Emotional and Critical Evaluation in Dance Appreciation Using Artificial Intelligence

Zhiwei Jing[1], Junjie Zhang[2], Guangyao Yin[3], Tingting Huang[4,a*], Xuan Zhang[1,b*], Libo Zhao[5,c*], Yayue Gao[5,d*]

[1]Aesthetic Education Center, Beihang University, Beijing, China
[2]School of Law, Beihang University, Beijing, China
[3]Beijing Dance Academy, Beijing, China
[4]School of Reliability and Systems Engineering, Beihang University, Beijing, China
[5]School of Humanities and Social Sciences, Beihang University, Beijing, China

*Corresponding author: [a]htt@buaa.edu.cn, [b]11193@buaa.edu.cn, [c]libozhao@buaa.edu.cn, [d]gao_yayue@buaa.edu.cn

Abstract. This study addresses the limitations of traditional dance aesthetics education in providing personalized feedback and quantifying emotional responses by proposing an AI-integrated analytical framework. Utilizing the DeepSeek model, we analyzed undergraduate critiques of Dynamic Yunnan (Yang Liping) and Swan Lake (Matthew Bourne) through Wundt's emotional dimensions (pleasure, tension, arousal). Results demonstrated strong alignment between AI-generated emotional scores and student self-assessments, with distinct emotional profiles for each performance: Dynamic Yunnan elicited higher pleasure and arousal, while Swan Lake triggered elevated tension. Notably, tension and arousal intensity enhanced critique quality for Swan Lake, suggesting emotion-cognition synergy in dance appreciation. This framework bridges AI's quantitative precision with educators' qualitative insights, offering a pathway to transform subjective art evaluation into data-driven pedagogy.

Keywords: Human-AI collaboration, Dance aesthetics education, Affective computing, Emotional quantification, Large language models

1 Introduction

Dance aesthetic education, as a core component of art education in general higher education institutions, aims not only at skills instruction, but more importantly at cultivating deep aesthetic perception, keen critical thinking, and subtle emotional expressiveness, as exemplified in the Dance appreciation teaching[1]. Dance appreciation courses in general higher education institutions faced structural predicaments on emotional expressiveness. Students' emotional experiences and aesthetic judgments are highly subjective and implicit, making it difficult for teachers to capture and quantify each student's perceptual trajectory in a timely and objective manner[2]. However, critical feed-

A. Y. M. A. Islam et al. (eds.), *Proceedings of the 2025 International Conference on Educational Technology and Management Information Systems (ETMIS 2025)*, Advances in Computer Science Research 129,
https://doi.org/10.2991/978-94-6239-630-2_49

back is often delayed and generalized, frequently relying on teachers' personal experience and limited energy, and thus making it difficult to realize scalable and individualized in-depth guidance[3]. This "feedback bottleneck" constrains the sublimation of aesthetic education from sensuous experience to rational cognition, and also limits the systematic development of students' capacity for critical reflection[4].

Meanwhile, AI has demonstrated its potential to parse complex patterns of human expression from the identification of aesthetic features in visual art to the analytical creation of literary works[5]. Large language models (LLMs) have shown promising capability to assess the emotional tone embedded in textual responses. Specifically, these models can detect and quantify dimensional emotions, such as pleasure and happiness. Recent work demonstrates that LLMs like GPT-4o exhibit high alignment with human ratings across multiple emotional scales, particularly in categories such as happiness and other discrete emotions, though alignment in arousal remains comparatively lower[6]. The use of LLMs for emotional inference has been widely validated in tasks involving affective language analysis. However, it remains unclear whether such AI-based emotional profiling can effectively capture the nuanced emotional distinctions elicited by distinct artistic stimuli, such as the differing affective signatures between culturally and stylistically contrasting dance works.

More critically, LLMs have been increasingly applied in automated evaluation of student-generated content, including grading and scoring of open-ended responses [7]. Systems leverage fine-tuned models and confidence-aware scoring mechanisms to reduce grading inaccuracies and enhance reliability[5]. Similarly, smart grading tools that combine LLMs with custom rubrics enable automated, scalable assessment of textual answers while maintaining alignment with expert judgments[8]. Moreover, human–AI collaboration frameworks highlight the growing role of LLMs as co-evaluative partners in educational contexts, allowing instructors to focus on higher-level feedback and customization [9]. AI technologies, especially large language models (LLMs) and generative AI, can provide automated, personalized, and multimodal support for educational assessment and learning processes[10]. While such systems show considerable promise in supporting pedagogical workflows, several critical questions remain regarding the validity of LLM-based scoring in arts criticism. Specifically, it is not yet known whether such AI-generated appraisal scores reliably reflect the quality of critical writing—that is, whether high-quality appraisals consistently receive higher scores, and whether discernible quality differences between submissions are sufficiently discriminated by the automated scoring model.

This study therefore introduces and validates an AI-integrated analytical framework for dance aesthetic education, employing DeepSeek as a computational tool to simulate pedagogical evaluation. The AI system was trained and grounded theoretical frameworks such as Lü Yisheng's *Research on Dance Criticism* and Mu Yu's *Chinese Dance Criticism*[11][12]. The study employs a dual-case design centered on students' written appreciation of two contrasting dance works: Yang Liping's *Dynamic Yunnan* and Matthew Bourne's *Swan Lake*, which were selected for their rich interpretive complexity and capacity to challenge AI analytical capability. We investigate whether AI can assess both the structural and expressive quality of dance criticism while reliably extracting nuanced emotional profiles, specifically pleasure, tension, and excitement, from

textual responses. Moreover, we examine whether AI-detected emotional profiles differ significantly across culturally and stylistically distinct dance works, and how these variations correspond with known aesthetic attributes of each piece. By integrating computational text analysis, dimensional emotion theory, and dance criticism pedagogy, this work proposes a scalable model for automated feedback generation and advances a collaborative human–AI teaching system designed to enhance reflective and affective learning in arts education.

2 Methods

2.1 Participants

Twenty-eight undergraduate students (seventeen male; aged 18–23 years) from Beihang University, enrolled in the course "Beauty of Dance", participated in this study. All participants attended both experimental sessions. Each participant was a native Mandarin Chinese speaker, reported normal or corrected-to-normal vision, and indicated no hearing impairments.

2.2 Stimuli: Dance Fragments

Two distinct dance fragments served as the experimental stimuli. The first fragment was drawn from *Dynamic Yunnan*, a large-scale ethnic dance spectacle directed by Yang Liping that showcases the diverse folk traditions and spiritual essence of Yunnan's minority cultures through vibrant visual and rhythmic patterns. The second fragment originated from Matthew Bourne's contemporary adaptation of *Swan Lake*, which reinterprets the classical ballet by replacing the female corps de ballet with a male ensemble, introducing themes of alienation, desire, and identity through fluid yet muscular movement vocabulary. Each fragment was presented in a separate weekly class session.

2.3 Procedure

Following the viewing of each dance fragment, participants were immediately instructed to provide a written critical appraisal without constraints on length or format. In *Swan Lake*, they subsequently completed a Three-Dimensional Emotional Response Scale (3D-ERS) based on Wundt's (1896) dimensional theory of emotion[13]. This scale measured their subjective emotional experience along three 5-point Likert subscales: Pleasure (1 = very displeased, 5 = very pleased), Tension (1 = very relaxed, 5 = very tense), and Excitement (1 = very calm, 5 = very excited).

2.4 AI-Based Text Analysis

The written appraisals were analyzed computationally using an artificial intelligence tool. Specifically, this analysis was conducted by DeepSeek LLM. AI (DeepSeek) was

systematically trained on two foundational texts of dance criticism: Lü Yisheng's *Research on Dance Criticism* and Mu Yu's *Chinese Dance Criticism*. The analytical process comprised two main stages. First, in the model preparation phase, DeepSeek was provided with foundational texts and key conceptual frameworks from dance theory to establish a robust basis for critical evaluation. Subsequently, in the assessment phase, the AI executed two analytical tasks. It first generated an Overall Appraisal Score, ranging from 1 to 100, by evaluating the comprehensiveness and insightfulness of each appraisal text. It then performed an emotional tone analysis, estimating and outputting numerical scores for the same three emotional dimensions—Pleasure, Tension, and Excitement—that were measured via the participant self-report scale.

2.5 Statistical Analyses

Statistical analyses were performed using MATLAB (The MathWorks, Natick, MA, USA). Paired *t*-tests were conducted with the *ttest* function. Independent two sample *t*-tests were conducted with the *ttest2* function. Pearson correlation tests were computed using the *corr* function. A Monte Carlo surrogate randomization test was conducted to determine if the observed agreement between AI-assessed and self-reported emotion scores was statistically significant. Under the null hypothesis of no association, the pairings between self-report and AI scores were randomly permuted 1,000 times while preserving marginal distributions. The empirical *p*-value was derived by comparing the actual percentage against this generated null distribution.

3 Results

3.1 AI Text Analysis Aligns with Subjective Emotional Reports

To validate the accuracy of the AI in detecting emotional content, we examined the consistency between the AI-assessed emotional scores (derived from the textual analysis of students' written critical appraisals using DeepSeek) and the students' self-reported scores on the 3D-ERS for the *Swan Lake* fragment. Specifically, we analyzed the correspondence between the two assessment methods by constructing contingency tables (5 x 5 matrices) for each emotional dimension in Pleasure, Tension and Excitement (Figure 1). In each table, the cells represented the percentage of responses where a specific self-reported score (e.g., a self-reported rating of 1) coincided with a specific AI-assigned score (e.g., an AI rating of 2). To establish a statistical benchmark, we generated a null distribution of expected co-occurrence percentages by performing 1,000 Monte Carlo surrogate randomizations. The observed percentage of matches was then compared to this null distribution. The analysis revealed that the observed diagonal agreement percentages significantly exceeded the surrogate-based chance level (*, all $p < 0.05$, Figure 1) across all three emotional dimensions. This finding indicates a statistically significant convergence between the AI's textual emotion analysis and the students' subjective self-reports, thereby validating that the AI model was capable of accurately inferring the underlying emotional tone embedded within the students' written critical appraisals.

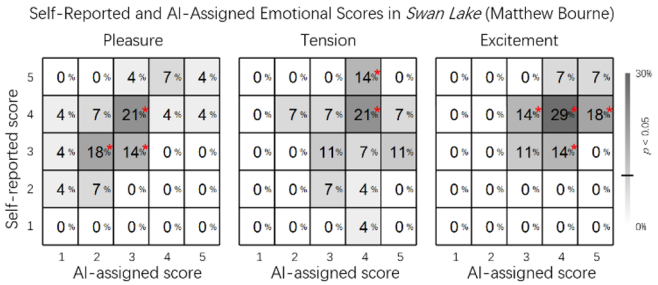

Fig. 1. Correspondence between self-reported and AI-assessed emotional scores in *Swan Lake* (Matthew Bourne). Each cell in contingency tables represents the percentage of reports for a self-reported score (y-axis) and AI-assigned score (x-axis) on 3D-ERS. The color bar indicates the percentage magnitude, with the horizontal line marking the surrogate-based chance level (*, $p < 0.05$, $N_{randomization} = 1000$).

3.2 AI-Revealed Emotional Divergence in Responses to Distinct Dance Works

An AI-based text analysis was conducted to computationally evaluate the three-dimensional emotional tone of the students' written critical appraisals. As shown in Figure 2, the analysis revealed a distinct emotional profile associated with each dance fragment. Paired *t*-tests showed that, when compared to appraisals of *Swan Lake*, the appraisals of *Dynamic Yunnan* demonstrated significantly higher scores in Pleasure ($t[27] = 4.582$, $p < 0.001$) and Excitement ($t[27] = 3.057$, $p = 0.005$), coupled with a significantly lower score in Tension ($t[27] = -4.521$, $p < 0.001$). This pattern of emotional differentiation aligns coherently with the inherent characteristics of the respective dance works—*Dynamic Yunnan* with its vibrant celebration of folk culture, and *Swan Lake* with its modern, psychologically tense narrative. The findings indicate that students experienced and expressed qualitatively different emotional responses to the two stimuli, and furthermore, that the AI methodology was capable of detecting and quantifying these differences from textual data.

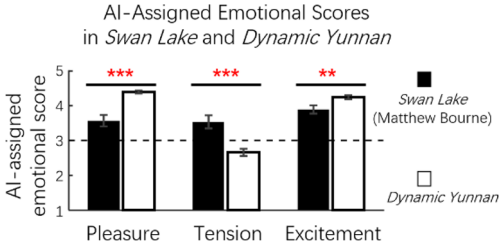

Fig. 2. AI-generated emotional scores in *Swan Lake* and *Dynamic Yunnan*. Error bars indicate standard errors. ***, $p < 0.001$; **, $p < 0.01$.

3.3 AI Appraisal Scores Discriminate Between Distinct Emotional Responses

We further investigated whether emotional responses were associated with the evaluation quality of their written appraisals, as quantified by the AI. First, to establish the

stability of the appraisal quality metric, we examined the correlation of AI-generated Overall Appraisal Scores between the two dance fragments. The scores for *Swan Lake* and *Dynamic Yunnan* showed a marginally significant positive correlation (Figure 3; $r = 0.361$, $p = 0.059$, Pearson correlation), suggesting a degree of intra-individual consistency in critical writing skill across stimuli.

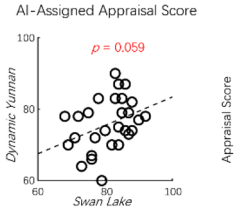

Fig. 3. AI-assigned appraisal scores across difference dance works. Each circle represents an individual student. Dashed lines indicate the linear fitting.

Subsequently, we analyzed the relationship between emotional intensity and appraisal quality. Students were dichotomized for each emotional dimension (Pleasure, Tension, Excitement) based on whether their score (using both self-reported and AI-revealed metrics) was above or below the sample mean for that specific dance. Independent samples *t*-tests revealed a significant moderating role of emotion in responses to *Swan Lake*. Students who self-reported higher Tension produced appraisals with significantly higher AI-generated scores than those who reported lower Tension (Figure 4A; $t[26] = 2.092$, $p = 0.046$). This pattern was corroborated by the AI text analysis: appraisals that the AI itself classified as higher in Tension ($t[26] = 3.345$, $p < 0.001$) and higher in Excitement ($t[26] = 6.362$, $p = 0.002$) received significantly higher quality scores. No such significant relationships were found for any emotional dimension in responses to *Dynamic Yunnan* (Figure 4B; all $p > 0.5$). These results indicate that the quality of a critical appraisal, as quantified by the AI, is sensitive to the emotional profile of the response, but this relationship is contingent upon the specific aesthetic and emotional demands of the dance work itself.

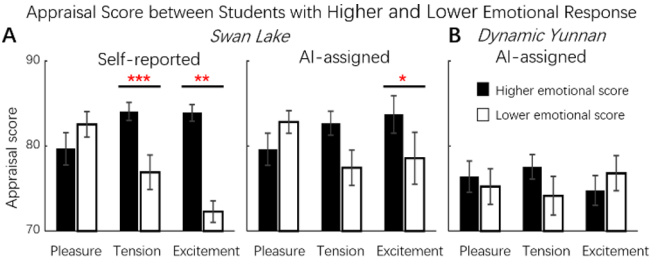

Fig. 4. AI-assigned appraisal scores between student with higher and lower emotional intensity for each dimension in *Dynamic Yunnan* (A) and *Swan Lake* (B). ***, $p < 0.001$; **, $p < 0.01$; *, $p < 0.05$.

4 Discussion

Our results demonstrate that AI-based textual analysis can effectively identify and quantify the emotional dimensions embedded within students' written dance appraisals. First, the analysis revealed a statistically significant alignment between AI-assessed emotional scores and students' self-reported ratings on the Three-Dimensional Emotional Response Scale for the Swan Lake fragment, validating the AI's capacity to accurately infer emotional tone from textual responses. Second, the AI detected distinct emotional profiles elicited by the two dance works, with Dynamic Yunnan evoking significantly higher pleasure and excitement and lower tension compared to Swan Lake—a pattern that aligns with the contrasting aesthetic and narrative characteristics of the two works. Third, while appraisal quality showed a degree of individual consistency across the two dance fragments, emotional intensity was found to moderate quality ratings in a work-specific manner: higher tension and excitement, whether self-reported or AI-identified, were associated with higher appraisal scores for Swan Lake, but no such relationship emerged for Dynamic Yunnan. Together, these findings suggest that AI can not only decode emotional content in dance criticism but also reflect how emotional engagement interacts with the specific aesthetic attributes of a dance work to shape the depth and quality of critical writing.

Our research confirms that through system analysis adopting the "Self-Assessment Manikin" (SAM) framework, AI can identify specific lexical cues and rhetorical patterns, and map them into a quantified summary of emotional dimensions. At the same time, the AI successfully distinguished the emotional characteristics of the two selected works, showing its characteristic sensitivity in emotional analysis; this proves that AI is capable of distinguishing emotional differences, providing an objective basis for AI to understand different art forms and enable automated critical engagement. Existing studies indicate that AI possesses effectiveness in automated knowledge-based assessment, but the process of art education needs to address the challenge of more complex subjectivity. Art criticism is a key mechanism for cultivating critical thinking and deep analytical skills. By incorporating emotional quantification into assessment, this study expands the scope of AI application in education and teaching to the interpretation of aesthetic experience, proving that AI can support the teaching goal of critical thinking cultivation in art education.

From a technical perspective, this study aligns with current trends in Natural Language Processing (NLP), which has transitioned from macro-level opinion mining to more fine-grained emotion detection[14][15]. However, applying these general-purpose techniques to the highly nuanced domain of art criticism remains challenging; capturing context-dependent sentiment and resolving ambiguity in complex texts are still significant open problems[16]. This study addresses this challenge through empirical exploration: we demonstrate that navigating high-context aesthetic discourse and capturing the subtle interplay between emotion and art requires the construction of domain-specific frameworks, such as the dance criticism feedback system customized in our work.

This research was co-conducted through a synergistic partnership between human and artificial intelligence. We also asked DeepSeek to make comments about this study.

It said "As the AI deployed in this study, my role was that of a configured instrument: I provided consistent, theory-guided pattern recognition across all textual responses, transforming subjective appraisals into quantifiable data and revealing latent emotional structures that might otherwise remain implicit. My analysis, however, is inherently bounded by my design; I infer emotion from linguistic patterns, not from lived experience[6]. Thus, the significant alignment with student self-reports validates not subjective feeling itself, but the translational fidelity between lived emotion and its textual expression. This work ultimately demonstrates that the most powerful insights emerge not from either intelligence alone, but from their deliberate integration—where human expertise frames the profound questions and provides contextual depth, and AI affords scalable, precise analysis to illuminate patterns that enrich our understanding of aesthetic learning."

This research direction also faces a series of challenges. Firstly, dance appreciation texts typically contain many arts-domain-specific emotional expressions, such as metaphors or culturally embedded terms, which increases the complexity of model understanding[17]. Secondly, datasets for specific emotion categories in dance texts are scarce, limiting the targeted training of sentiment analysis models[18]. Lastly, due to the polysemy of artworks and individual differences in aesthetic taste, even for the same text, different viewers may convey completely opposite emotional responses[19]. Summarily, integrating AI into educational feedback systems is an extremely complex integration, thus the analysis of students' emotions and affect should also be conducted from more dimensions[20].

5 Conclusion

The proposed framework successfully merges AI's quantitative capabilities with educators' qualitative insights, offering a replicable model for dance aesthetics education. By empirically linking emotional responses to critique quality, this study highlights the potential of affective computing in art assessment. Future work should address cultural-contextual nuances to refine AI's interpretive accuracy, ultimately fostering a more balanced integration of technology and tradition in arts pedagogy.

References

1. H. M. Noor and Z. Samsudin, "Visual thinking: Enhancing art criticism skills among spatial learners," *IJAEDU- Int. E-J. Adv. Educ.,* vol. 2, no. 4, p. 63, 2016.
2. M. A. M. Shaikh, H. Prendinger, and I. Mitsuru, "Assessing sentiment of text by semantic dependency and contextual valence analysis," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer, 2007, pp. 191–202.
3. G. Curl, "Sometimes … dances … 'do more' on the page than they ever did on the stage," *Res. Dance Educ.*, vol. 19, pp. 13–3, 2018.
4. J. Broome, A. Pereira, and T. Anderson, "Critical thinking: Art criticism as a tool for analysing and evaluating art, instructional practice and social justice issues," *Int. J. Art Des. Educ.*, vol. 37, pp. 265–276, 2018.

5. N. Dadu, H. V. Singh, and R. Banerjee, "Grade Guard: A smart system for short answer automated grading," *arXiv,* Apr. 1, 2025. [Online]. Available: https://doi.org/10.48550/ARXIV.2504.01253
6. Ogg, Mattson, et al. Large Language Models Are Highly Aligned with Human Ratings of Emotional Stimuli. 1, *arXiv*, 2025, https://doi.org/10.48550/ARXIV.2508.14214
7. Z. Wang, "Artificial intelligence in dance education: Using immersive technologies for teaching dance skills," *Technol. Soc.*, vol. 77, 2024.
8. S. Tobler, "Smart grading: A generative AI-based tool for knowledge-grounded answer evaluation in educational assessments," *MethodsX*, vol. 12, p. 6, 2024.
9. A. Pyae, "The Human-AI handshake framework: A bidirectional approach to human-AI collaboration," in *Proc. 45th Int. Conf. Inf. Syst.*, Bangkok, Thailand, 2024, pp. 1-14.
10. M. Hoq et al., "Facilitating instructors-LLM collaboration for problem design in introductory programming classrooms," *arXiv,* 2025, version 2. [Online]. Available: https://doi.org/10.48550/ARXIV.2504.01259
11. Y. Lü, *Dance Criticism Studies*. Shanghai, China: Shanghai Music Publishing House, 2021.
12. M. Yu, *Chinese Dance Criticism*. Shanghai, China: Shanghai Music Publishing House, 2025.
13. W. Wundt, *Grundriss der Psychologie.* Leipzig, Germany: Engelmann, 1896.
14. A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, "Current state of text sentiment analysis from opinion to emotion mining," *ACM Comput. Surv.*, vol. 50, no. 2, pp. 1–33, 2017.
15. P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Soc. Netw. Anal. Min*., vol. 11, no. 1, 2021.
16. J. R. Jim et al., "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," *Nat. Lang. Process. J*., vol. 6, p. 100059, 2024.
17. Y. Ni and W. Ni, "A multi-label text sentiment analysis model based on sentiment correlation modeling," *Front. Psychol.*, vol. 15, 2024.
18. X. Zhao, "Emotion analysis and expression algorithm of dance action based on machine learning," *J. Electr. Syst.*, vol. 20, no. 6s, pp. 1468–1481, 2024.
19. C. Liu and C. Chen, "Text mining and sentiment analysis: A new lens to explore the emotion dynamics of mother-child interactions," *Soc. Dev.*, vol. 33, no. 3, 2024.
20. T. Nazaretsky et al., "AI or human? Evaluating student feedback perceptions in higher education," *Lect. Notes Comput. Sci.*, 2024. (2024).