denzler, and Christoph Redies, "Jenaesthetics subjective dataset: analyzing paintings by subjective scores," in *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13*. Springer, 2015, pp. 3–19.

[5] Anna Fekete, Matthew Pelowski, Eva Specker, David Brieber, Raphael Rosenberg, and Helmut Leder, "The vienna art picture system (vaps): A data set of 999 paintings and subjective ratings for art and aesthetics research.," *Psychology of Aesthetics, Creativity, and the Arts*, vol. 17, no. 5, pp. 660, 2023.

[6] Jin Xin, Qiao Qianqian, Lu Yi, Wang Huaye, Gao Shan, Huang Heng, and Li Guangdong, "Paintings and drawings aesthetics assessment with rich attributes for various artistic categories," in *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2024.

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

[8] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes, "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*. Springer, 2016, pp. 662–679.

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

[10] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang, "Perceptual quality assessment of smartphone photography," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3677–3686.

[11] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming, "Rethinking image aesthetics assessment: Models, datasets and benchmarks.," in *IJCAI*, 2022, pp. 942–948.

[12] Shuai He, Anlong Ming, Yaqi Li, Jinyuan Sun, ShunTian Zheng, and Huadong Ma, "Thinking image color aesthetics assessment: Models, datasets and benchmarks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21838–21847.

[13] Victoria Yanulevskaya, Jasper Uijlings, Elia Bruni, Andreza Sartori, Elisa Zamboni, Francesca Bacci, David Melcher, and Nicu Sebe, "In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings," in *Proceedings of the 20th ACM international conference on multimedia*, 2012, pp. 349–358.

[14] Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, and Paul L Rosin, "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*, 2023, pp. 22388–22397.

[15] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 457–466.

[16] Mingzhe Guo, Zhipeng Zhang, Heng Fan, Liping Jing, Yilin Lyu, Bing Li, and Weiming Hu, "Rethinking image aesthetics assessment: Models, datasets and benchmarks," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, Jul 2022.

[17] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9375–9383.

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

[19] Xiaodan Zhang, Xinbo Gao, Lihuo He, and Wen Lu, "Mscan: Multimodal self-and-collaborative attention network for image aesthetic prediction tasks," *Neurocomputing*, vol. 430, pp. 14–23, 2021.

[20] Xiaodan Zhang, Yuan Xiao, Jinye Peng, Xinbo Gao, and Bo Hu, "Confidence-based dynamic cross-modal memory network for image aesthetic assessment," *Pattern Recognition*, vol. 149, pp. 110227, 2024.

[21] Fei Gao, Yuhao Lin, Jiaqi Shi, Maoying Qiao, and Nannan Wang, "Aesmamba: universal image aesthetic assessment with state space models," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 7444–7453.

[22] Zhuo Li, Xingao Yan, Xuebin Wei, and Feng Shao, "Iaaclip: Image aesthetics assessment via clip," *Electronics*, vol. 14, no. 7, pp. 1425, 2025.

[23] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al., "Paddleocr 3.0 technical report," *arXiv preprint arXiv:2507.05595*, 2025.

[24] Shuai He, Anlong Ming, Shuntian Zheng, Haobin Zhong, and Huadong Ma, "Eat: An enhancer for aesthetics-oriented transformers," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1023–1032.