Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Exhibition-Based Learning as a Promotion Model for Junior High School Art Club Curriculum Development

Ji Wu

sherawies231@outlook.com

No.2 Junior High School, Chang'an District, Xi'an City, Shaanxi Province, Xi'an, 710100, China
https://orcid.org/0009-0000-6668-7675

Jing Zhang

School of Teacher Development, Shaanxi Normal University Xi'an Aerospace City No.1 High School, Xi'an, 710100, China,

Di Wu

No.3 Junior High School, Chang'an District, Xi'an City, Shaanxi Province, Xi'an, 710100, China

Yiming Peng

Xi'an Chang-an District No.3 Middle School, Xi'an, 710100, China

Research Article

Keywords: exhibition-based learning, multimodal fusion, AI art evaluation, course recommendation, cross-school transfer

Posted Date: September 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10610307/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Exhibition–Based Learning as a Promotion Model for Junior High School Art Club Curriculum Development

Ji Wu*

No.2 Junior High School, Chang'an District, Xi'an City, Shaanxi Province, Xi'an, 710100, China, jiwu9488@Gmail.com

Jing Zhang

School of Teacher Development, Shaanxi Normal University；Xi'an Aerospace City No.1 High School, Xi'an, 710100, China, 675408638@qq.com

Di Wu

No.3 Junior High School, Chang'an District, Xi'an City, Shaanxi Province, Xi'an, 710100, China, 452489152@qq.com

Yiming Peng

Xi'an Chang-an District No.3 Middle School, Xi'an, 710100, China，821569149@qq.com

Abstract

This paper proposes a multimodal AI evaluation and curriculum model to solve the problem of subjective art evaluation, weak curriculum connection and low reuse rate of art parameters between schools. The data consists of artwork images, curatorial texts, peer reviews, audience feedback and teacher records from 120 students in 4 schools. Pretrained ConvNeXt and semantic encoders are fused by masked gated attention and eight teachers give rubric–based labels with ICC(2, k) = 0.86. Due to the very small size of the data, the measures of grouped splitting, freezing of the backbone layers, regularization, five–seeds training and leave–one–school–out validation are used to prevent overfitting. The model's MAE is 0.49, while that of ViT–B/16 and CLIP–ViT/B32 are 0.70 and 0.51, respectively, and the model's expert correlation is 0.87, and its topic–matching accuracy is 86.8% (p=0.018 for ViT–B/16, p=0.037 for CLIP–ViT/B32). Calibration helps to increase the cross–school alignment of topics to 83.4% and decreases teacher assessment time by 44.9%, which does not lend itself to universal generalization, but to deployment.

CCS CONCEPTS

Applied computing → Arts and humanities → Fine arts

Keywords

exhibition–based learning; multimodal fusion; AI art evaluation; course recommendation; cross–school transfer

1 Introduction

With the development of deep learning, image aesthetic evaluation is gradually moving from expert evaluation to deep learning. Daryanavard Chounchenani et al. used deep neural networks for image classification, quality prediction and performance estimation [1]. Soydaner and Wagemans proposed to predict both image attributes and aesthetic scores using multi-task convolutional networks to get better structured representation of the evaluation outcomes [2]. In personalized aesthetic assessment, Adilova and Shamoi added fuzzy logic and colour preferences, ensuring adaptation to individual differences [3]. Redies et al. created a toolbox for quantitative image analysis for composition, colour and complexity [4]. Based on these studies, the object of this work is the junior high school art club, and a model for art assessment through AI, matching the topics of student works and updating the parameters in an iterative manner is created to share with other schools.

2 Exhibition-Based Learning Data Construction

In the 14 week long exhibition project "Campus Memories and Urban Life", 120 students from four schools created 240 pieces of artwork together with 120 curatorial descriptions, 360 peer reviews, 500 audience comments and 24 teaching records [5]. Multiple records were from the same learner and to prevent inflating the effective sample size the student was used as the sampling unit. The artwork images and the curatorial texts, peer reviews, audience feedback and teacher records were illuminated using student, artwork, theme and school identifiers. Images were taken at a fixed lighting condition at 300dpi and then cropped to 512 × 512 pixels and then white-balance corrected and normalized. The text data was preprocessed, segmented and sentiment tagged. Data augmentation was only applied to the training samples and grouped splitting, regularization, transfer learning, and cross-school validation was used to mitigate overfitting.Each student's exhibition-based learning sample for Week t is represented as follows:

$$X_i^t = \left\{ I_i^t, T_i^t, P_i^t, A_i^t, F_i^t \right\} \tag{1}$$

where X_i^t represents the multimodal sample of the i th student in t ; I_i^t represents the image features of the artwork; T_i^t represents the features of the curatorial description text; P_i^t represents the peer review features; A_i^t represents the audience feedback features; and F_i^t represents the teacher guidance record features. The data is indexed by student, artwork, theme, and school identities for the purposes of AI evaluation, to match with a topic, and to optimize across schools. The course comprised nine stages, which constituted the implementation process depicted in Figure 1: Theme introduction, visual exploration, experimentation of materials, production of art works,

curatorial comments, peer review, exhibition presentation, feedback from audience and teacher reflection.

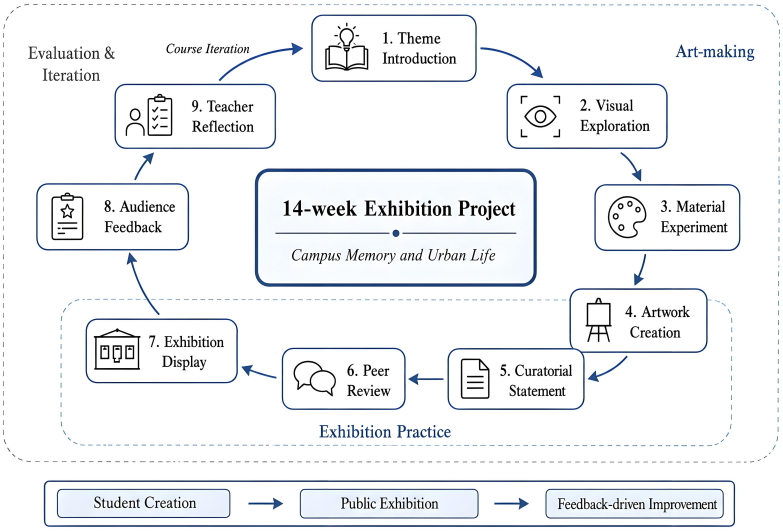


Figure 1: Closed–Loop Diagram of the Exhibition–Based Learning Art Club Course Implementation

3 Design of an AI Art Assessment Model

3.1 Image Feature Extraction

Images of students' artwork are retrieved from an exhibition based learning dataset. All images are subjected to perspective correction, background cropping, white balance adjustment, and normalization to 512 x 512 [6] before input. Given a set of 240 artwork images, ConvNeXt–Tiny is initialized with the pretrained weights of ImageNet. The first two layers are frozen, and only the high level, light attention, and fusion scoring layer are finely tuned. The model retains four levels of feature output, and adds light attention modules in phase 3 and 4 to improve response to theme areas, color gradation, and edge structures. Image features are not directly mapped to the overall score, but are divided into four visual measures: composition integrity, colour harmony, material texture representation, and visual coherence.

The image encoding process is represented as follows:

$$\mathbf{v}_i = f_\theta\left(\mathrm{N}\left(I_i\right)\right) \tag{2}$$

where I_i represents the original image of the th student work (i); $\mathrm{N}(\cdot)$ denotes image normalization and cropping operations; $f_\theta(\cdot)$ denotes a visual feature extraction network with parameters θ ; and $\mathbf{v}_i \in \mathbb{R}^d$ is the output visual feature vector of the work, which is subsequently fed into the scoring fusion layer.

3.2 Text Semantic Analysis

Text fields are set up for curatorial description, peer review, and audience feedback. Then, after removing the invalid symbols, repeating phrases, and low information comments, we use the pretrained semantic encoder to extract the sentence vector. In the case of curatorial descriptions, emphasis is placed on the identification of key words, cultural explanations, and creative intention terms; for peer review, the specificity of proposals and emotions are extracted; and for audience feedback, the level of understanding and interaction preferences are extracted. [7] The similarity of textual semantics and the theme vector of "Campus Memories and Urban Life" is calculated to determine if the expression of the work deviates from the aims of the exhibition.

The thematic match score is calculated as follows:

$$M_i = \frac{\mathbf{s}_i \cdot \mathbf{q}}{\| \mathbf{s}_i \|_2 \| \mathbf{q} \|_2} \tag{3}$$

where $\mathbf{s}_i = g_\phi(T_i)$ represents the text vector T_i of the curatorial text by the i th student, as output by the semantic encoder $g_\phi(\cdot)$; this vector; $\mathbf{q}$ denotes the exhibition theme vector; M_i denotes the text–theme match score, with higher values indicating greater alignment between the curatorial description and the course theme.

3.3 Calculation of Artistic Expression Scores

The score is not simply a concatenation of raw inputs, but is produced via a type of "late" fusion. ConvNeXt generates a 768 dimensional image vector, and the semantic encoder is used to represent the teacher records, curatorial text, peer reviews, and audience feedback. These vectors are projected to 128 dimensions, layer–normalized and associated with a modality–availability mask. A "gated attention" layer learns sample–specific weights, after which to which the weighted vectors are fed to a two–layer regression head. Noisy ratings and small sample size are less of a problem with Huber loss, dropout, and weight decay, as the supervised target is the adjudicated teacher score[8].

The composite score for a student's work (i) is defined as:

$$S_i = \alpha V_i + \beta M_i + \gamma P_i + \delta A_i + \eta C_i, \quad \alpha + \beta + \gamma + \delta + \eta = 1 \tag{4}$$

where S_i is the comprehensive artistic expression score; V_i is the visual image score; M_i is the thematic alignment score; P_i is the peer review effectiveness score; A_i is the audience feedback score; C_i is the series

completeness score; and $\alpha, \beta, \gamma, \delta, \eta$ is a trainable weight used to dynamically adjust evaluation priorities according to different exhibition phases.

4 Optimization of the Club Course Promotion Model

4.1 Intelligent Course Theme Matching

The theme library for the course includes candidate themes, such as "Campus Memories", "Urban Life", "Digital Cultural Heritage", "Public Space Observations", and "Material Redesign", and each theme is linked to resource tags, material requirements, task difficulty, exhibition space requirements, and a vector representing the instructor's guidance experience [9]. The system reads the students' pre-test works, interest questionnaires, keywords from curatorial texts and school resource fields, and uses Top-K ranking to generate course theme recommendations; K is set to 3, and candidate themes with scores below 0.60 are excluded from the course proposal.

The theme matching score is calculated as:

$$R(q_j) = \lambda_1 E_j + \lambda_2 B_j + \lambda_3 D_j + \lambda_4 T_j - \lambda_5 L_j \tag{5}$$

In the formula, q_j represents the j th candidate topic; $R(q_j)$ is the topic recommendation score; E_j is the student interest match; B_j is the match between school materials and space; D_j is the availability of digital cultural resources; T_j is the match for teacher guidance capabilities; L_j is the implementation difficulty; $\lambda_1 \sim \lambda_5$ is the weight coefficient, and it satisfies $\sum_{r=1}^{5} \lambda_r = 1$. The ranking results are used to generate a 14-week sequence of club course topics.

4.2 Parameter Optimization for Cross-School Scaling

For cross-school implementation, leave-one-school-out calibration is adopted across the four participating schools. In each round, three schools jointly form the source domain $D_s^{(k)}$, while the remaining school is treated as the target domain $D_t^{(k)}$. The input variables include the initial score of the student, the frequency of teacher guidance, the facilities, the availability of materials, the exhibition area, and the audience-feedback scale. Instead of directly transferring one core school configuration, the model recalibrates task difficulty, theme openness, teacher intervention frequency, and AI scoring weights for each held-out school [10]. The optimization objective is:

$$\min_{\Theta} \mathrm{L} = \mathrm{L}_{score} + \mu \mathrm{MMD}\left(\mathrm{D}_s, \mathrm{D}_t\right) + \nu \mathrm{L}_{stab} \tag{6}$$

where Θ is the generalization parameter to be optimized; L_{score} is the AI scoring error; $MMD(D_s, D_t)$ is the difference in feature distributions between the core school and the collaborative schools; L_{stab} is the cross-school scoring stability loss; μ and V are adjustment coefficients, set to 0.35 and 0.20, respectively.

As shown in Figure 2, the sample distribution of core schools is concentrated in the regions of high topic matching and high work completion, and collaborative schools have scattered samples with low matching. The parameter optimisation module lowers the difficulty of the task, adds text scaffolding, adjusts evaluation weights, and thus makes the target domain samples converge to a stable generalisation area.

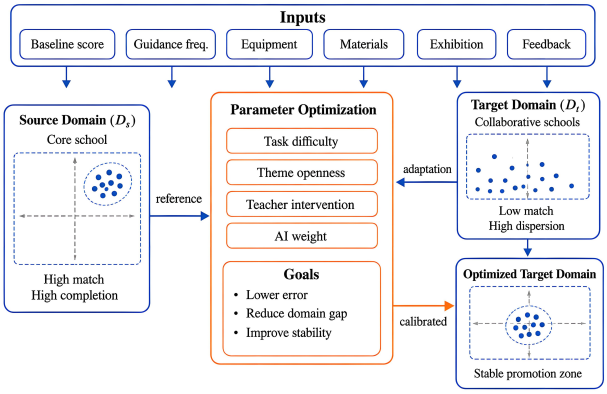


Figure 2: Cross-domain transfer analysis diagram

4.3 Model Iterative Updates

For model iteration, each exhibition cycle is the unit of iteration. The following data are saved in a versioned database: AI – teacher score deviations, low completion works, low confidence texts, modality conflicts, and audience feedback. Updates are generated when there is a disparity of more than 0.80 points, or a topic match of less than 0.60 and/or completion of less than 0.65. Since only 120 students are involved, the image and language backbones are not trained, only the projection, gated-fusion and theme-matching layers are trained. The learning rate is 1×10^{-4}, batch size 16, dropout 0.30, and weight decay 1×10^{-4}. Monitoring of overfitting is done by validation MAE, five seed variance, held-out-school error and manual-review demand, and early stopping is done using 8 epochs.

5 Experimental Results and Application Validation

5.1 Experimental Setup

The experiment was conducted with the 14th week "Campus Memories and Urban Life" project with one core

school and 3 partner schools. 240 artworks were collected and only 120 students were independent, so that the data set could be used for transfer learning and fusion–layer training, but not for training a visual backbone from scratch. Students' records were clustered together prior to their assignment to the three partitions to avoid having the same student's records appear in more than one partition. ConvNeXt and the text encoder were manually pretrained, drop–out was set at 0.30, weight–decay was $1 \times 10–4$ and early–stopping was 8 epochs. Training variance was estimated using five random seeds. 20 reference works were used with calibrated 10–point rubrics for the eight art teachers to score the 5 dimensions. Inter–rater reliability reached ICC(2, k)=0.86 (95% CI: 0.80 – 0.91). There was an agreement that cases with a difference of two or more points were rescored in blind fashion and where there was no resolution, a rubric discussion was conducted and the median awarded.Table 1 summarizes the configuration.

Table 1 Experimental Data and Parameter Configurations

Data Type	Total	Training Set	Validation Set	Test Set	Model Purpose
Images of Student Work	240	168	36	36	Visual Feature Extraction
Curatorial Commentary	120 copies	84	18	18	Thematic Semantic Matches
Peer–reviewed texts	360 entries	252	54	54	Feedback Validity Calculation
Audience Feedback Records	500	350	75	75	Public Communication Skills Evaluation
Teacher Teaching and Research Records	24 entries	17	3	4	Promotion Parameter Calibration

Note: Data partitions were created at the student level. Augmented versions of an artwork were retained within the same partition as the original image.

5.2 Analysis of Evaluation Results

Instead of just internal ablations, our evaluation results showed that the proposed fusion network outperforms three baselines: ViT–B/16 fine–tuned on images, CLIP–ViT/B32 using image – curatorial–text similarity, and MMBT that uses visual tokens and all textual records. The same student–grouped partitions, preprocessing and 5 random seeds were used in all the models. The metrics were MAE, expert–scores correlation, topic–matching accuracy,

manual–review rate and evaluation time. Two thousand bootstrap resamples were used to construct 95% confidence intervals, and paired permutation tests were used to compare predictions of the proposed model with the best baseline. Holm correction controlled comparisons were performed.Results are reported in Table 2.

Table 2 Performance and Statistical Comparison with Strong Multimodal Baselines

Evaluation Method	MAE (95% CI)	Expert Correlation	Topic Matching Accuracy/%	Manual Review Rate/%	Evaluation Time/s
ViT–B/16	0.67 (0.61 – 0.74)	0.79	77.9	24.7	8.5
CLIP–ViT/B32	0.58 (0.52 – 0.64)	0.83	82.9	19.6	8
MMBT	0.54 (0.49 – 0.60)	0.85	84.6	17.3	9.1
Proposed Gated Fusion Model	0.49 (0.44 – 0.55)	0.87	86.8	14.8	7.2

Note: The p–values were calculated against MMBT, which was the strongest baseline. The proposed model achieved p=0.018 for MAE and p=0.031 for topic–matching accuracy after Holm correction.

ViT–B/16 yielded good visual representation but was still limited if the theme relied on the curatorial intention and the audience's interpretation of the theme (Table 2). CLIP–ViT/B32 decreased MAE to 0.58 by aligning image with curatorial text, while MMBT to 0.54 by using textual evidence. The proposed gated fusion model achieved an MAE of 0.49 and a 95% confidence interval of [0.44, 0.55] and topic–matching accuracy of 86.8% and confidence interval of [82.1%, 91.0%]. The reduction in the MAE from MMBT was significant based on the paired permutation test (p=0.018) and the increase in the topic–matching accuracy was significant (p=0.031) after Holm correction. The range of MAE for the five seeds was 0.021, suggesting small instability with the limited number of seeds. The model pushed cases of low confidence and/or cross–modal conflict to teachers for manual review during deployment, which reduced the manual–review rate to 14.8%. In Figure 3, evidence is provided by artwork, curatorial statements, peer reviews, audience feedback, and teacher records, moving assessment from an individual grading process to a creation – interpretation – presentation – feedback process that is traceable.

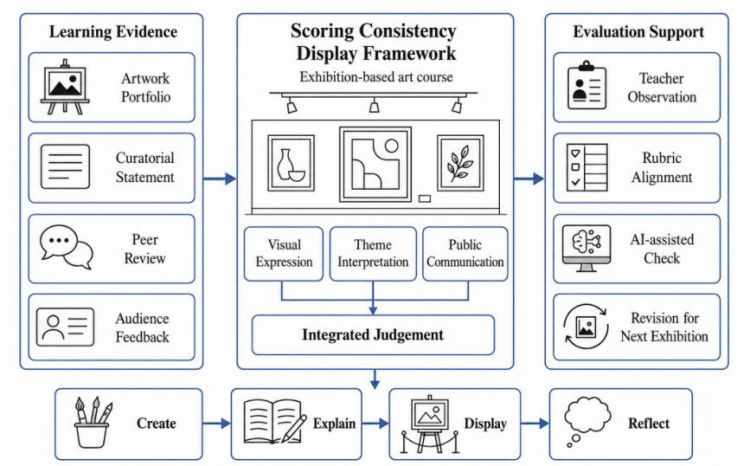


Figure 3: Distribution of Scoring Consistency for Exhibition–Based Learning

5.3 Comparison of Scaling Effects

To assess cross–school transfer, leave–one–school–out validation was used instead of a fixed source – target split. The four rounds included one school as a target and the other three schools as source–domain training schools with a resource descriptor only, but no target labels. The procedure was tried with variations in equipment, material available, frequency of teacher guidance, space for exhibition and participation of audience. Topic matching, series completion, curatorial awareness, evaluation time, low–completion proportion and coefficient of variation were indicators. The mean values, standard deviations and worst school result were recorded. Table 3 presents the outcomes.

Table 3: Leave–One–School–Out Results Before and After Parameter Calibration

Evaluation Metrics	Before Calibration	After Calibration	Worst Target After Calibration
Topic Matching Rate/%	72.6 ± 4.9	83.4 ± 3.3	78.7
Series Completion Rate/%	69.8 ± 5.3	78.1 ± 4.1	72.4
Curatorial Concept Score/Points	6.72 ± 0.48	7.54 ± 0.34	7.03
Teacher Evaluation Time/min · item^{-1}	4.9 ± 0.7	2.7 ± 0.4	3.3
Low–completion Samples/%	24.2 ± 5.6	15.6 ± 4.0	21.1
Cross–school Stability Coefficient	0.136 ± 0.018	0.091 ± 0.013	0.109

The average topic–matching rate during leave–one–school–out testing, after calibration, was 83.4% (before

calibration, 72.6%) while the worst was 78.7% as shown in Table 3. The percentage of series completion rose to 78.1% and the cross-school coefficient of variation dropped from 0.136 to 0.091. These changes reflect that schools experienced less distribution shift due to resource-aware task difficulty, text scaffolding, teacher intervention and scoring weights. Time of teacher evaluation decreased from 4.9 to 2.7 min · item−1 due to low confidence cases only being subject to manual review. There was some variation in performance, however, between the four target schools, with the greatest differences being found where the material resources and activities for the audience were most divergent. Thus, the findings can be considered applicable to the sampled urban junior high context and not universally transferable. It is important to note that figure 4 does not indicate complete domain invariance, but rather reduced dispersion. Cross validation on more rural, private and regionally varied schools, as well as prospective deployment cycles, are still needed before making the claim to generalization under real teaching conditions.

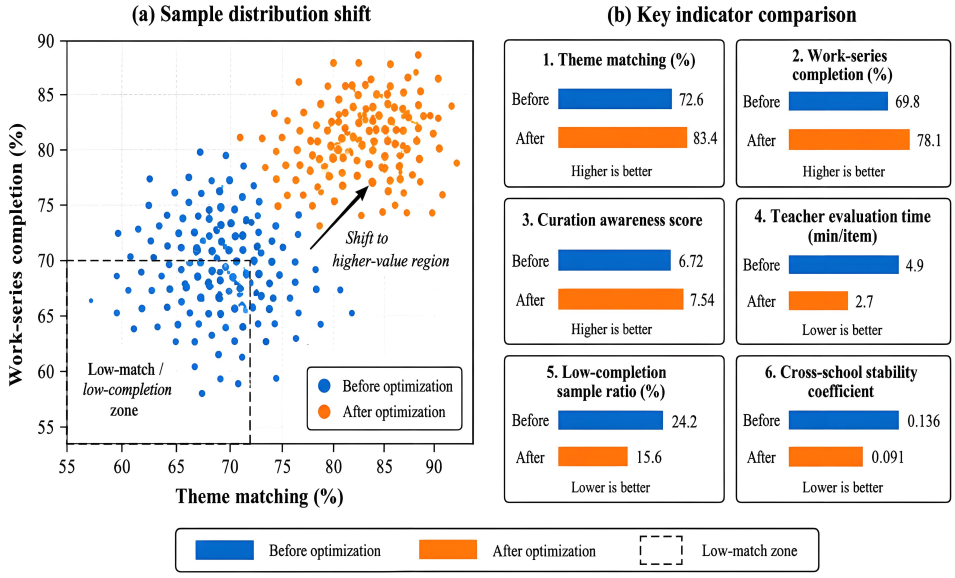


Figure 4: Multi-indicator Comparison Panel Chart

6 Conclusions

The model of AI-based evaluation and curriculum scaling in the exhibition-based learning model applied to junior high school art clubs was created. Artwork images, curatorial texts, peer reviews, audience feedback and teacher records were independently encoded and combined with masked gated attention, which enables the model to dynamically adjust contributions from modalities in the individual samples. The error rate was lower than that of

the ViT, CLIP, and MMBT baselines, and the effect was maintained with bootstrap and permutation analyses and pretrained encoders, with grouped splitting and regularization. There was also more consistent calibration across the four schools included in the leave-one-school-out validation. However, transferability across educational context and the possibility of unrestricted deep-model training cannot be demonstrated with 120 students. The reported results thus provide evidence of feasibility, but not generalization. Future work will include increasing the number of school types, regions, and exhibition themes, as well as longitudinal cycles, preregistering scoring rubrics, repeating reliability analysis, and testing federated, explainable and lightweight multimodal models in the classroom in an independent school setting.

References

[1] Daryanavard Chounchenani M, Shahbahrami A, Hassanpour R, et al. Deep learning-based image aesthetic quality assessment—a review [J]. ACM Computing Surveys, 2025, 57(7): 1—36.

[2] Soydaner D, Wagemans J. Multi-task convolutional neural network for image aesthetic assessment [J]. IEEE Access, 2024, 12: 4716—4729.

[3] Adilova A, Shamoi P. Personalized Aesthetic Assessment: Integrating Fuzzy Logic and Color Preferences [J]. IEEE Access, 2024, 12: 97646—97663.

[4] Redies C, Bartho R, Koßmann L, et al. A toolbox for calculating quantitative image properties in aesthetics research[J]. Behavior Research Methods, 2025, 57(4): 117.

[5] Bizzozzero N, Bendidi I, Risser-Maroix O. Prompt Performance Prediction for Image Generation[C]//2024 IEEE International Conference on Image Processing (ICIP). IEEE, 2024: 2208—2214.

[6] Kirstain Y, Polyak A, Singer U, et al. Pick-a-pic: An open dataset of user preferences for text-to-image generation[J]. Advances in Neural Information Processing Systems, 2023, 36: 36652—36663.

[7] Cosentino G, Giannakos M. Multisensory Interaction and Analytics to Enhance Smart Learning Environments: A Systematic Literature Review[J]. IEEE Transactions on Learning Technologies, 2023, 16(3): 414—430.

[8] Chejara P, Prieto L P, Dimitriadis Y, et al. The Impact of Attribute Noise on the Automated Estimation of Collaboration Quality Using Multimodal Learning Analytics in Authentic Classrooms[J]. Journal of Learning Analytics, 2024, 11(2): 73-90.

[9] Ouhaichi H, Spikol D, Vogel B. Rethinking MMLA: Design Considerations for Multimodal Learning Analytics Systems [C]//Proceedings of the Tenth ACM Conference on Learning@Scale. 2023: 354—359.

[10] Bayoudh K. A survey of multimodal hybrid deep learning for computer vision: Architectures, applications, trends, and challenges[J]. Information Fusion, 2024, 105: 102217.