

Context-Aware Fine-Grained Ranking of Art Exam Works Under Shared Evaluation Conditions

Jiahao Li

Hubei University of Technology

Yingjie Zhang

Hong Kong Polytechnic University

Zhongwei Huang

Hubei University of Technology

Haoze Chen

Hubei University of Technology

Zongming Tan

University of Melbourne

Baohua Tan

tbh@hbut.edu.cn

Hubei University of Technology

Chao Chen

University of Nottingham

Research Article

Keywords: Image aesthetic assessment, Art exam, Group-conditional ranking, Fine-grained assessment

Posted Date: September 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10164848/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Context-Aware Fine-Grained Ranking of Art Exam Works Under Shared Evaluation Conditions

Jiahao Li^{1,2,3†}, Yingjie Zhang^{4,5†}, Zhongwei Huang^{6†},
Haoze Chen^{5†}, Zongming Tan^{7†}, Baohua Tan^{1,2,3*}, Chao Chen^{8*}

¹School of Science (School of Chip Industry), Hubei University of Technology, Wuhan, 430068, China.

²National “111 Research Center” Microelectronics and Integrated Circuits, Hubei University of Technology, Wuhan, China.

³Faculty of Engineering, University of Nottingham, Nottingham, NG7 2RD, United Kingdom.

⁴Department of Applied Physics, The Hong Kong Polytechnic University, Hong Kong, China.

⁵Detroit Green Technology Institute, Hubei University of Technology, Wuhan, 430068, China.

⁶School of Computer Science and Artificial Intelligence, Hubei University of Technology, Wuhan, 430068, China.

⁷Faculty of Science, University of Melbourne, Victoria, 3010, Australia.

⁸School of Computer Science, University of Nottingham, Nottingham, NG8 1BB, United Kingdom.

*Corresponding author(s). E-mail(s): tbh@hbut.edu.cn;
Baohua.Tan@nottingham.ac.uk; chao.chen@nottingham.ac.uk;

[†]These authors contributed equally to this work.

Abstract

Image aesthetic assessment has made substantial progress, yet most existing approaches evaluate individual images independently and assume that scores are comparable across broad visual domains. This assumption is less suitable for art examination scenarios, where works are judged within the same proposition, batch and assessment context, and where the key challenge is to recover fine-grained relative quality differences among visually and semantically similar submissions. To address this problem, this paper constructs the Art Exam Dataset, comprising 3360 works from 93 examination themes across sketch, quick

sketch and colour painting. The dataset is organised by examination groups, enabling the study of intra-group ranking under shared evaluation conditions. Building on this setting, the paper proposes a group-conditioned assessment framework that augments single-image aesthetic prediction with dual-path reference encoding. One path models the relationship between each work and the group mean reference to capture context-aware deviations, while the other emphasises residual-relative differences between individual samples and their group context. An adaptive routing fusion module then combines the baseline prediction with the context-aware and residual-relative correction candidates through learned fusion weights. Experiments conducted under a group-disjoint protocol show that the proposed framework outperforms representative image aesthetic assessment baselines in terms of SRCC, PLCC and pairwise ranking accuracy. Ablation studies further indicate that separating context-aware and residual-relative correction paths, together with sample-specific routing, contributes to improved ranking consistency. The work provides a dataset, task formulation and methodological reference for intelligent assessment of art exam works, while also highlighting the need for richer expert annotations and more transparent evaluation protocols in future research.

Keywords: Image aesthetic assessment, Art exam, Group-conditional ranking, Fine-grained assessment

1 Introduction

With the continuous advancement of image style transfer techniques [1–3] and the rapid development of artificial intelligence-generated content (AIGC) [4], users can now generate artistic images in diverse styles more conveniently [5, 6]. Related generated works have also gained substantial attention and even won awards in art competitions [7]. This trend has increased the demand for reliable quality assessment methods for artistic images. Automated or semi-automated aesthetic assessment methods can serve as auxiliary tools for manual assessment, and can be used for preliminary work screening, intra-group quality ranking, and evaluation result analysis, thereby providing references for large-scale work management, teaching analysis, and training. Unlike open-domain artistic images, art exam works usually have stronger task constraints and clearer evaluation dimensions. For example, sketch, quick sketch, and color painting works are often created under the same theme, similar composition requirements, or unified proposition, thus requiring models to identify fine-grained quality differences under high semantic consistency.

At present, studies on the assessment of art exam works are still relatively limited. Some existing studies apply deep convolutional features to the classification and evaluation of sketch works, treating sketch assessment as a fine-grained classification problem with several levels, such as excellent, good, medium, and poor [8, 9]. These methods divide the quality of works into a few coarse-grained levels and rely on limited supervised data. Wang et al. [10] proposed the large-scale sketch portrait dataset SHPD and the evaluation network SPAAN, using degraded aesthetic operations and

self-supervised learning to enable the model to perceive aesthetic features. These works indicate that the assessment of specific types of artworks has begun to attract attention. However, in art exam scenarios, works are often assessed as groups under the same theme, and the assessment of these works is usually not conducted independently. Instead, it is formed through intra-group comparison under shared evaluation conditions, resulting in a relative order. Therefore, several deficiencies remain in this scenario. Existing IAA datasets rarely consider art exam-related works. In the AVA, AADB, and BAID datasets [11–13], sketch and exam-related artworks are limited [10]. Traditional aesthetic datasets usually contain highly diverse visual content, and samples that can be compared under the same semantic condition are relatively limited. Moreover, most evaluation processes assess single works independently, which ignores the differentiated information within the same group and makes it difficult for models to capture fine-grained differences among works under the same theme.



Fig. 1 Sample groups from the AED dataset. Each row corresponds to one art exam category, and samples within the same row come from the same group with a shared exam context.

To address the insufficient data support for art-exam work assessment and the limited adaptability of existing image aesthetic assessment methods to this scenario, we first formulate art-exam work assessment as a group-conditioned aesthetic assessment problem and construct the Art Exam Dataset (AED). AED contains 3,360 art-exam works from 93 exam themes, covering three major categories: sketch, quick sketch, and color painting. Unlike general aesthetic datasets that mainly support independent image-level assessment, AED preserves the natural group structure of real art exams. As shown in Fig. 1, works are organized by exam themes, and samples within the same group share the same proposition, category, and shared assessment context.

Based on this group-organized benchmark, we further define an evaluation and training formulation for art-exam aesthetic assessment. Specifically, we adopt a group-disjoint evaluation protocol, where works from the same exam theme are not split into both the training and test sets. This protocol allows the evaluation to reflect the model’s generalization ability to unseen propositions and assessment contexts. In addition, because different exam themes involve different propositions, their absolute scores are not assumed to be directly comparable across groups. Therefore, we use group-normalized mid-rank targets to transform the intra-group score order into stable supervision aligned with the intra-group ranking objective. These designs make AED a consistent benchmark setting for future research on group-conditioned aesthetic assessment.

This formulation further motivates a model design that assesses each work with reference to its group context. Accordingly, we propose the Dual-path Reference Encoder with Adaptive Routing Network (DPAR-Net) as an initial reference architecture. Instead of treating each work as an isolated image, DPAR-Net uses the shared assessment context within each group. The Group-Mean Reference Path captures common group-level information under the same proposition, while the Residual-Relative Path describes sample-wise deviations from that group context. The Adaptive Routing Fusion module combines the baseline prediction with the context-aware and residual-relative correction candidates to obtain the final group-wise prediction. In this way, DPAR-Net provides a concrete implementation for the proposed group-conditioned intra-group ranking task.

In summary, the main contributions of our work can be categorized into three aspects:

- (1) We formulate art-exam work assessment as a group-conditioned aesthetic assessment problem and construct AED, a group-organized benchmark containing 3,360 works from 93 exam themes across sketch, quick sketch, and color painting. AED preserves the natural group structure of art exams and supports fine-grained intra-group aesthetic comparison under shared propositions, categories, and assessment contexts.
- (2) We design a group-disjoint evaluation protocol for AED, where works from the same exam theme are kept in the same split to test generalization to unseen propositions. To match the relative nature of art-exam assessment, we evaluate models by intra-group ranking and train them with group-normalized mid-rank targets derived from the original scores.
- (3) We propose DPAR-Net as a concrete implementation of group-conditioned aesthetic assessment and validate it through systematic experiments on AED. The results show that DPAR-Net achieves better intra-group ranking correlation and pairwise comparison accuracy than representative image aesthetic assessment baselines.

2 Related Work

2.1 Image Aesthetic Assessment

Early image aesthetic assessment mainly relied on manually designed features combined with traditional machine learning methods [14, 15]. These methods evaluated

image quality by extracting aesthetic-related features such as composition, color, and texture. However, they had clear limitations in feature representation ability and generalization performance. With the development of deep neural networks and large-scale supervised datasets, image aesthetic assessment has achieved substantial progress. Talebi et al. [16] changed the prediction target from a single aesthetic score to a score distribution based on the AVA dataset, thereby taking evaluation uncertainty into account. Hosu et al. [17] designed a multi-scale pooling fusion model to extract multi-level aesthetic features, improving performance on the AVA dataset. He et al. [18] proposed an evaluation network based on thematic style, indicating that images with different styles may require different evaluation criteria, and designed the TAD66K dataset containing images of different styles. To address the insufficient research on artistic image styles, Yi et al. [13] proposed the large-scale artistic image dataset BAID, and designed an unsupervised learning branch to capture aesthetic differences among images. Behrad et al. [19] designed a tagging strategy that retains composition, high-resolution content, aspect ratio, and multi-scale context, and assigned high-resolution tags to visually important regions, achieving strong performance on multiple datasets. In addition, recent work has begun to explore the effectiveness of new visual architectures such as state space models in general IAA [20], while traditional aesthetic cues such as color composition and spatial organization have also been shown to be valuable for image aesthetic assessment [21]. With the rapid development of vision-language models, multimodal aesthetic assessment methods are also evolving [22–27]. These methods obtain text-image representations through MLLMs or combine user comments with visual information, enabling the model to make progress toward interpretability in this field.

Although existing methods have made significant progress in single-image aesthetic modeling, most of them assume that image labels can be compared globally. This assumption is usually applicable to general IAA evaluation, but it faces challenges in art exam scenarios. Specifically, scoring scales may vary across different propositions and batches, while works under the same proposition share a natural evaluation context. Therefore, we address the problem of how to shift from independent absolute assessment of single images to group-relative modeling within shared evaluation conditions.

2.2 Relational Modeling and Conditional Assessment

Existing studies have shown that human aesthetic judgment is not always made in isolation. Comparative relations, reference samples, and contextual information often help form more stable and consistent evaluations. Based on this observation, some works have begun to improve aesthetic prediction performance from the perspectives of relational modeling and conditional assessment [28–35].

Ma et al. [28] adaptively integrated local and global patch fusion to generate stronger aesthetic predictions. Liu et al. [29] constructed multiple patch regions and inferred regional relationships, thereby improving the performance of aesthetic assessment. She et al. [30] proposed a hierarchical layout-aware graph convolutional network (HLA-GCN) for unified image aesthetic assessment. By constructing graph structures in both coordinate space and latent space and performing graph reasoning, HLA-GCN

explicitly models layout relationships and spatial organization information between image regions. These methods indicate that aesthetic judgment not only depends on local textures or global semantics, but is also closely related to the organizational relationships among internal elements of the image. However, their focus still mainly remains within a single image. Yang et al. [31] constructed a reference image set through semantic retrieval, and achieved multi-image joint learning by combining graph relationship reasoning. Shi et al. [32] then introduced a multi-reference learning strategy with semantic and style consistency for artistic image scenarios, jointly modeling the relationship between the target image and reference images. These works collectively illustrate that cross-image comparison and reference learning can provide discriminative information beyond independent modeling of a single image. However, the reference samples in the above methods are mostly constructed a posteriori through retrieval or pairing strategies, while the art exam scenarios studied in our work naturally form groups of works based on the same proposition and shared evaluation conditions.

In relation to reference learning, ranking modeling is also often used to depict relative quality relationships in aesthetic assessment. Lee et al. [36] modeled image aesthetic assessment as a pairwise comparison problem, indicating that pairwise comparison can be applied simultaneously to aesthetic score prediction and preference relationship modeling. Zhou et al. [37] combined regression learning with pairwise ranking for personalized image aesthetic assessment. Recently, Cheng et al. [38] constructed a pairwise comparison task for image color temperature aesthetic assessment, indicating that relative ranking is helpful for modeling fine-grained aesthetic differences. This also provides a basis for constructing supervisory signals from the perspective of intra-group relative relationships.

Apart from relational modeling, existing studies have shown that images of different themes, styles, or content types often correspond to different evaluation criteria. Therefore, a fixed and uniform scoring function is not always optimal. Tang et al. [39] designed an attention module to fully utilize the complementary information between photographic visual attribute features and artistic aesthetic features. The TANet work [18] also indicates that theme or style perception can enhance the adaptability of image aesthetic assessment, and shows that the evaluation function itself has certain contextual dependence. However, these conditional assessment methods mainly focus on themes, styles, attributes, or user preferences, while the group-organized evaluation context in art exam scenarios has not been sufficiently explored.

2.3 Image Aesthetic Assessment Datasets

The development of public datasets has significantly advanced image aesthetic assessment research. Early works mainly relied on large-scale general benchmarks such as AVA [11], which contains over 250,000 images and provides multi-dimensional annotations such as aesthetic ratings, semantic categories, and photographic styles, laying the foundation for aesthetic classification, score regression, and score distribution prediction. Subsequently, AADB [12] further introduced photographic attributes such as composition and color beyond the overall aesthetic rating, while TAD66K [18]

expanded the evaluation scenario from the perspective of themes, enabling models to learn aesthetic rules based on perceptual cues and thematic differences.

As research subjects gradually expanded from natural images to artistic images and paintings, related datasets began to focus more on artistic expression itself. As a large-scale aesthetic assessment dataset for artistic images, BAID [13] expanded the research foundation in the context of artistic works. APDD [40] further introduced multi-category paintings and multi-dimensional aesthetic attribute annotations, enabling models to learn more domain-specific artistic expression features. At the same time, AACP [41] and SHPD [10] respectively targeted more specific task scenarios such as children’s paintings and sketch portraits, indicating that image aesthetic assessment is gradually moving from general modeling toward specialization and professionalism. Recent work [42] has also begun to focus on fine-grained aesthetic differences in natural images, cropped images, and AIGC images, further demonstrating that assessment problems in fine-grained scenarios remain highly challenging [43].

However, most of these datasets still focus primarily on predicting the absolute aesthetic quality of a single image. There remains a research gap in aesthetic data organized as groups of works under a unified evaluation context. The art exam scenario naturally possesses this condition. For the same proposition and the same batch of works, the content constraints and scoring criteria are highly consistent. Fine-grained comparison within the group is the core issue in this scenario, and the original scores of different groups may not share a unified absolute scale. This motivates the construction of AED, a group-organized art exam dataset covering sketch, quick sketch, and color painting. Details are provided in Section 3.1.

3 Methods

This section first introduces the construction process of the AED dataset in Section 3.1, and then formally defines the group-conditional fine-grained assessment task for art exam works in Section 3.2. Different from traditional single-image aesthetic assessment, we focus on modeling relative quality within groups under shared evaluation conditions, that is, restoring the relative ranking relationship between works within the same exam theme and batch.

Based on this, we propose DPAR-Net, whose overall structure is shown in Fig. 2(a). DPAR-Net is designed for the group-organized data format of art exam works and explicitly utilizes group-context information for fine-grained quality modeling. Specifically, the model first extracts the feature z_i of each sample using a shared backbone, and a single-image prediction head maps z_i to the baseline prediction b_i . Subsequently, the Dual-path Reference Encoder introduced in Section 3.3 uses the group features to generate two correction candidates: a context-aware correction c_i and a residual-relative correction r_i . Finally, the Adaptive Routing Fusion module introduced in Section 3.4 combines the baseline prediction with the two correction candidates to obtain the final group-wise prediction.

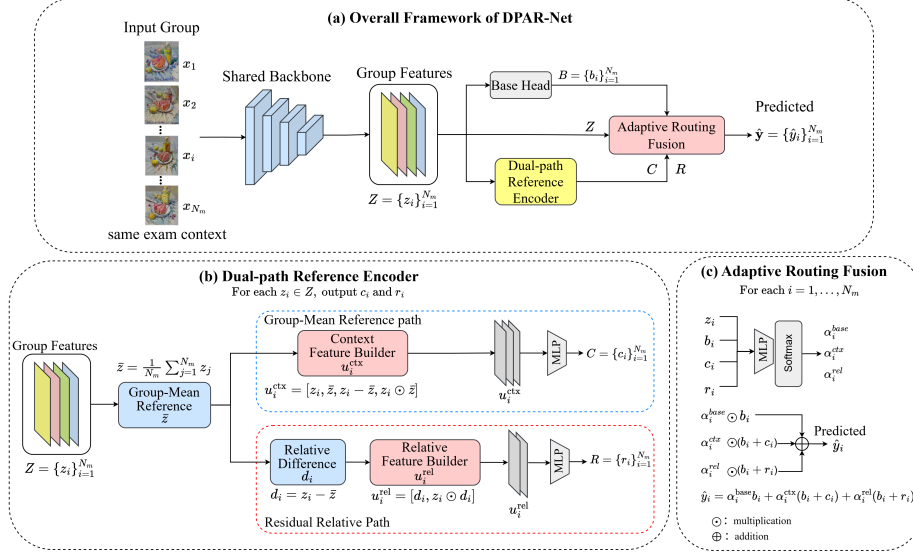


Fig. 2 Overview of DPAR-Net. (a) Overall framework of the proposed method. Given an input group from the same exam context, a shared backbone extracts group features, the base head predicts baseline scores, and the Dual-path Reference Encoder estimates context-aware and residual-relative correction candidates. The Adaptive Routing Fusion module then combines the baseline prediction with these correction candidates to produce the final group-wise prediction. (b) Structure of the Dual-path Reference Encoder, which contains a Group-Mean Reference Path and a Residual Relative Path. (c) Adaptive Routing Fusion, which combines the baseline prediction with the context-aware and residual-relative correction candidates to produce the final prediction.

3.1 Art Exam Dataset

To study the quality assessment of artworks in art exam scenarios, an Art Exam Dataset (AED) specifically designed for the art exam scenario was constructed. AED was constructed from archived mock art-examination records collected by a collaborating institution. These mock examinations follow a format similar to regional joint art examinations, where multiple candidates complete their works under the same proposition, category, and assessment conditions. The dataset contains 3,360 works from 93 examination themes, covering three categories: sketch, quick sketch, and color painting. In AED, each examination theme is treated as one group. Fig. 3 and Table 1 show the sample number distribution and detailed statistical information of the 93 groups in AED.

The dataset is permitted for non-commercial academic research. Since all identity information has been anonymized and the data come from archived mock examinations, we cannot reliably determine whether works from the same student appear across different groups. The original images have varying resolutions and aspect ratios because they were digitized from different archived examination records. Some images are horizontally oriented, while others are vertically oriented. The image widths range from 363 to 1707 pixels, and the heights range from 455 to 1897 pixels, with a median

size of approximately 1235×1376 pixels. However, images within the same examination group usually have similar acquisition conditions and comparable resolutions. During model training, all images are resized to 224×224 pixels following the same preprocessing protocol. Finally, we cleaned and rechecked the collected data to ensure that the samples in each group were valid, and then constructed AED.

Each work is associated with a final score generated by the examination assessment process. These final scores are the available assessment outcomes of the mock art examination, rather than objective measurements of artistic quality. Since judge-level rating records, expert comments, detailed rubrics, and score aggregation procedures are unavailable, we do not make assumptions about the specific score formation mechanism. In this study, the final scores are used to derive the relative ordering embedded in the art-exam assessment process within each group. Therefore, AED is designed to support the study of assessment-derived intra-group ranking under shared examination conditions, rather than absolute prediction of artistic quality across all groups. Finally, AED is permitted for non-commercial academic research and can be publicly released.

Table 1 Statistics of AED by category.

Category	Images	Groups	Group size mean	Min./Max. group size	Score mean $\pm$ std
Sketch	1066	33	32.3	9/72	83.88 ± 13.42
Quick sketch	1045	29	36.0	6/72	82.17 ± 13.68
Color	1249	31	40.3	7/132	86.71 ± 11.19

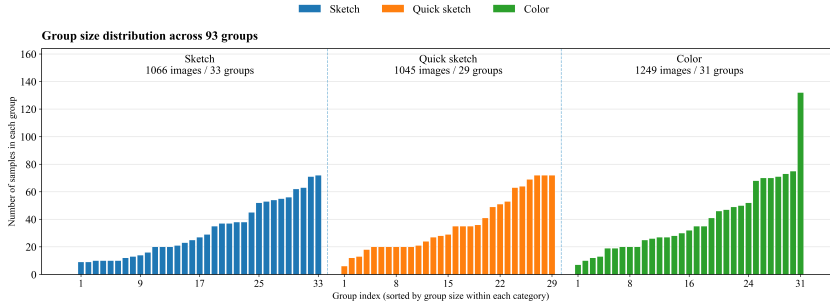


Fig. 3 Group size distribution of the AED dataset. The 93 groups are divided into three art exam categories: sketch, quick sketch, and color painting. Bars indicate the number of samples in each group, and groups are sorted by group size within each category.

3.2 Problem formulation

Unlike traditional image aesthetic assessment tasks, we do not focus on independently scoring individual images on a unified scale across the entire dataset. Instead, we

address group-conditional fine-grained assessment for art exam works. In AED, each group corresponds to a set of works from the same exam batch and the same proposition. Since works within the same group are highly consistent in subject matter, composition style, and semantic expression, the samples within the group have natural comparability. In contrast, different groups often correspond to different propositions and evaluation contexts, and their score labels may not share a fully consistent comparison scale. Therefore, the original scores across different groups cannot be strictly compared directly. Let AED be represented as:

$$\mathcal{D} = \{\mathcal{G}_m\}_{m=1}^M, \quad \mathcal{G}_m = \{(x_i^{(m)}, s_i^{(m)})\}_{i=1}^{N_m}, \quad (1)$$

where $\mathcal{G}_m$ denotes the m -th group, $x_i^{(m)}$ denotes the i -th image in this group, $s_i^{(m)}$ denotes the original score label of the i -th image, and N_m is the number of samples in this group. Since we do not assume a strictly consistent global scoring scale across different groups, the original score $s_i^{(m)}$ is mainly used as the source of intra-group relative quality relationships, rather than as an absolute supervisory signal that can be directly aligned across groups. During training, we further convert $s_i^{(m)}$ into an intra-group relative supervision target, as defined in Section 3.5.

Given an input group $\mathcal{G}_m$, the goal of the model is to predict a set of relative quality scores:

$$\hat{\mathbf{y}}^{(m)} = \{\hat{y}_i^{(m)}\}_{i=1}^{N_m}. \quad (2)$$

Although the final output of the model is still represented as scalar scores, for the AED task, these scores are not intended to recover absolute exam scores that are directly comparable across groups. Instead, they are used to recover the relative ranking relationships and local hierarchical structure among works within the same shared evaluation conditions. Based on this setting, we regard each group as a local evaluation context and model group-context correction to learn fine-grained quality differences among works. For clarity, the main notation used in the following model description is summarized in Table 2.

Table 2 Main notation used in the DPAR-Net formulation.

Symbol	Meaning
G_m	The m -th artwork group under the same exam proposition
$x_i^{(m)}$	The i -th artwork in group G_m
$s_i^{(m)}$	Original final exam score of $x_i^{(m)}$
$t_i^{(m)}$	Group-normalized mid-rank target
z_i	Visual feature of artwork x_i
$\bar{z}$	Group mean reference feature
b_i	Baseline single-image prediction
c_i	Context-aware correction from the Group-Mean Reference Path
r_i	Residual-relative correction from the Residual Relative Path
$\alpha_i^{\text{base}}, \alpha_i^{\text{ctx}}, \alpha_i^{\text{rel}}$	Routing weights for the three prediction paths
$\hat{y}_i$	Final predicted relative score

3.3 Dual-path Reference Encoder

To fully utilize the comparable relationships among works within the same group, we design a Dual-path Reference Encoder, as shown in Fig. 2(b). Given the group features extracted by the shared backbone, the encoder estimates two types of correction information for each sample: a context-aware correction and a residual-relative correction. The encoder consists of two paths, namely the Group-Mean Reference Path and the Residual Relative Path. For simplicity, the group superscript (m) is omitted in this subsection, and z_i denotes the feature of the i -th work in the current group.

The Group-Mean Reference Path aims to model each work with respect to the overall group context. Specifically, we first calculate the group mean reference $\bar{z}$ by averaging the features of all works in the group. The relation between each sample and the group reference is then represented in the learned feature space by combining the sample feature, the group reference, their residual difference, and their element-wise interaction. The context-aware correction c_i is obtained as:

$$\begin{aligned}\bar{z} &= \frac{1}{N_m} \sum_{j=1}^{N_m} z_j, \\ u_i^{\text{ctx}} &= [z_i, \bar{z}, z_i - \bar{z}, z_i \odot \bar{z}], \\ c_i &= \text{MLP}_{\text{ctx}}(u_i^{\text{ctx}}),\end{aligned}\tag{3}$$

where $\odot$ denotes element-wise multiplication. This path allows the correction head to compare the current sample with the group-level reference in the learned feature space.

In addition to the context-aware representation, we further introduce a Residual Relative Path to emphasize sample-wise deviation from the group context. This path isolates the residual deviation and constructs a compact residual-relative representation:

$$\begin{aligned}d_i &= z_i - \bar{z}, \\ u_i^{\text{rel}} &= [d_i, z_i \odot d_i], \\ r_i &= \text{MLP}_{\text{rel}}(u_i^{\text{rel}}),\end{aligned}\tag{4}$$

Here, d_i denotes the residual deviation between the current sample and the group mean reference, and r_i denotes the residual-relative correction predicted by the Residual Relative Path.

In summary, the Group-Mean Reference Path provides a context-aware correction by jointly modeling the sample feature and the group mean reference, while the Residual Relative Path provides a residual-relative correction by focusing on the sample deviation from the group context. These two correction candidates are then combined with the baseline prediction by the Adaptive Routing Fusion module for fine-grained intra-group ranking.

3.4 Adaptive Routing Fusion

The Dual-path Reference Encoder produces two correction candidates for each sample: the context-aware correction c_i and the residual-relative correction r_i . A straightforward way to use these candidates is to directly add them to the baseline prediction.

However, direct additive fusion fixes the coefficients of both correction terms and does not provide an explicit mechanism for learning their relative contribution. Another possible solution is to use an unconstrained MLP to regress the final prediction from the baseline output and the correction candidates, but such a fusion function does not explicitly preserve the baseline prediction as an anchor and does not provide directly interpretable correction weights.

Based on this consideration, we design an Adaptive Routing Fusion module, as shown in Fig. 2(c). The module keeps the prediction in a baseline-anchored residual form and learns sample-conditioned weights for the two correction candidates. For the i -th sample, we construct three candidate prediction paths: the baseline path b_i , the context-corrected path $b_i + c_i$, and the residual-corrected path $b_i + r_i$. The routing MLP takes z_i , b_i , c_i , and r_i as input, and generates normalized routing weights for these paths:

$$\begin{aligned} [\alpha_i^{\text{base}}, \alpha_i^{\text{ctx}}, \alpha_i^{\text{rel}}] &= \text{softmax}(\text{MLP}_{\text{route}}([z_i, b_i, c_i, r_i])), \\ \hat{y}_i &= \alpha_i^{\text{base}} b_i + \alpha_i^{\text{ctx}} (b_i + c_i) + \alpha_i^{\text{rel}} (b_i + r_i), \\ &= b_i + \alpha_i^{\text{ctx}} c_i + \alpha_i^{\text{rel}} r_i. \end{aligned} \quad (5)$$

where α_i^{base} , α_i^{ctx} , and α_i^{rel} denote the routing weights for the baseline path, context-corrected path, and residual-corrected path, respectively. The three weights are normalized by the softmax function and sum to 1.

This formulation keeps the final prediction anchored to the baseline output while allowing the two correction candidates to contribute through learned non-baseline routing weights. In this way, Adaptive Routing Fusion provides a structured residual fusion form in which c_i and r_i are introduced as weighted correction terms rather than directly added with fixed coefficients.

3.5 Training Objective

The goal of AED is to recover relative quality relationships among artworks within the same exam group, rather than to fit absolute scores that may not be directly comparable across different exam themes. Therefore, for each group, we convert the original final exam scores into group-normalised mid-rank targets and use them as the regression supervision.

Let the current training batch correspond to a complete group $G_m = \{(x_i^{(m)}, s_i^{(m)})\}_{i=1}^{N_m}$, where $s_i^{(m)}$ is the original expert score of the i -th sample. The samples are ranked within the group according to their raw scores in ascending order, and tied scores receive the average rank:

$$r_i^{(m)} = \text{MidRank}(s_i^{(m)}, \{s_j^{(m)}\}_{j=1}^{N_m}),$$

where $\text{MidRank}(\cdot)$ denotes the intra-group rank with average ranks assigned to tied scores.

Based on the mid-rank, the group-normalised target is defined as

$$t_i^{(m)} = \begin{cases} 0, & N_m = 1, \\ \frac{r_i^{(m)} - 1}{N_m - 1}, & N_m > 1, \end{cases} \quad t_i^{(m)} \in [0, 1].$$

Under this definition, $t_i^{(m)} = 0$ corresponds to the lowest-ranked artwork in the group, while $t_i^{(m)} = 1$ corresponds to the highest-ranked artwork. This target preserves intra-group relative quality and reduces the influence of score-scale differences across exam groups.

For an input image $x_i^{(m)}$, the model outputs a prediction $\hat{y}_i^{(m)}$, which is supervised by the group-normalised target. We use the Smooth L1 loss as the regression objective:

$$\mathcal{L}_{reg}^{(m)} = \frac{1}{N_m} \sum_{i=1}^{N_m} \text{SmoothL1} \left(\hat{y}_i^{(m)}, t_i^{(m)} \right).$$

Since each training batch corresponds to a complete group, this loss performs relative position regression within the group. This supervision is consistent with AED’s group-conditioned assessment setting, where evaluation is defined within the same exam proposition rather than across unrelated themes. The complete training and inference procedure of DPAR-Net is summarized in Algorithm 1.

Algorithm 1 Training and inference of DPAR-Net

Require: Grouped artworks $\mathcal{D} = \{G_m\}_{m=1}^M$, where $G_m = \{(x_i^{(m)}, s_i^{(m)})\}_{i=1}^{N_m}$

- 1: **for** each training group G_m **do**
- 2: Convert original scores $\{s_i^{(m)}\}_{i=1}^{N_m}$ into group-normalized mid-rank targets $\{t_i^{(m)}\}_{i=1}^{N_m}$
- 3: Extract feature z_i for each artwork $x_i^{(m)} \in G_m$
- 4: Compute the group mean reference $\bar{z} = \frac{1}{N_m} \sum_{j=1}^{N_m} z_j$
- 5: Predict baseline output b_i for each artwork
- 6: Predict context-aware correction c_i and residual-relative correction r_i
- 7: Compute routing weights $\alpha_i^{\text{base}}, \alpha_i^{\text{ctx}}, \alpha_i^{\text{rel}}$
- 8: Obtain final prediction $\hat{y}_i = \alpha_i^{\text{base}} b_i + \alpha_i^{\text{ctx}} (b_i + c_i) + \alpha_i^{\text{rel}} (b_i + r_i)$
- 9: Update model parameters using Smooth L1 loss between $\hat{y}_i$ and $t_i^{(m)}$
- 10: **end for**
- 11: **for** each test group G_m **do**
- 12: Extract group features and compute the group mean reference $\bar{z}$
- 13: Compute b_i, c_i, r_i , and routing weights for each artwork
- 14: Output $\hat{y}_i$ for intra-group ranking and evaluation
- 15: **end for**

4 Experiments

4.1 Experimental setup

To evaluate the effectiveness of the proposed method on the fine-grained quality assessment task for art-exam works, we conduct five-fold group-disjoint evaluation on AED. The five folds are constructed using a group-level greedy balancing strategy. First, all images are assigned to their corresponding exam groups, and each exam group is treated as an indivisible splitting unit. The fold assignment is then performed at the group level. For each category, namely Sketch, Quick Sketch, and Color Painting, exam groups are sorted by the number of images in descending order. The groups are sequentially assigned to the fold that currently has the smallest number of images in the same category. If multiple folds have the same category-specific image count, the group is assigned to the fold with the smaller total number of images. This procedure makes the folds approximately balanced in terms of total image number and category coverage while preserving the group-disjoint constraint. In each evaluation round, one fold is used as the test set, and the remaining four folds are used for training. The same hyperparameter configuration is used across all folds, and the test fold is used only for final evaluation rather than hyperparameter tuning. Table 3 reports the detailed split statistics.

Table 3 Statistics of the five-fold group-disjoint evaluation. Each entry denotes the number of groups / images.

Fold	Train	Test	Sketch	Quick Sketch	Color
Fold 1	76 / 2696	17 / 664	6 / 209	5 / 203	6 / 252
Fold 2	74 / 2690	19 / 670	7 / 217	6 / 206	6 / 247
Fold 3	73 / 2675	20 / 685	7 / 216	6 / 215	7 / 254
Fold 4	74 / 2688	19 / 672	7 / 216	6 / 210	6 / 246
Fold 5	75 / 2691	18 / 669	6 / 208	6 / 211	6 / 250

All experiments are conducted on a single NVIDIA GeForce RTX 3080 Ti GPU. The software environment is PyTorch 2.1.0 with CUDA 11.8. The backbone network is ResNet-18, initialized with ImageNet-pretrained weights. The batch size is set to one group, and all images are cropped to 224×224 . Only random horizontal flipping is used for data augmentation. During training, we use the Adam optimizer with an initial learning rate of 1×10^{-4} , combined with a cosine annealing learning rate schedule. The minimum learning rate is set to 1×10^{-6} , and each fold is trained for 80 epochs. Unless otherwise specified, the same training configuration is used for all comparison experiments.

4.2 Evaluation Metrics

We use three metrics to evaluate model performance, including Spearman Rank Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), and

Pairwise Accuracy (Pair Acc). Since the AED task focuses on intra-group relative quality assessment, all metrics are first computed within each group and then macro-averaged across groups.

For the m -th group, SRCC is computed as:

$$\text{SRCC}^{(m)} = \frac{\sum_{i=1}^{N_m} (\rho_i^{\hat{y}} - \bar{\rho}^{\hat{y}})(\rho_i^s - \bar{\rho}^s)}{\sqrt{\sum_{i=1}^{N_m} (\rho_i^{\hat{y}} - \bar{\rho}^{\hat{y}})^2} \sqrt{\sum_{i=1}^{N_m} (\rho_i^s - \bar{\rho}^s)^2}}, \quad (6)$$

where N_m is the number of samples in the m -th group, $\rho_i^{\hat{y}} = \text{rank}(\hat{y}_i^{(m)})$ and $\rho_i^s = \text{rank}(s_i^{(m)})$ denote the intra-group ranks of the predicted score and the ground-truth score, respectively. If tied scores exist, their average rank is used. $\bar{\rho}^{\hat{y}}$ and $\bar{\rho}^s$ denote the mean ranks. A higher SRCC indicates that the model can better recover the relative quality order of works within the same evaluation context.

PLCC is computed as:

$$\text{PLCC}^{(m)} = \frac{\sum_{i=1}^{N_m} (\hat{y}_i^{(m)} - \bar{\hat{y}}^{(m)})(s_i^{(m)} - \bar{s}^{(m)})}{\sqrt{\sum_{i=1}^{N_m} (\hat{y}_i^{(m)} - \bar{\hat{y}}^{(m)})^2} \sqrt{\sum_{i=1}^{N_m} (s_i^{(m)} - \bar{s}^{(m)})^2}}, \quad (7)$$

where $\hat{y}_i^{(m)}$ and $s_i^{(m)}$ denote the predicted score and the ground-truth score of the i -th sample, respectively, and $\bar{\hat{y}}^{(m)}$ and $\bar{s}^{(m)}$ are their group-wise mean values. A higher PLCC indicates that the model can better capture the linear quality variation of works within the same group.

To directly evaluate whether the model can correctly judge the relative order of any two works within the same exam context, we further use Pair Acc as an evaluation metric.

$$\mathcal{P}_m = \{(i, j) \mid 1 \leq i < j \leq N_m, s_i^{(m)} \neq s_j^{(m)}\} \quad (8)$$

Let $\mathcal{P}_m$ denote the set of sample pairs with different ground-truth scores in the m -th group. Pair Acc is defined as:

$$\text{PairAcc}^{(m)} = \frac{1}{|\mathcal{P}_m|} \sum_{(i,j) \in \mathcal{P}_m} \mathbf{1} \left[(\hat{y}_i^{(m)} - \hat{y}_j^{(m)})(s_i^{(m)} - s_j^{(m)}) > 0 \right], \quad (9)$$

where $\mathbf{1}[\cdot]$ is the indicator function. If the predicted order is consistent with the ground-truth order, the indicator value is 1; otherwise, it is 0. A higher Pair Acc indicates that the model can more accurately recover pairwise relative quality relationships within the group.

Unless otherwise specified, the reported SRCC, PLCC, and Pair Acc are obtained by first computing the corresponding metric within each group and then taking the macro average across groups:

$$\text{Metric} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \text{Metric}^{(m)}, \quad (10)$$

where $\mathcal{M}$ denotes the set of evaluation groups. It should be noted that all evaluation metrics are computed within each group and then macro-averaged across groups. This protocol is stricter than global evaluation over all samples, because it focuses on works under the same exam context. Therefore, the absolute metric values may be lower than those reported on conventional IAA benchmarks, but they more faithfully reflect the group-conditional assessment objective of AED.

4.3 Performance comparison

We compare DPAR-Net with representative image aesthetic assessment methods, including NIMA [16], MLSP [17], TANet [18], SAAN [13], and HLA-GCN [30]. We additionally include a simple statistical baseline and two ranking-oriented baselines. The statistical baseline, ResNet18-Frozen + Ridge, extracts ImageNet-pretrained ResNet-18 features with the feature extractor frozen, and trains a ridge regression model to predict the group-normalized mid-rank target. The ridge coefficient is selected by inner group-disjoint cross-validation. The two ranking-oriented baselines, Baseline + Pairwise-Rank and Baseline + Listwise-Rank, use the same architecture as Baseline but replace the mid-rank regression objective with ranking-oriented losses. Specifically, Baseline + Pairwise-Rank adopts a RankNet-style pairwise logistic loss, and Baseline + Listwise-Rank adopts a ListNet-style top-one loss. The pairwise preferences and listwise rank distributions are derived from the group-normalized mid-rank targets. All methods are evaluated using the same data splits and evaluation protocol.

Table 4 reports the overall comparison on AED. DPAR-Net achieves the highest mean performance on all three metrics. Compared with Baseline, DPAR-Net improves SRCC, PLCC, and Pair Acc from 0.4189, 0.4226, and 0.6813 to 0.4596, 0.4624, and 0.7035, respectively. This shows that introducing group-context modeling brings consistent mean improvements over the ResNet-18 baseline under the same evaluation setting. The ResNet18-Frozen + Ridge baseline obtains lower performance, with SRCC, PLCC, and Pair Acc of 0.3298, 0.3561, and 0.6351, respectively. This indicates that fixed pretrained features combined with a linear statistical model and group-normalized supervision are not sufficient to address AED. Among the two ranking-oriented variants, Baseline + Pairwise-Rank obtains results close to Baseline, while Baseline + Listwise-Rank obtains lower mean performance. DPAR-Net achieves higher mean performance than both variants, suggesting that standard pairwise or listwise ranking formulations alone are less suited to the group-conditioned assessment setting of AED than explicit group-context modeling.

Compared with general image aesthetic assessment methods, DPAR-Net also obtains higher mean performance. Compared with NIMA, DPAR-Net improves SRCC, PLCC, and Pair Acc by 0.0505, 0.0493, and 0.0284, respectively. Compared with MLSP and HLA-GCN, which enhance single-image representation through multi-level features or regional relationship modeling, DPAR-Net also shows higher mean performance, especially on Pair Acc. This suggests that stronger single-image feature modeling alone is not sufficient for recovering fine-grained intra-group quality relationships in AED. Compared with TANet and SAAN, which introduce theme or style information, DPAR-Net further improves the overall results. In particular, compared with SAAN, DPAR-Net improves SRCC, PLCC, and Pair Acc by 0.0242, 0.0170, and

0.0181, respectively. These results indicate that theme- or style-aware modeling can be helpful, but existing IAA methods still mainly assess each image independently and do not explicitly model relative differences among works under the same proposition.

Therefore, the comparison in this section evaluates the applicability of general IAA methods to AED and further examines ranking-oriented alternatives under the same group-disjoint evaluation setting. Overall, the results suggest that explicit group-context modeling provides a more suitable solution for recovering intra-group relative quality relationships in AED.

Table 4 Comparison of DPAR-Net with baseline, image aesthetic assessment methods, statistical baseline, and ranking-oriented baselines on AED.

Model	Params	SRCC $\uparrow$	PLCC $\uparrow$	Pair Acc $\uparrow$
Baseline	11.2 M	0.4189 ± 0.0450	0.4226 ± 0.0302	0.6813 ± 0.0152
<i>Image aesthetic assessment methods</i>				
NIMA	21.8 M	0.4091 ± 0.0504	0.4131 ± 0.0365	0.6751 ± 0.0218
MLSP	64.9 M	0.3924 ± 0.0375	0.3621 ± 0.0241	0.6412 ± 0.0158
TANet	13.8 M	0.4299 ± 0.0389	0.4317 ± 0.0290	0.6786 ± 0.0149
SAAN	26.0 M	0.4354 ± 0.0436	0.4454 ± 0.0317	0.6854 ± 0.0299
HLA-GCN	44.3 M	0.4133 ± 0.0388	0.4276 ± 0.0390	0.6496 ± 0.0257
<i>Statistical baseline</i>				
ResNet18-Frozen + Ridge	—	0.3298 ± 0.0500	0.3561 ± 0.0525	0.6351 ± 0.0209
<i>Ranking-oriented baselines</i>				
Baseline + Pairwise-Rank	11.2 M	0.4184 ± 0.0444	0.4286 ± 0.0330	0.6832 ± 0.0179
Baseline + Listwise-Rank	11.2 M	0.3540 ± 0.0715	0.3662 ± 0.0726	0.6499 ± 0.0303
DPAR-Net (Ours)	13.2 M	0.4596 ± 0.0477	0.4624 ± 0.0239	0.7035 ± 0.0189

4.4 Ablation Study

To compare different architectural configurations of DPAR-Net, we construct several variants based on the same single-image baseline. Baseline+GMRP adds only the Group-Mean Reference Path, while Baseline+RRP adds only the Residual Relative Path. Baseline+SingleRefMLP merges the feature inputs of the two reference paths into a single-path representation and uses one MLP to predict a unified correction term. Baseline+GMRP+RRP keeps the two separate correction heads but directly adds both correction terms to the baseline prediction. Baseline+GMRP+RRP+MLPFusion replaces Adaptive Routing Fusion with an unconstrained MLP fusion module. DPAR-Net (Full) uses Adaptive Routing Fusion to combine the baseline prediction and the two correction candidates in a baseline-anchored residual form.

The results are reported in Table 5. Compared with the baseline, using GMRP alone does not bring improvement, while RRP slightly improves SRCC and Pair Acc, suggesting that the residual-relative representation may contain useful group-relative

information for intra-group ranking. Baseline+SingleRefMLP performs worse than Baseline+GMRP+RRP, indicating that representing the two correction candidates with separate heads is more suitable than merging them into a single correction branch in this setting. Baseline+GMRP+RRP improves over the single-path variants, but its performance remains lower than that of DPAR-Net (Full), suggesting that directly adding the correction terms with fixed coefficients is not the most effective fusion strategy. We further compare Adaptive Routing Fusion with ordinary MLP fusion. Baseline+GMRP+RRP+MLPFusion keeps the two correction heads but uses an MLP to regress the final prediction from the feature and prediction-related inputs. DPAR-Net (Full) achieves higher SRCC, PLCC, and Pair Acc than this variant, suggesting that Adaptive Routing Fusion, which keeps the prediction anchored to the baseline output, is more suitable for this benchmark than the ordinary MLP fusion used in Baseline+GMRP+RRP+MLPFusion.

Overall, the ablation results show that DPAR-Net (Full) achieves the best performance among the tested variants. These results support using DPAR-Net as an initial reference architecture for AED.

Table 5 Results of the ablation study of DPAR-Net on AED.

Model	SRCC $\uparrow$	PLCC $\uparrow$	Pair Acc $\uparrow$
Baseline	0.4189 $\pm$ 0.0450	0.4226 $\pm$ 0.0302	0.6813 $\pm$ 0.0152
Baseline+GMRP	0.4074 $\pm$ 0.0741	0.4136 $\pm$ 0.0524	0.6788 $\pm$ 0.0328
Baseline+RRP	0.4224 $\pm$ 0.0417	0.4213 $\pm$ 0.0412	0.6872 $\pm$ 0.0206
Baseline+SingleRefMLP	0.4099 $\pm$ 0.0676	0.4024 $\pm$ 0.0599	0.6796 $\pm$ 0.0313
Baseline+GMRP+RRP	0.4347 $\pm$ 0.0569	0.4404 $\pm$ 0.0332	0.6871 $\pm$ 0.0265
Baseline+GMRP+RRP+MLPFusion	0.4207 $\pm$ 0.0467	0.4201 $\pm$ 0.0414	0.6839 $\pm$ 0.0139
DPAR-Net (Full)	0.4596 $\pm$ 0.0477	0.4624 $\pm$ 0.0239	0.7035 $\pm$ 0.0189

4.5 Routing and Fusion Diagnostics

To examine how Adaptive Routing Fusion is used in practice, we visualize the learned routing weights in Fig. 4. Fig. 4(a) shows the overall routing weights of the three prediction paths. The baseline path receives the dominant routing weight, whereas the context-corrected and residual-corrected paths are assigned very small weights. This indicates that the learned fusion module follows a conservative baseline-anchored pattern, and that the final prediction is mainly determined by the baseline output.

We further inspect the non-baseline routing weights under different sample and group conditions in Fig. 4(b)–(d). Since AED does not provide independent human-labeled sample difficulty annotations, we do not define sample difficulty as a separate ground-truth dimension. Instead, we analyze the routing weights across intra-group rank ranges, art categories, and group sizes. The rank ranges in Fig. 4(b) describe positions in the assessment-derived intra-group ordering, where Top 20% denotes the highest-scoring works within each group and Bottom 20% denotes the lowest-scoring works.



Fig. 4 Visualization of the learned routing weights in Adaptive Routing Fusion. Panel (a) shows the overall routing weights of the three prediction paths, including the baseline path α_i^{base} , the context-corrected path α_i^{ctx} , and the residual-corrected path α_i^{rel} . Panels (b)–(d) report the non-baseline routing weights α_i^{ctx} and α_i^{rel} under different sample and group conditions, including intra-group rank range, art category, and group size. The intra-group rank ranges are defined using the ground-truth group-normalized mid-rank targets, where larger target values indicate higher within-group scores. Panels (b)–(d) use magnified y-axes to show the small non-baseline routing weights.

Overall, the non-baseline weights remain small across the examined conditions, including different rank ranges, categories, and group sizes. This suggests that the model mainly relies on the baseline prediction, including for clearly high- or low-scoring rank ranges. The non-baseline weights show only weak variations across the examined conditions. For example, the residual-relative weight is slightly larger in the middle rank ranges, the color painting subset shows relatively larger non-baseline weights than the other categories, and larger groups show slightly higher non-baseline weights. These trends are used only as descriptive diagnostics of the learned routing behavior and should not be interpreted as evidence of large direct correction effects at inference time.

Given this baseline-dominant routing pattern, the performance improvement of the full model should not be interpreted as being mainly caused by large direct corrections at inference time. To further analyze the role of routing in this setting, we conduct a controlled diagnostic experiment, as reported in Table 6. The Baseline model is trained with the single-image prediction loss $L(b, y)$. DPAR-Net (base-loss only) keeps the full DPAR-Net architecture and still computes b_i , c_i , r_i , and the routing weights, but applies the training loss only to the baseline-path output b_i , rather than to the fused prediction $\hat{y}_i$. DPAR-Net (Full) optimizes the fused prediction $\hat{y}_i = b_i + \alpha_i^{ctx}c_i + \alpha_i^{rel}r_i$.

As shown in Table 6, DPAR-Net (base-loss only) performs similarly to the Baseline, indicating that simply adding the dual-path encoder and routing module without supervising the fused output does not improve the model by itself. In contrast, DPAR-Net (Full), trained with the fused prediction, achieves higher SRCC, PLCC, and Pair Acc. These results suggest that the gain is not mainly due to inference-time routing-based path selection. Instead, it is related to using the fused prediction $\hat{y}_i$ as the supervised output during training.

Table 6 Controlled diagnostic results for the fused-prediction training formulation.

Model	SRCC	PLCC	Pair Acc
Baseline	0.4189 ± 0.0450	0.4226 ± 0.0302	0.6813 ± 0.0152
DPAR-Net (base-loss only)	0.4140 ± 0.0470	0.4190 ± 0.0323	0.6804 ± 0.0173
DPAR-Net (Full)	0.4596 ± 0.0477	0.4624 ± 0.0239	0.7035 ± 0.0189

4.6 Qualitative Evaluation

To further observe the intra-group ranking performance of the model on specific samples, Fig. 5 shows the prediction results of Baseline and DPAR-Net on three representative groups. It should be noted that the outputs of the model during training and prediction are normalized relative quality scores. For better visualization, the ground-truth scores and predicted scores are converted into intra-group ranks in the figure, where a smaller rank value indicates a higher ranking position. For samples with tied ground-truth scores, interval ranks are used in the figure.

From the overall results, the predicted rankings of DPAR-Net are closer to the GT ranks than those of Baseline, especially for samples in the middle-quality range. In the Sketch group, Baseline can identify samples at the two ends of the ranking, but it produces obvious misorderings for middle-level works. For example, the sample with a GT rank of 5/10 is predicted as 9/10, while the sample with a GT rank of 6/10 is predicted as 3/10. In contrast, DPAR-Net produces predictions that are closer to the ground-truth order for these samples. In the Quick Sketch group, Baseline underestimates some high-quality samples and overestimates some low-quality samples, indicating that single-image features may introduce local ranking deviations in fine-grained quick sketch assessment. DPAR-Net alleviates these ranking shifts for most samples, although overestimation still occurs for a few low-quality samples. In the Color group, Baseline shows local disorder in the middle and lower ranking ranges, while DPAR-Net produces rankings that are closer to the GT ranks for neighboring-level samples.

These qualitative results show that DPAR-Net produces more stable relative ranking relationships under shared evaluation conditions in these examples. Meanwhile, the remaining prediction deviations for a few samples also indicate that distinguishing adjacent quality levels in art exam works is still challenging.

(a) Sketch							
GT		1/10	2/10	4/10	5/10	6/10	10/10
Baseline		1/10	4/10	6/10	9/10	3/10	10/10
DPAR-Net		1/10	2/10	4/10	8/10	5/10	10/10
(b) Quick Sketch							
GT		1-2/20	5/20	10/20	17/20	19-20/20	19-20/20
Baseline		8/20	17/20	13/20	9/20	20/20	6/20
DPAR-Net		4/20	12/20	10/20	17/20	19/20	7/20
(c) Color							
GT		1-2/20	7-8/20	11-12/20	15-16/20	17-18/20	19-20/20
Baseline		7/20	3/20	14/20	11/20	19/20	20/20
DPAR-Net		4/20	6/20	11/20	15/20	18/20	20/20

Fig. 5 Qualitative comparison of group-wise ranking results. Examples are selected from three art exam categories: (a) sketch, (b) quick sketch, and (c) color painting. For each sample, GT denotes the ground-truth rank within the corresponding group, while Baseline and Ours denote the predicted ranks produced by the baseline model and DPAR-Net, respectively. A smaller rank indicates higher quality. Rank intervals are used for samples with tied ground-truth scores. Bold values indicate the prediction closer to the ground-truth rank between Baseline and Ours.

4.7 Category-wise Analysis

To analyze the performance of the model on different art-exam categories, we further evaluate Baseline and DPAR-Net on the Sketch, Quick Sketch, and Color Painting subsets. Specifically, in each fold, SRCC, PLCC, and Pair Acc are computed for each category according to the evaluation protocol in Section 4.2. The results are then summarized over the five folds, and the mean performance together with the 95% confidence interval is reported in Table 7. Each result is presented as the mean value followed by the 95% confidence interval in parentheses. The confidence intervals are computed over the five folds using Student’s t -distribution.

The category-wise results show that DPAR-Net achieves higher mean performance than Baseline across all three categories and all three metrics. For the Sketch category, DPAR-Net obtains SRCC, PLCC, and Pair Acc values of 0.5329, 0.5332, and 0.7520, respectively, compared with 0.4989, 0.5247, and 0.7306 for Baseline. For the Quick Sketch category, DPAR-Net improves the mean SRCC from 0.4848 to 0.5042, PLCC from 0.4663 to 0.5017, and Pair Acc from 0.7023 to 0.7112. For the Color Painting category, DPAR-Net improves SRCC, PLCC, and Pair Acc from 0.3671, 0.3686, and

0.6524 to 0.4224, 0.4432, and 0.6819, respectively. These consistent mean improvements suggest that the proposed group-conditioned modeling strategy is not restricted to a single art-exam category.

The results also reveal category-dependent performance differences. Compared with Sketch, Color Painting obtains lower SRCC and Pair Acc for both Baseline and DPAR-Net, indicating that ranking color painting works is more challenging in AED. The confidence intervals further show the uncertainty of the category-wise estimates across folds, with relatively wider intervals observed for Quick Sketch and Color Painting. Meanwhile, DPAR-Net consistently achieves higher mean performance across all categories and metrics. Therefore, the category-wise confidence intervals provide a more transparent basis for judging the generality of the reported improvements.

Table 7 Category-wise performance with 95% confidence intervals over five folds.

Category	Metric	Baseline	DPAR-Net
Sketch	SRCC	0.4989 (0.4076, 0.5902)	0.5329 (0.4080, 0.6578)
Sketch	PLCC	0.5247 (0.4071, 0.6423)	0.5332 (0.3981, 0.6683)
Sketch	Pair Acc	0.7306 (0.6599, 0.8013)	0.7520 (0.6657, 0.8383)
Quick Sketch	SRCC	0.4848 (0.2646, 0.7050)	0.5042 (0.2965, 0.7119)
Quick Sketch	PLCC	0.4663 (0.2111, 0.7215)	0.5017 (0.2310, 0.7724)
Quick Sketch	Pair Acc	0.7023 (0.6101, 0.7945)	0.7112 (0.6152, 0.8072)
Color	SRCC	0.3671 (0.1074, 0.6268)	0.4224 (0.1906, 0.6542)
Color	PLCC	0.3686 (0.1300, 0.6072)	0.4432 (0.2487, 0.6377)
Color	Pair Acc	0.6524 (0.5385, 0.7663)	0.6819 (0.5794, 0.7844)

4.8 Group Size Analysis

To further analyze the influence of group size on model performance, we divide the test groups according to the number of samples in each group. Specifically, $N = 30$ is used as the threshold, and the test groups are divided into two subsets: $N \leq 30$ and $N > 30$. This threshold is selected according to the distribution of group sizes in the test set, so that the two subsets are relatively balanced in terms of the number of groups. The $N \leq 30$ subset contains 48 groups with 899 images, while the $N > 30$ subset contains 45 groups with 2461 images. The results are shown in Table 8.

Table 8 Performance comparison under different group sizes.

Group Size	Model	SRCC $\uparrow$	PLCC $\uparrow$	Pair Acc $\uparrow$
$N \leq 30$	Baseline	0.3809 ± 0.0999	0.3696 ± 0.0927	0.6741 ± 0.0398
$N \leq 30$	DPAR-Net	0.4329 ± 0.1074	0.4130 ± 0.0810	0.7053 ± 0.0484
$N > 30$	Baseline	0.4598 ± 0.0373	0.4838 ± 0.0521	0.6900 ± 0.0136
$N > 30$	DPAR-Net	0.4928 ± 0.0450	0.5215 ± 0.0757	0.7046 ± 0.0174

The experimental results show that DPAR-Net outperforms Baseline under both group-size settings. For small groups with $N \leq 30$, DPAR-Net achieves clear improvements on all three metrics, with SRCC increasing from 0.3809 to 0.4329, PLCC from 0.3696 to 0.4130, and Pair Acc from 0.6741 to 0.7053. This indicates that even when the number of intra-group samples is relatively limited, DPAR-Net can still recover relative quality relationships among works more effectively than the single-image Baseline.

For larger groups with $N > 30$, DPAR-Net also maintains a stable advantage. Although the improvement margin is slightly smaller than that in small groups, the overall results still show that the proposed method remains effective when more samples are available to form group-context representations.

Overall, group size has a certain influence on task difficulty and model performance. Compared with larger groups, small groups show larger performance fluctuations, suggesting that when fewer samples are available within a group, the assessment results are more easily affected by sample distribution and supervision sparsity. Nevertheless, DPAR-Net consistently outperforms Baseline under both group-size settings, demonstrating its stability and generalization ability under different group-size conditions.

5 Limitations

Although the proposed method achieves good overall performance on AED, it still has several limitations. First, AED currently provides only final exam scores, without judge-level rating records or dimension-specific evaluation annotations. Therefore, the model cannot directly learn fine-grained expert judgment rules in terms of structural accuracy, detail completeness, and overall coordination. In addition, art assessment itself contains a certain degree of subjectivity, and the final scores may inevitably include subjective bias. Therefore, the model should be interpreted as approximating the relative ordering embedded in the art-exam assessment process, rather than learning objective artistic quality.

We also analyze several failure cases, as shown in Fig. 6. The model may overestimate some works with strong visual impact or salient color features, while underestimating works that are less visually striking but more stable in facial depiction or structural relationships. This indicates that although the current model improves overall ranking consistency under the group-conditioned setting, it is still insufficient in learning professional evaluation dimensions and high-level assessment factors. Future work should therefore incorporate expert comments, dimension-specific evaluation criteria, rubric-based labels, judge-level scores, and multi-rater uncertainty as essential extensions. These richer forms of annotation and supervision are necessary for more rigorous computational modeling of art-exam assessment [44, 45].

6 Conclusions and future work

In this study, we introduce AED, a group-conditioned art exam dataset containing 3,360 images, and formulate art exam assessment as a fine-grained intra-group ranking problem under shared evaluation conditions. AED complements a scenario that has been largely overlooked by existing single-image aesthetic assessment datasets, namely



Fig. 6 Failure cases with large prediction ranking errors. Below each sample, Pred represents the model’s predicted within-group ranking, and GT represents the true within-group ranking; a smaller ranking value indicates higher quality. For samples with tied true scores, a ranking interval is used.

the assessment of multiple works created under the same proposition, category, and examination conditions. We further provide DPAR-Net as an initial reference architecture for this benchmark. Experimental results show that DPAR-Net achieves higher mean SRCC, PLCC, and Pair Acc than several representative image aesthetic assessment methods and ranking-oriented baselines under the group-disjoint evaluation protocol.

Our experiments reveal two key observations. First, directly applying traditional single-image aesthetic assessment models to art exam assessment still has limitations. Second, group-conditioned modeling is useful for recovering fine-grained relative ordering among works created under shared evaluation conditions. These findings suggest that art exam assessment should be studied not only as single-image aesthetic prediction, but also as a group-based visual comparison problem.

This work provides a feasible direction for preliminary screening, intra-group ranking, and auxiliary assessment of art exam works, and also offers a data foundation and methodological reference for future research on intelligent assessment under shared evaluation conditions. More broadly, AED can support future work in visual computing and computer graphics by providing a structured benchmark for artistic image understanding, fine-grained visual comparison, and group-conditioned visual analysis. In future work, we will first validate the model in small-scale and controlled exam scenarios, and continue to collect more theme-based works to further evaluate and improve the generalization ability of the model. We will also attempt to incorporate expert comments and higher-level evaluation criteria, and explore the fusion of visual features with professional evaluation language, so as to improve the practicality and interpretability of the model in art education auxiliary assessment scenarios.

CRediT authorship contribution statement

Jiahao Li: Methodology, Software, Writing – original draft. **Yingjie Zhang:** Methodology, Validation. **Zhongwei Huang:** Investigation, Resources. **Haoze Chen:** Data curation, Validation. **Zongming Tan:** Data curation, Investigation, Validation. **Bao-hua Tan:** Supervision, Project administration, Funding acquisition, Writing – review & editing. **Chao Chen:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The image archive of the Art Exam Dataset (AED) is publicly available at <https://github.com/farawaylx/AED-Dataset>.

Acknowledgments

Thank all the authors for their hard work in this project. And thank to the Hubei University of Technology and the National “111 Research Center” for Microelectronics and Integrated Circuits for providing research facilities and all the equipment, and the University of Nottingham for providing relevant research software.

Funding

This work was supported partly by National Natural Science Foundation of China (Grants no. 62306106). China Scholarship Council, the Funding of The Innovation and Entrepreneurship Program for College Students of the Ministry of Education of China (Grants no. 202510500019), the Key Research Project of Xinjiang Autonomous Region-Joint Task Special Project between Provincial Department and Local Government, China (Grants no. 20261152731), the University Innovation Fund of Ministry of Education of China (2022BL052), the National Ministry of Education industry-university cooperation project (220900487303708), the Major Project of Research on Philosophy and Social Science of Higher Education Institutions in Hubei Province (Grants no. 21ZD054), the Science and Technology Innovation Talent Program of Hubei Province, China (Grants no. 2023DJC119), and the Major Project of Hubei Key Laboratory of intelligent transportation technology and device in Hubei Polytechnic University(Grants no. 2022XZ106).

References

- [1] Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and

- Pattern Recognition (CVPR), pp. 2414–2423 (2016). <https://doi.org/10.1109/CVPR.2016.265>
- [2] Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1510–1519 (2017). <https://doi.org/10.1109/ICCV.2017.167>
 - [3] Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5873–5881 (2019). <https://doi.org/10.1109/CVPR.2019.00603>
 - [4] Zhang, C., Lei, K., Jia, J., Ma, Y., Hu, Z.: Ai painting: An aesthetic painting generation system. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 1231–1233 (2018). <https://doi.org/10.1145/3240508.3241386>
 - [5] Xu, B., Cheng, X., Tan, H., Han, Y., Zhang, J.: Effisketch3d: Data-efficient 3d generation from sketches via dual-domain pre-trained priors. *The Visual Computer* **42**(1), 21 (2026) <https://doi.org/10.1007/s00371-025-04247-3>
 - [6] Pan, J., Li, F.W.B., Yang, B., Nan, F.: Clpfusion: A latent diffusion model framework for realistic chinese landscape painting style transfer. *Computer Animation and Virtual Worlds* **36**(3), 70053 (2025) <https://doi.org/10.1002/cav.70053>
 - [7] Gault, M.: An AI-generated artwork won first place at a state fair fine arts competition, and artists are pissed. *Vice*. Accessed: 2026-04-27 (2022)
 - [8] Li, C., Sun, S., Min, X., Wang, W., Tang, Z.: Application of deep convolutional features in sketch works classification and evaluation. *Journal of Computer-Aided Design and Computer Graphics* **29**(9) (2017)
 - [9] Yu, S., Lin, Z., Liang, J., Shu, G., Yu, J., Zhu, A.: Sketch works ranking based on improved transfer learning model. *Multimedia Tools and Applications* **80**, 33663–33678 (2021) <https://doi.org/10.1007/s11042-021-11305-0>
 - [10] Wang, G., Ke, Y., Yang, S., Wang, K., Guo, W., Qin, F.: A novel framework for aesthetic assessment of portrait sketches via multi-feature integration and self-supervised learning. *Expert Systems with Applications* **292**, 128659 (2025) <https://doi.org/10.1016/j.eswa.2025.128659>
 - [11] Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2408–2415 (2012). <https://doi.org/10.1109/CVPR.2012.6247954>
 - [12] Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C.: Photo aesthetics ranking

- network with attributes and content adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016). <https://doi.org/10.48550/arXiv.1606.01621> . arXiv:1606.01621
- [13] Yi, R., Tian, H., Gu, Z., Lai, Y.-K., Rosin, P.L.: 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 (CVPR), pp. 22388–22397 (2023). <https://doi.org/10.1109/CVPR52729.2023.02144>
 - [14] Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: Computer Vision – ECCV 2006, pp. 288–301. Springer, Berlin, Heidelberg (2006). https://doi.org/10.1007/11744078_23
 - [15] Nishiyama, M., Okabe, T., Sato, I., Sato, Y.: Aesthetic quality classification of photographs based on color harmony. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 33–40 (2011). <https://doi.org/10.1109/CVPR.2011.5995539>
 - [16] Talebi, H., Milanfar, P.: Nima: Neural image assessment. *IEEE Transactions on Image Processing* **27**(8), 3998–4011 (2018) <https://doi.org/10.1109/TIP.2018.2831899>
 - [17] Hosu, V., Goldlücke, B., Saupe, D.: Effective aesthetics prediction with multi-level spatially pooled features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9367–9375 (2019). <https://doi.org/10.1109/CVPR.2019.00960>
 - [18] He, S., Zhang, Y., Xie, R., Jiang, D., Ming, A.: Rethinking image aesthetics assessment: Models, datasets and benchmarks. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI), pp. 942–948 (2022). <https://doi.org/10.24963/ijcai.2022/132>
 - [19] Behrad, F., Tuytelaars, T., Wagemans, J.: Charm: The missing piece in vit fine-tuning for image aesthetic assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7815–7824 (2025). <https://doi.org/10.1109/CVPR52734.2025.00732>
 - [20] Gao, F., Lin, Y., Shi, J., Qiao, M., Wang, N.: Aesmamba: Universal image aesthetic assessment with state space models. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 7444–7453 (2024). <https://doi.org/10.1145/3664647.3681011>
 - [21] Yang, B., Zhu, C., Li, F.W.B., Wei, T., Liang, X., Wang, Q.: Iaacs: Image aesthetic assessment through color composition and space formation. *Virtual Reality & Intelligent Hardware* **5**(1), 42–56 (2023) <https://doi.org/10.1016/j.vrih.2022.06.006>

- [22] Cao, S., Ma, N., Li, J., Li, X., Shao, L., Zhu, K., Zhou, Y., Pu, Y., Wu, J., Wang, J., Qu, B., Wang, W., Qiao, Y., Yao, D., Liu, Y.: ArtiMuse: Fine-Grained Image Aesthetics Assessment with Joint Scoring and Expert-Level Understanding. arXiv:2507.14533 (2025). <https://doi.org/10.48550/arXiv.2507.14533>
- [23] Sheng, X., Li, L., Chen, P., Wu, J., Dong, W., Yang, Y., Xu, L., Li, Y., Shi, G.: Aesclip: Multi-attribute contrastive learning for image aesthetics assessment. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 1117–1126 (2023). <https://doi.org/10.1145/3581783.3611969>
- [24] Liu, Y., Wen, Z., Li, L., Jing, P., Fan, D.: Migf-net: Multimodal interaction-guided fusion network for image aesthetics assessment. Pattern Recognition **172**, 112401 (2026) <https://doi.org/10.1016/j.patcog.2025.112401>
- [25] Xu, X., Huang, S., Li, W., Wang, F.: Attribute-guided aesthetic assessment for artistic images based on multimodal hybrid network. Expert Systems with Applications **299**, 129954 (2026) <https://doi.org/10.1016/j.eswa.2025.129954>
- [26] Meng, F., Han, D., Shen, X., Chen, C., Wang, Q.: Enhancing image-text matching through contextual fine-grained alignment. The Visual Computer **42**(1), 70 (2026) <https://doi.org/10.1007/s00371-025-04250-8>
- [27] Peng, T., Yin, W., Liu, J., Li, L., Hu, X.: Deconfounded fashion image captioning with transformer and multimodal retrieval. Virtual Reality & Intelligent Hardware **7**(2), 127–138 (2025) <https://doi.org/10.1016/j.vrih.2024.08.002>
- [28] Ma, S., Liu, J., Chen, C.W.: 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 (CVPR), pp. 722–731 (2017). <https://doi.org/10.1109/CVPR.2017.84>
- [29] Liu, D., Puri, R., Kamath, N., Bhattacharya, S.: Composition-aware image aesthetics assessment. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 3558–3567 (2020). <https://doi.org/10.1109/WACV45572.2020.9093412>
- [30] She, D., Lai, Y.-K., Yi, G., Xu, K.: Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8471–8480 (2021). <https://doi.org/10.1109/CVPR46437.2021.00837>
- [31] Yang, Z., Li, L., Yang, Y., Li, Y., Lin, W.: Multi-level transitional contrast learning for personalized image aesthetics assessment. IEEE Transactions on Multimedia **26**, 1944–1956 (2024) <https://doi.org/10.1109/TMM.2023.3290479>
- [32] Shi, T., Chen, C., Li, X., Hao, A.: Semantic and style based multiple reference learning for artistic and general image aesthetic assessment. Neurocomputing

- [33] Li, L., Huang, Y., Wu, J., Yang, Y., Li, Y., Guo, Y., Shi, G.: Theme-aware visual attribute reasoning for image aesthetics assessment. *IEEE Transactions on Circuits and Systems for Video Technology* **33**(9), 4798–4811 (2023) <https://doi.org/10.1109/TCSVT.2023.3249185>
- [34] Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: MUSIQ: Multi-scale Image Quality Transformer. *arXiv:2108.05997* (2021). <https://doi.org/10.48550/arXiv.2108.05997>
- [35] Shi, T., Chen, C., Wu, Z., Hao, A., Fang, Y.: Improving image aesthetic assessment via multiple image joint learning. *ACM Transactions on Multimedia Computing, Communications, and Applications* **20**, 1–24 (2024) <https://doi.org/10.1145/3687128>
- [36] Lee, J.-T., Kim, C.-S.: Image aesthetic assessment based on pairwise comparison: A unified approach to score regression, binary classification, and personalization. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1191–1200 (2019). <https://doi.org/10.1109/ICCV.2019.00128>
- [37] Zhou, J., Zhang, Q., Fan, J.-H., Sun, W., Zheng, W.-S.: Joint regression and learning from pairwise rankings for personalized image aesthetic assessment. *Computational Visual Media* **7**, 241–252 (2021) <https://doi.org/10.1007/s41095-021-0207-y>
- [38] Cheng, J., Li, C., He, S., Chen, T., Ming, A.: Thinking aesthetics assessment of image color temperature: Models, datasets and benchmarks. *Proceedings of the AAAI Conference on Artificial Intelligence* **40**(5), 3246–3254 (2026) <https://doi.org/10.1609/aaai.v40i5.37319>
- [39] Tang, H., Chen, Y., Liang, X., Chen, L., Huang, P., Tang, Z.: Artistic image aesthetics assessment assisted by photographic visual attributes. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2025). <https://doi.org/10.1109/ICASSP49660.2025.10890249>
- [40] Jin, X., Qiao, Q., Lu, Y., Wang, H., Gao, S., Huang, H., Li, G.: Paintings and drawings aesthetics assessment with rich attributes for various artistic categories. In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 7672–7680 (2024). <https://doi.org/10.24963/ijcai.2024/849>
- [41] Jiang, S., Li, N., Shi, C., Guo, L., Wang, C., Li, C.: Aacp: Aesthetics assessment of children’s paintings based on self-supervised learning. *Proceedings of the AAAI Conference on Artificial Intelligence* **38**(3), 2534–2542 (2024) <https://doi.org/10.1609/aaai.v38i3.28030>

- [42] Yang, Z., Wang, J., Zhang, Z., Xie, P., Sheng, X., Chen, P., Li, L.: Fine-grained Image Aesthetic Assessment: Learning Discriminative Scores from Relative Ranks. arXiv:2603.03907 (2026). <https://doi.org/10.48550/arXiv.2603.03907>
- [43] Wang, Z., Li, T., Qin, Y., Sheng, B.: Contrastive decoupling: Dynamic regularization for enhanced fine-grained image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–15 (2026) <https://doi.org/10.1109/TCSVT.2026.3661760>
- [44] Zhu, T., Li, L., Chen, P., Wu, J., Yang, Y., Li, Y.: Emotion-aware hierarchical interaction network for multimodal image aesthetics assessment. *Pattern Recognition* **154**, 110584 (2024) <https://doi.org/10.1016/j.patcog.2024.110584>
- [45] Huang, Y., Sheng, X., Yang, Z., Yuan, Q., Duan, Z., Chen, P., Li, L., Lin, W., Shi, G.: Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In: *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5911–5920 (2024). <https://doi.org/10.1145/3664647.3680649>