Construction of a Multimodal Learning-Based Emotion Computing Model for Dance Movements, Clothing, and Aesthetic Education

Shuanghe Li, Yanlin Deng
Beijing Coll Finance & Commerce, Beijing 101101, China.
Corresponding author: Shuanghe Li (e-mail: 15652102268@163.com)

ABSTRACT Modeling aesthetic perception from dynamic visual information requires simultaneous understanding of human motion and the physical behavior of textile materials, making multimodal feature fusion a critical challenge in intelligent perception systems. Similar information integration strategies have also attracted increasing attention in electromagnetic sensing and heterogeneous signal interpretation, where multiple data sources must be jointly analyzed to characterize complex targets. This study proposes a multimodal emotion computing framework (DC-AEM) that combines dance kinematics and garment dynamics for computational aesthetic assessment. The framework employs a GCN-LSTM network to encode spatio-temporal skeletal motion and constructs a dedicated costume representation by integrating color distribution, texture descriptors, shape characteristics, and optical-flow-based dynamic features through a convolutional architecture. A cross-modal attention mechanism is introduced to adaptively associate movement patterns with context-dependent fabric behavior, generating a unified representation for predicting gracefulness, power, and fluidity. Experimental evaluation on a ballet performance dataset demonstrates that the proposed model significantly outperforms movement-only and simple fusion baselines, achieving superior correlation and lower prediction error across all aesthetic dimensions. The results confirm that dynamic textile characteristics provide essential complementary information for affective perception and establish an effective multimodal analysis strategy that may offer methodological reference for intelligent visual sensing, electromagnetic perception systems, and data-driven costume design.

INDEX TERMS multimodal learning, cross-modal attention, garment dynamics, affective computing, electromagnetic perception

I. INTRODUCTION

DANCE, as a universal form of artistic expression, embodies a complex integration of human motion and visual artistry, serving as a cornerstone of cultural heritage and aesthetic education [1,2]. The appreciation of a dance performance is a deeply subjective and holistic process, in which the audience interprets not only the choreographer's intent and the dancer's skill but also the powerful nonverbal narrative conveyed by the interplay between body and attire [3,4]. The costume, far from being a mere accessory, is an active participant in the performance; it can amplify, constrain, or transform the perception of movement [5,6]. The color of a leotard can evoke passion, the drape of a skirt can accentuate the fluidity of a turn, and the texture of a fabric can suggest character and mood [7]. Understanding and quantifying this complex synergy between dance movement and costume presents a significant challenge, yet it holds the key to unlocking new paradigms in digital arts, performance analysis, and aesthetic pedagogy [8].

Rapid advancements in artificial intelligence, particularly in computer vision and deep learning, have provided powerful tools for analyzing complex human behaviors [9]. Affective computing, a field dedicated to developing systems that can recognize, interpret, and simulate human emotions, has made significant strides in areas such as facial expression analysis and speech sentiment recognition [10]. However, the application of these techniques to the nuanced domain of dance aesthetics remains relatively nascent. Early computational approaches to dance analysis have predominantly focused on movement kinematics, employing techniques such as pose estimation and motion capture to analyze action recognition, style classification, and skill assessment [11]. While valuable, these studies often overlook the integral role of the visual spectacle, particularly the costume, thereby

neglecting a critical channel of affective information. The costume is not just a static visual element but a dynamic object whose properties—material, form, and behavior in motion— contribute significantly to overall aesthetic judgment.

This research addresses this critical gap by proposing a novel affective computing model that explicitly integrates dance movement and costume analysis through a multimodal learning framework. We argue that a truly comprehensive understanding of dance aesthetics can only be achieved by treating movement and costume as coupled, synergistic modalities. The core hypothesis of our work is that costume features, when analyzed in conjunction with kinematic data, can significantly improve the accuracy and robustness of computational models designed to predict human aesthetic responses to dance. To test this hypothesis, we construct a deep learning architecture that processes spatio-temporal movement data and dynamic costume features through separate, specialized neural network streams. These streams are then intelligently fused using an attention mechanism that allows the model to learn which aspects of the costume are most relevant to specific movements in predicting a particular aesthetic quality. By mapping these complex inputs to a defined affective space, our model provides a quantitative, objective measure of aesthetic perception. The contribution of this paper is threefold: first, we introduce a novel multimodal dataset and annotation scheme for dance aesthetic analysis; second, we develop and validate a new deep learning model, DC-AEM, that effectively fuses movement and costume information; and third, we provide empirical evidence demonstrating the critical importance of the costume modality in the computational assessment of dance, paving the way for new applications in intelligent costume design and data-driven aesthetic education.

II. RELATED WORK

A. COMPUTATIONAL ANALYSIS OF DANCE MOVEMENT

The quantitative analysis of dance has a rich history, evolving from early biomechanical studies to modern data-driven approaches. The advent of markerless motion capture, particularly skeleton-based pose estimation algorithms such as OpenPose, has revolutionized the field [12]. These tools allow for the extraction of 2D or 3D skeletal keypoints from video, providing a rich representation of human kinematics. Researchers have leveraged this data for various tasks. For instance, deep learning models, especially Recurrent Neural Networks (RNNs) and their variants such as Long Short-Term Memory (LSTM) networks, have been widely used to model the temporal dependencies in movement sequences for dance genre classification and quality assessment [13,14]. More recently, Graph Convolutional Networks (GCNs) have shown exceptional promise by modeling the human skeleton as a graph, thereby explicitly capturing the spatial relationships between body joints [15]. The combination of GCNs and LSTMs has become a state-of-the-art method for skeleton- based action recognition, providing a powerful foundation for our movement analysis stream [15]. However, these works primarily focus on the geometric and dynamic properties of the body itself, seldom incorporating external visual cues such as apparel.

B. AFFECTIVE COMPUTING AND COMPUTATIONAL AESTHETICS

Affective computing aims to bridge the gap between human emotion and computational technology [16]. In the visual domain, computational aesthetics focuses on predicting the aesthetic appeal of images or videos [17]. Early works relied on hand-crafted features based on principles of photography and art theory, such as color harmony, the rule of thirds, and texture [18]. With the rise of deep learning, Convolutional Neural Networks (CNNs) have become the dominant tool, learning complex aesthetic features directly from large-scale datasets of images with human-provided ratings. These models have been successfully applied to tasks such as photo ranking and style analysis [19,20]. While powerful, applying these models directly to dance is challenging because aesthetic judgment is not based on a single static frame but on the dynamic evolution of visual elements over time [21]. Furthermore, existing research has not adequately addressed the specific affective contributions of clothing within a dynamic, performance-based context.

C. TEXTILE AND COSTUME IN COMPUTER VISION

The analysis of clothing in images and videos is a significant area of research in computer vision, with direct relevance to the textile and fashion industries [22]. Key tasks include clothing parsing (segmenting an image into different garment items), attribute recognition (identifying categories, colors, textures), and virtual try- on [23,24]. Texture analysis, using methods such as Gray-Level Co-occurrence Matrices (GLCM) or deep texture descriptors, is particularly relevant for quantifying fabric properties from visual data [25]. Some research has explored the physics-based simulation of fabric drape and motion, which is crucial for virtual fashion design. However, the majority of this work focuses on recognizing or simulating garments in relatively static or everyday scenarios [26,27]. The unique challenge in dance is to understand the affective consequence of a costume's dynamic behavior—how the fabric's movement, driven by the dancer's actions, influences aesthetic perception. Our work is among the first to computationally link quantifiable dynamic costume features to high-level aesthetic and emotional ratings in a performance art context.

D. MULTIMODAL LEARNING

Many real-world phenomena are better understood by considering information from multiple sources or modalities. Multimodal learning is a branch of machine learning that focuses on building models that can process and relate information from heterogeneous data types, such as video, audio, and text. A key challenge in this field is feature

fusion—the process of combining information from different modalities [28,29]. Fusion strategies range from simple concatenation (early fusion) to the combination of outputs from separate models (late fusion). More sophisticated techniques, such as cross-modal attention mechanisms, have recently gained prominence. Attention allows a model to dynamically weight the importance of features from one modality based on the context provided by another [30,31]. This is particularly suitable for our task, as the relevance of a costume's features (e.g., its flow) is highly dependent on the dancer's current movement (e.g., a leap versus a stationary pose). Our adoption of a cross-modal attention framework is therefore a principled choice to model the intricate interplay between dance and costume.

III. METHODOLOGY

Our proposed methodology is centered around the Dance Movement-Costume-Aesthetic Education Model (DC-AEM), a multimodal deep learning framework designed to predict affective scores from dance videos.

The overall architecture, depicted in Figure 1, involves two parallel feature extraction streams—one for movement and one for costume—followed by an attention-based fusion module and a final regression head for affective prediction.

A. DATA ACQUISITION AND ANNOTATION

To train and evaluate our model, we curated a specialized dataset, termed "BalletAesthetics-200," consisting of 200 video clips, each 10–15 seconds in duration. The clips were sourced from high-definition recordings of professional solo ballet performances, ensuring high visual quality and consistency in artistic standards.

Specifically, all video clips were standardized to a resolution of 1,920 $\times$ 1,080 pixels and a frame rate of 30 fps.

To ensure dataset diversity, the distribution of costume categories was balanced, comprising approximately 35% classical tutus, 35% romantic tutus, and 30% leotards. Furthermore, given the dataset size, we employed data augmentation techniques, including random horizontal flipping, random cropping, and scale jittering, during the training phase to enhance the model's generalization capability and robustness.

For ground truth annotation, we recruited a panel of 20 participants, comprising 10 domain experts (dancers, choreographers, and costume designers) and 10 laypersons with an interest in dance. In selecting the aesthetic dimensions for annotation, we focused on concepts that are fundamentally expressed through the physical and dynamic interplay between the dancer's body and their attire, which is the central focus of our computational model. Therefore, we chose "Gracefulness," "Power," and "Fluidity" as our primary dimensions.

These were selected over more abstract or narrative concepts (e.g., "emotional expression" or "storytelling") because they are more directly observable from the visual evidence of kinematics and fabric dynamics. This targeted selection is crucial for creating a robust ground truth for a model that analyzes physical motion, and it also aids in achieving higher inter-rater reliability among annotators. Participants were asked to watch each video clip and rate it on a 7-point Likert scale across the three primary aesthetic dimensions: Gracefulness (smooth, controlled, and elegant motion), Power (dynamic strength, athleticism, and impact of movement), and Fluidity (seamlessness of transitions and the perceived flow of the dancer and their costume). The final ground truth score for each dimension was calculated as the average rating across all participants. Inter-rater reliability was assessed using Krippendorff's alpha, yielding a satisfactory coefficient of $\alpha = 0.78$, indicating substantial agreement among the annotators. Further analysis was conducted to compare the agreement between the domain experts and laypersons. The Krippendorff's alpha for the expert-only group was 0.82, while that for the layperson-only group was 0.74. The inter-group alpha was calculated to be 0.76, suggesting that while the experts demonstrated slightly higher internal consistency, the aesthetic judgments of both groups were largely aligned. This finding confirms the holistic nature of dance aesthetics and validates our approach of using a combined panel for annotation.

IV. MULTIMODAL FEATURE EXTRACTION

A. MOVEMENT MODALITY: SPATIO-TEMPORAL GRAPH REPRESENTATION

The movement feature extraction pipeline aims to capture the detailed kinematics of the dancer's body. To eliminate the influence of camera distance and the dancer's absolute position, these raw coordinates were subsequently normalized to a range of [–1, 1] relative to the center of the frame. This results in a sequence of skeletal graphs for each video clip. To effectively model both the spatial configuration of the body and its temporal evolution, we utilized a GCN-LSTM architecture.

The GCN component processes the skeletal data at each frame. Specifically, we adopted the spatial graph convolution formulation proposed by Kipf and Welling to aggregate features from neighboring joints, capturing the topological structure of the skeleton within each static time step. It treats the skeleton as a graph G= (V, E), where V is the set of joints (nodes) and E is the set of natural connections between them (edges, i.e., limbs) are defined. The GCN learns to aggregate features from neighboring joints, capturing complex spatial dependencies such as the relative orientation of limbs. The output of the GCN for each frame is a feature vector that represents the body's posture. The sequence of these feature vectors is then fed into an LSTM network. The LSTM, with its inherent ability to model long-range temporal dependencies, learns the dynamic patterns of the dance sequence, producing a final context vector f_{mov} that encapsulates the overall movement characteristics. Our GCN-LSTM model consisted of a 3-layer GCN with 256 output channels per layer, followed by a 2-layer LSTM with 128 hidden units. We applied the ReLU activation function

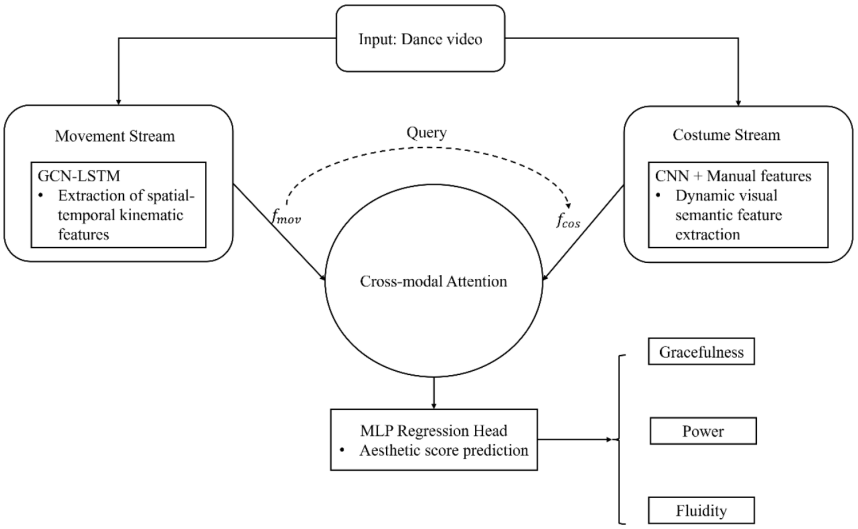

FIGURE 1. Conceptual diagram of the proposed DC-AEM model. The architecture consists of two parallel streams for processing movement (GCN-LSTM) and costume (CNN and handcrafted features) data. A cross-modal attention mechanism fuses these features, and a final MLP regression head predicts the aesthetic scores for Gracefulness, Power, and Fluidity

after each GCN layer (while the LSTM utilized its standard internal activations) and included a dropout rate of 0.3 to prevent overfitting.

B. COSTUME MODALITY: DYNAMIC VISUAL-SEMANTIC FEATURES

The costume feature extraction stream is designed to quantify the visual and dynamic properties of the dancer's attire, which is critical for linking the analysis to textile properties. This process involves three main steps:

- Costume Segmentation: To isolate the costume from the dancer and background, we fine-tuned a pre-trained Mask R-CNN model on a small, manually annotated subset of our data. This network generates a binary mask for the costume region in each frame.
- Static Feature Extraction: For each masked costume region, we extracted a set of features designed to represent its material and form. These included a 128-bin HSV color histogram to capture the dominant colors and their distribution; texture features derived from a GLCM, including contrast, correlation, and energy, to represent the fabric's surface properties; and Hu moments to provide a scale- and rotation-invariant description of the costume's overall shape.
- Dynamic Feature Extraction: To capture how the costume moves, we computed the dense optical flow between consecutive segmented costume masks. The magnitude and orientation of the flow vectors were summarized into a histogram, providing a quantitative measure of the costume's dynamic behavior (e.g., the amount of "billowing" or "swirling"). Specifically, we constructed a 2D joint histogram with 8 bins for orientation (covering 360°) and 10 bins for magnitude (normalized to a range of 0–1). The resulting 80-bin histogram was then flattened into a feature vector. This representation captures the joint distribution of movement speed and direction, effectively quantifying complex dynamics such as uniform flowing (a peak in one orientation bin) versus turbulent swirling (a wider distribution across multiple bins).

For each frame, the features extracted in the steps above—the 128-bin HSV color histogram, the 3 GLCM texture features, the 7 Hu moments, and the 80-bin dynamic optical flow histogram—are concatenated to form a single, comprehensive feature vector of 218 dimensions. This vector represents the complete state of the costume at that specific moment. For a video clip with T frames, this process generates a feature sequence of shape $T \times 218$, which serves as the input to a one-dimensional Convolutional Neural Network (1D CNN).

The 1D CNN applies its convolutional filters along the temporal (T) dimension, enabling it to learn salient short-term patterns from the evolution of the costume's visual and dynamic properties (e.g., the rapid change in texture and flow during a spin). The resulting feature maps are then passed through a global average pooling layer to produce the final, fixed-size feature vector, which encapsulates the costume's overall contribution.

C. ATTENTION-BASED MULTIMODAL FUSION

A simple concatenation of movement and costume features fails to capture the nuanced, context-dependent relationship between them. Therefore, we implemented a cross-modal attention mechanism. This mechanism allows the model to dynamically weight the importance of costume features based on the current movement context.

Specifically, the movement feature vector f_{mov} acts as the query, while the costume feature vector f_{cos} serves as both the key and the value. This design choice reflects the causal nature of the performance: the aesthetic relevance of the costume's dynamic state (e.g., swirling vs. static) is conditioned on the dancer's kinematic action. Therefore, movement

serves as the contextual "Query" to retrieve the most salient visual attributes from the costume ("Key/Value"). We opted for a single-head, unidirectional attention mechanism to maintain model parsimony and prevent overfitting on our specialized dataset, focusing specifically on how costume features support movement rather than modeling complex bidirectional dependencies. The attention weights are calculated as:

$$\alpha = \text{softmax}\left(\frac{(f_{\text{mov}}W_Q)(f_{\text{cos}}W_K)^T}{\sqrt{d_k}}\right) \quad (1)$$

where, W_Q and W_K are learnable and d_k is the dimension of the key vector. The attended costume feature vector f'_{cos}, is then computed by applying these weights to the original costume features:

$$f'_{\text{cos}} = \alpha(f_{\text{cos}}W_V) \quad (2)$$

This attended costume feature vector, which emphasizes costume properties most relevant to the given movement, is then concatenated with the original movement feature vector f_{mov}. Specifically, the movement feature vector f_{mov} has a dimension of 128 (matching the LSTM hidden size), and the costume feature vector f_{cos} acts as a 256-dimensional embedding. The learnable weight matrices are defined as $W_Q \in \mathbb{R}^{128\times d_k}$, $W_K \in \mathbb{R}^{256\times d_k}$, and $W_V \in \mathbb{R}^{256\times 256}$, where we set the attention key dimension $d_k = 128$. Consequently, the final fused vector f_{fused} is formed by concatenating the 128-dimensional movement vector and the 256-dimensional attended costume vector, resulting in a total dimensionality of 384 (128 + 256). This final fused vector, $f_{\text{fused}} = [f_{\text{mov}}; f'_{\text{cos}}]$, contains a rich, context-aware representation of the performance.

The fused feature vector, f_{fused}, of 384 dimensions is then passed through a two-layer MLP regression head. The first layer has 256 neurons, and the second has 128, both using the ReLU activation function. The final output layer consists of three neurons, corresponding to the three aesthetic scores, with no activation function.

D. AFFECTIVE PREDICTION AND MODEL TRAINING

The fused feature vector f_{fused} is passed through a multilayer perceptron (MLP), which acts as the regression head. The MLP has three output neurons, corresponding to the predicted scores for Gracefulness, Power, and Fluidity.

The model was trained end-to-end using the mean squared error (MSE) loss function, which measures the average squared difference between the predicted scores and the ground truth human ratings. We used the Adam optimizer with a learning rate of 1×10^{-4}. The dataset was split into training (70%), validation (15%), and testing (15%) sets.

V. RESULTS

To validate the effectiveness of our proposed DC-AEM model and to investigate the specific contribution of the costume modality, we conducted a series of experiments. We evaluated the model's performance using two standard metrics for regression tasks: the Pearson correlation coefficient (PCC), which measures the linear relationship between predicted and actual scores, and the mean absolute error (MAE), which measures the average magnitude of the errors. Specifically, evaluations were performed at the video level (one predicted score per clip) on the test set. For the overall performance metrics (Table 1), the reported PCC and MAE represent the average values calculated across the three aesthetic dimensions.

A. ABLATION STUDY

The central hypothesis of this work is that the integration of costume features is crucial for accurate affective prediction. To test this, we performed an ablation study by comparing the performance of our full multimodal model (DC-AEM) with two single-modality baselines:

- Movement-Only Model: This model uses only the GCN-LSTM stream (f_{mov}) for prediction.
- Costume-Only Model: This model uses only the CNN-based costume feature stream (f_{cos}) for prediction.
- Simple Fusion Model (No Attention): This model concatenates the movement feature vector (f_{mov}) and the costume feature vector (f_{cos}) directly. The resulting combined vector is then fed into the final regression head for prediction. This baseline is designed to assess the performance gain from multimodality without the benefit of the attention mechanism.

The results, averaged across the three affective dimensions, are presented in Table 1.

TABLE 1. Performance Comparison of Different Model Configurations

Model Configuration	Pearson Correlation Coefficient (PCC)	Mean Absolute Error (MAE)
Costume-Only	0.48	1.15
Movement-Only	0.76	0.62
Simple Fusion (No Attention)	0.83	0.49
DC-AEM (Full Model)	0.89	0.38

As shown in Table 1, the single-modality models establish baseline performances. The Movement-Only model performs reasonably well (PCC of 0.76), confirming that kinematics are the primary carrier of aesthetic information, while the Costume-Only model performs significantly worse. This result is expected and highlights the core hypothesis of our work: the aesthetic contribution of the costume is not independent but deeply contextual. The dynamic features of a garment, such as the pattern of its flow or swirl, are largely meaningless without the specific bodily movements that generate them; therefore, their predictive power in isolation is inherently limited. Crucially, the Simple Fusion (No Attention) model shows a notable improvement over the Movement-Only model, achieving a PCC of 0.83. This demonstrates that simply combining costume and movement features provides significant complementary information.

However, our full DC-AEM model, which employs the cross-modal attention mechanism, achieves the best performance by a clear margin (PCC of 0.89 and MAE of 0.38). This final step in performance gain—from 0.83 to 0.89 in PCC— isolates the contribution of the attention mechanism, strongly supporting our hypothesis that intelligently weighting the features based on context is critical for achieving the most accurate predictions.

To further confirm the critical role of dynamic costume features, we conducted an additional ablation study.

We built a variant of the full DC-AEM model that used only the static costume features (color, texture, and shape) and omitted the dynamic optical flow features. The performance of this Static Costume-Only variant was a PCC of 0.81, notably lower than the full DC-AEM model's 0.89. This result provides direct evidence that the dynamic properties of the costume are not only complementary but essential for accurate aesthetic prediction.

B. PERFORMANCE ON INDIVIDUAL AFFECTIVE DIMENSIONS

We further analyzed the performance of our DC-AEM model on each of the three aesthetic dimensions separately. The results are detailed in Table 2.

TABLE 2. DC-AEM Performance on Individual Affective Dimensions

Affective Dimension	Pearson Correlation Coefficient (PCC)	Mean Absolute Error (MAE)
Gracefulness	0.91	0.35
Power	0.85	0.43
Fluidity	0.92	0.36

The model demonstrates excellent performance across all dimensions, with the highest correlations achieved for Fluidity (0.92) and Gracefulness (0.91). This suggests that the model is particularly adept at capturing the interplay between smooth, elegant movements and the dynamic behavior of the costume. For instance, the flowing motion of a romantic tutu in a slow turn, captured by our dynamic costume features, likely contributes significantly to the high Fluidity score. The performance on Power is also strong (PCC = 0.85), though slightly lower. This may be because Power is often conveyed more through sharp, athletic kinematics, with the costume playing a more secondary, albeit still important, role (e.g., the clean lines of a simple leotard accentuating the dancer's musculature). These results highlight the model's ability to learn nuanced relationships between multimodal features and high-level aesthetic concepts.

VI. DISCUSSION

The empirical results presented in this study provide compelling evidence for the efficacy of a multimodal approach to the computational analysis of dance aesthetics. The significant performance gain of the full DC- AEM model over the Movement-Only baseline, especially when contrasted with the poor performance of the Costume-Only model, demonstrates that the costume's features are not powerful in isolation but are critically synergistic. This finding suggests that the costume is not a passive decoration but an active, information-rich component whose aesthetic relevance is unlocked and defined by the context of the dancer's movements.

Our model's success can be attributed to its ability to learn the complex, synergistic relationship between how a dancer moves and what the dancer wears. The cross-modal attention mechanism, in particular, allows the model to infer, for example, that the billowing of a skirt is highly relevant when assessing the Fluidity of a spin but less so during a static pose. This learned context-awareness is a critical step toward building computational systems that can interpret artistic performance with human-like nuance.

The findings have profound implications for the field of textile and costume design. The feature extraction pipeline developed for the costume modality provides a quantitative framework for characterizing not just the static properties of a garment (color, texture, shape) but also its dynamic behavior. By correlating these quantitative features with predicted aesthetic outcomes, our model can function as a powerful analytical tool for designers. For instance, a designer could computationally explore how changing a fabric's weight (which would alter its dynamic flow features) might impact the perceived Gracefulness of a specific choreographic sequence. This research provides a foundational step toward a new paradigm of intelligent costume design, where creative intuition is augmented by data-driven insights, enabling the creation of garments that are optimally tuned to the expressive goals of a performance. However, we acknowledge that the generalizability of these findings is currently limited to the specific context of solo ballet performances, and further validation is required for other dance genres. This moves beyond traditional textile testing, which focuses on physical properties like tensile strength or drape coefficient, to computationally modeling the affective properties of textiles in their intended application context. Furthermore, this research contributes significantly to the domain of aesthetic education. One of the greatest challenges in teaching art appreciation is the subjective nature of aesthetic judgment. Our model provides a potential pathway toward creating more objective and personalized educational tools. An application built upon the DC-AEM could provide students with real-time feedback, highlighting specific moments in a performance where the synergy between movement and costume is particularly effective in conveying a certain emotion or aesthetic quality. It could deconstruct a performance by visualizing the model's attention weights, showing a student which specific bodily actions and costume elements are contributing to a feeling of "Power" or "Gracefulness." This could help provide new computational insights into the artistic process and thus potentially cultivate a deeper, more analytical appreciation for dance and design.

Despite the promising results, this study has certain limitations that suggest avenues for future work. Our dataset, while carefully curated, is limited to solo ballet performances. Future research should expand the dataset to include a wider variety of dance genres (e.g., modern, folk, hip-hop), as well as ensemble performances, to test the model's generalizability. The costume features, while effective, could be further enriched by incorporating physics-based material properties (e.g., stiffness, weight) if such data were available, creating an even tighter link between textile science and affective computing. Finally, the defined aesthetic dimensions of Gracefulness, Power, and Fluidity represent only a subset of the possible human responses to dance. Future work could explore a more extensive affective space, including discrete emotions (e.g., joy, sadness) and more complex concepts (e.g., "narrative intensity").

VII. CONCLUSION

In this paper, we addressed the challenge of computationally modeling the holistic aesthetic experience of dance by proposing a novel multimodal deep learning framework, the DC-AEM. Our work establishes a new approach that, for the first time, concurrently analyzes the dancer's kinematic movement and the visual- semantic features of their costume. By processing these two modalities in parallel and integrating them through a sophisticated cross-modal attention mechanism, our model successfully learned the intricate interplay that governs human aesthetic perception. The experimental results demonstrated that this multimodal approach is not only viable but significantly superior to methods that rely on movement analysis alone. The findings confirm that the costume is a critical information channel, providing essential context that refines and enhances the interpretation of the dancer's movements. This research provides a foundational methodology for the objective and quantitative assessment of solo ballet performance and offers a preliminary analytical tool for intelligent costume design within this specific context. This framework provides a solid basis for future research to expand upon and test for generalizability across a broader range of performance styles. It also lays the groundwork for future applications in digital aesthetic education, which will require expansion and validation across more diverse dance genres. By bridging the gap between artistic expression and computational intelligence, this work contributes to a deeper and more nuanced computational understanding of dance as a performing art, while clearly demonstrating the need for future research to test the generalizability of this approach across the broad spectrum of performance styles.

AUTHOR CONTRIBUTIONS

Shuanghe Li and Yanlin Deng designed the study; all authors conducted the study; Shuanghe Li and Yanlin Deng collected and analyzed the data. Yanlin Deng and Shuanghe Li participated in drafting the manuscript, and all authors contributed to critical revision of the manuscript for important intellectual content. All authors gave final approval of the version to be published. All authors participated fully in the work, took public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or completeness of any part of the work were appropriately investigated and resolved.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

AVAILABILITY OF DATA AND MATERIALS

The datasets used and/or analysed during the current study were available from the corresponding author on reasonable request.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This survey was conducted in compliance with Ethics Committee of Beijing College of Finance and Commerce.

Participants were informed of the study's purpose and data usage prior to participation, and responses were collected anonymously. No personally identifiable information was stored.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] C. Hou, "The Fusion of Modern Dance and Traditional Dance: Balancing Innovation and Heritage," Journal of Education, Humanities, and Social Research, vol. 2, no. 1, pp. 137-144, 2025, doi: 10.71222/skerj284.

[2] R. Huang, L. Zhang, and Y. Li, "Transforming dance education in China: Enhancing sustainable development and cultural preservation," Research in Dance Education, pp. 1-29, 2025, doi: 10.1080/14647893.2025.2524151.

[3] S. Banes, "Terpsichore in Sneakers: Post-Modern Dance," Middletown, CT, USA: Wesleyan University Press; 1987.

[4] A. Carter and J. O'Shea, "The Routledge Dance Studies Reader Second Edition," London, UK: Routledge; 2010.

[5] D. Barbieri, "Costume in Performance: Materiality, Culture, and the Body," London, UK: Bloomsbury Publishing; 2017.

[6] J. Imparato, "Relations between body and clothing in performance: Costume as an activator of bodily actions," Studies in Costume & Performance, vol. 6, no. 2, pp. 171-184, 2021, doi: 10.1386/scp_00045_1.

[7] B. Faber, "Color Psychology and Color Therapy," New York, NY, USA: McGraw-Hill; 1950.

[8] A. Kraut, "Choreographing Copyright: Race, Gender, and Intellectual Property Rights in American Dance," New York, NY, USA: Oxford University Press; 2015.

[9] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," Pattern Recognition Letters, vol. 119, pp. 3-11, 2019, doi: 10.1016/j.patrec.2018.02.010.

[10] R. W. Picard, "Affective Computing," Cambridge, MA, USA: MIT Press; 2000.

[11] I. Rallis, A. Voulodimos, N. Bakalos, E. Protopapadakis, N. Doulamis, and A. Doulamis, "Machine learning for intangible cultural heritage: A review of techniques on dance analysis," Visual Computing for Cultural Heritage, pp. 103-119, 2020, doi: 10.1007/978-3-030-37191-3_6.

[12] M. Joshi and S. Chakrabarty, "An extensive review of computational dance automation techniques and applications," Proceedings of the Royal Society A, vol. 477, no. 2251, Art. no. 20210071, 2021, doi: 10.1098/rspa.2021.0071.

[13] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, et al., "Deep learning-based human pose estimation: A survey," ACM Computing Surveys, vol. 56, no. 1, pp. 1-37, 2023, doi: 10.1145/3603618.
[14] L. V. Wang and Wu H-i, "Biomedical Optics: Principles and Imaging," Hoboken, NJ, USA: John Wiley & Sons; 2007.
[15] S. Yan, Y. Xiong, D. Lin, and editors, "Spatial temporal graph convolutional networks for skeleton-based action recognition," Proceedings of the AAAI Conference on Artificial Intelligence; 2–7 February 2018; New Orleans, LA, USA. Palo Alto, CA, USA: AAAI Press; 2018, doi: 10.1609/aaai.v32i1.12328.
[16] S. Afzal, H. A. Khan, M. J. Piran, and J. W. Lee, "A comprehensive survey on affective computing: Challenges, trends, applications, and future directions," IEEE Access, vol. 12, pp. 96150-96168, 2024, doi: 10.1109/ACCESS.2024.3422480.
[17] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," IEEE Signal Processing Magazine, vol. 34, no. 4, pp. 80-106, 2017, doi: 10.1109/MSP.2017.2696576.
[18] R. Datta, D. Joshi, J. Li, J. Z. Wang, and editors, "Studying aesthetics in photographic images using a computational approach," European Conference on Computer Vision. Berlin, Heidelberg: Springer; 2006, doi: 10.1007/11744078_23.
[19] X. Lu, Z. Lin, H. Jin, J. Yang, J. Z. Wang, and editors, "Rapid: Rating pictorial aesthetics using deep learning," Proceedings of the 22nd ACM International Conference on Multimedia; 3-7 November 2014; Orlando, FL, USA. New York, NY, USA: ACM; 2014, doi: 10.1145/2647868.2654927.
[20] L. Mai, H. Jin, F. Liu, and editors, "Composition-preserving deep photo aesthetics assessment," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 27-30 June 2016; Las Vegas, NV, USA. New York, NY, USA: IEEE; 2016, [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/papers/Mai_Composition-Preserving_Deep_Photo_CVPR_2016_paper.pdf.
[21] S. Fan, B. L. Koenig, Q. Zhao, and M. S. Kankanhalli, "A deeper look at human visual perception of images," SN Computer Science, vol. 1, no. 1, pp. 58, 2020, doi: 10.1007/s42979-019-0061-5.
[22] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, T. L. Berg, and editors, "Where to buy it: Matching street clothing photos in online shops," Proceedings of the IEEE International Conference on Computer Vision; 7-13 December 2015; Santiago, Chile. New York, NY, USA: IEEE; 2015, doi: 10.1109/ICCV.2015.382.
[23] X. Han, Z. Wu, Z. Wu, R. Yu, and L. Davis, "Viton: An image-based virtual try-on network," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 18-23 June 2018; Salt Lake City, UT, USA. New York, NY, USA: IEEE; 2018, doi: 10.1109/CVPR.2018.00787.
[24] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, and editors, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 27-30 June 2016; Las Vegas, NV, USA. New York, NY, USA: IEEE; 2016, doi: 10.1109/CVPR.2016.124.
[25] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, no. 6, pp. 610-621, 1973, doi: 10.1109/TSMC.1973.4309314.
[26] K.-J. Choi and H.-S. Ko, "Stable but responsive cloth," ACM SIGGRAPH 2005 Courses. New York, NY, USA: Association for Computing Machinery; 2005. p. 1-es, doi: 10.1145/1198555.1198571.
[27] D. Baraff and A. Witkin, "Large steps in cloth simulation," Seminal Graphics Papers: Pushing the Boundaries. New York, NY, USA: Association for Computing Machinery; 2023. Volume 2, p. 767-778, doi: 10.1145/3596711.3596792.
[28] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, 2018, doi: 10.1109/TPAMI.2018.2798607.
[29] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," IEEE Access, vol. 7, pp. 63373-63394, 2019, doi: 10.1109/ACCESS.2019.2916887.
[30] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, editors, et al., "Show, attend and tell: Neural image caption generation with visual attention," Proceedings of the 32nd International Conference on Machine Learning, PMLR; 6-11 July 2015; Lille, France, 2015, doi: 10.48550/arXiv.1502.03044.
[31] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 12, no. 5, pp. 1-32, 2021, doi: 10.1145/3465055.