Multi-Modal Correlation Modeling and Aesthetic Education Effectiveness Evaluation of Visual and Action Semantics in Ethnic Dance Costumes

Lei Chen, Meng Zheng
General Education Center, Guangdong University of Technology, Guangzhou 510006, Guangdong, China
Corresponding author: Meng Zheng (e-mail: LY826567756@163.com)

ABSTRACT Establishing reliable associations between visual appearance and dynamic behavior is essential for multimodal perception systems that require semantic consistency across heterogeneous data sources. This study proposes a cultural semantic-driven framework for modeling the relationship between ethnic dance costumes and movement expression through cross-modal representation learning. Clothing images are encoded using CLIP to extract texture, color, and structural features, while spatiotemporal characteristics of dance movements are captured by ST-GCN from threedimensional skeletal sequences. Cultural keywords are introduced as semantic anchors, and a cross-modal InfoNCE objective aligns visual and motion embeddings within a shared feature space under weak supervision. To improve interpretability, cross-modal attention and Grad-CAM are employed to generate association heatmaps that reveal the correspondence between costume elements and movement semantics. Experimental evaluations on representative ethnic dance datasets demonstrate superior semantic consistency, lower semantic alignment error, robust interpretability under increasing motion complexity, and improved agreement with human cognitive judgments compared with mainstream multimodal models. Educational intervention experiments further confirm significant gains in cultural knowledge acquisition and aesthetic judgment stability. By integrating visual feature extraction, semantic alignment, and interpretable multimodal reasoning into a unified architecture, the proposed framework provides a transferable computational strategy for image–motion information fusion and perception-oriented intelligent systems, offering methodological insights for multidisciplinary engineering applications involving multimodal sensing, feature transmission, and semantic association analysis.

INDEX TERMS cross-modal learning, semantic alignment, CLIP, ST-GCN, multimodal perception

I. INTRODUCTION

ETHNIC dance is an important carrier of multi-ethnic cultural expression, and its costumes and movements form a highly coordinated symbol system [1,2]. Clothing serves as a visual representation of cultural identity, utilizing the material properties of textile fibers and intricate weaving techniques to create a structural metaphor for the rhythm, emotions, and ritual meanings of human movement [3]; And dance movements activate the dynamic semantics of clothing through body language, jointly completing the transmission of cultural information [4-6]. The deep coupling relationship between vision and action constitutes the core material of ethnic aesthetic education practice [7,8]. Thoroughly analyzing the semantic correlation between dance movements and the structural logic of traditional textiles contributes to the digital preservation of intangible cultural heritage while providing theoretical support for building an intelligent aesthetic education system rooted in textile-based cultural sensitivity.

There is a clear disconnect in current research on multimodal analysis of dance: visual modalities focus on image classification or style transfer, while action modalities are mostly used for behavior recognition or action synthesis, with very little alignment between the two at the cultural semantic level [9,10]. On the one hand, clothing analysis often remains at the level of surface feature extraction, lacking binding with specific action contexts, resulting in cultural symbols being simplified into decorative elements [11,12]; On the other hand, although action modeling can capture spatiotemporal dynamics, it lacks the cultural context constraints provided by clothing, making it difficult to restore the symbolic connotations of actions [13,14]. A deeper issue is that the semantic expression of ethnic dance heavily relies on non explicit cultural knowledge, and

these associations cannot be automatically captured through general visual or motion models [15,16]. Existing cross-modal methods often rely on strictly paired graphic or visual data, and have weak generalization ability in scenarios such as ethnic dance where annotation is scarce, semantic granularity is high, and cultural specificity is strong [17,18]. Even if large-scale pre-trained models are introduced, their general semantic space is difficult to cover the metaphorical systems unique to ethnic cultures [19,20]. In addition, the vast majority of models lack interpretable mechanisms, and the decision-making process is like a black box, unable to indicate which part of the clothing triggers the judgment of specific action semantics, resulting in insufficient credibility in aesthetic education teaching [21,22]. The multiple limitations of modal isolation, shallow semantics, cultural embeddedness, and lack of interpretation make it difficult for technology to touch the essential structure of visual and motor coordination in ethnic dance, seriously restricting the ability of digital aesthetic education to restore and transmit deep cultural logic.

This article aims to establish a fine-grained semantic correlation model between clothing visual and dance movements, and evaluate its effectiveness in aesthetic education scenarios. Different from existing works, this study transfers the image text alignment ability of CLIP (Contrastive Language-Image Pretraining) to the cross-modal task of ethnic clothing action, constructs semantic anchors through cultural keywords, and guides the alignment of visual and action features in the shared embedding space. The innovation is reflected in three aspects: firstly, this article constructs a dual stream feature alignment framework with cultural labels as a bridge. The clothing end uses CLIP to extract visual representations with cultural orientation, and the action end captures the spatiotemporal evolution of joint sequences through ST-GCN (Spatial Temporal-Graph Convolutional Network) and aligns them with the same cultural semantic vector; Secondly, this article designs a cross-modal InfoNCE (Information Noise Contrastive Estimation) loss to achieve weakly supervised fine-grained matching without strict paired data conditions, improving the robustness of the model to sparse labeling; Thirdly, cross-modal attention and Grad-CAM (Gradient-weighted Class Activation Mapping) can be integrated to generate clothing action correlation heat maps, visualizing the activation areas of cultural semantics and providing interpretable basis for aesthetic feedback.

II. CLOTHING ACTION CULTURE SEMANTIC ALIGNMENT MODELING

Figure 1 shows the architecture of the clothing action multimodal semantic association model. The system includes input branches: clothing images are segmented by Mask R-CNN (Mask Region-based Convolutional Neural Network) [23,24] and visual features are extracted through CLIP [25]; The action sequence is modeled spatiotemporal dynamics using ST-GCN [26-28]; Cultural keywords are encoded into semantic vectors through text encoding. The core module achieves fine-grained semantic alignment between clothing and action in a shared embedding space through cross-modal contrastive learning, and innovatively adopts momentum update mechanism [29,30] and InfoNCE loss. The interpretability module generates a heatmap of clothing action associations, intuitively revealing the activation areas of cultural symbols. The final output evaluates the effectiveness of aesthetic education from dimensions such as cognitive accuracy and cultural understanding, forming a complete closed loop from technological implementation to educational application.

A. VISUAL FEATURE EXTRACTION AND SEMANTIC ENCODING OF ETHNIC COSTUMES

For the clothing area in ethnic dance images, Mask R-CNN is first used for instance level segmentation to accurately extract the clothing foreground mask and eliminate interference from irrelevant parts of the background and human body. The segmentation results are cropped and input into the CLIP visual encoder, which is based on the ViT-B/32 (Vision Transformer Base/32) [31,32] architecture and fine tuned on the ImageNet-21K dataset to adapt to the distribution of ethnic cultural visual features. After being processed by the encoder, each clothing image is mapped into a 512 dimensional visual embedding vector, preserving high-order texture, structure, and color semantics.

To establish alignment between visual representation and cultural semantics, a set of clothing culture keywords annotated by experts is constructed, covering dimensions such as ethnic affiliation, ritual attributes, and pattern symbolism. It should be noted that the cultural keywords were jointly determined by three scholars of ethnic dance based on the Chronicle of Chinese Ethnic Minority Dances and field survey data. Each costume sample was associated with only 1–3 keywords most directly related to its movement context (e.g., "peacock spreading its tail—praying for blessings," "drum surface pattern—ancestor worship"), and semantic focus was ensured through expert consistency testing (Kappa ≥ 0.78). While the CLIP text encoder cannot parse the metaphor itself, by anchoring polysemous symbols to specific keywords within the performance context, cross-cultural ambiguity can be effectively avoided. These keywords are used as text input and fed into the CLIP text encoder, which also shares the semantic space based on the ViT-B/32 architecture to generate corresponding semantic vectors. This process ensures comparability between images and text within a unified embedding space, laying the foundation for subsequent cross-modal correlations. To enhance the consistency between visual features and cultural semantics, contrast alignment constraints are adopted during the fine-tuning stage. Given a clothing image and its corresponding keywords, the similarity of its positive sample pairs is measured by cosine similarity:

$$s(v,t) = \frac{v^T t}{\|v\| \cdot \|t\|} \tag{1}$$

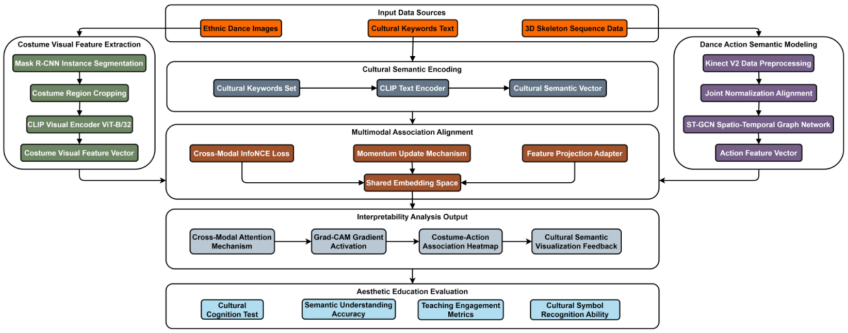

FIGURE 1. Architecture of Clothing Action Multimodal Semantic Association Model

v is the image encoding output, and t is the text encoding output. Maximizing the similarity of positive sample pairs during training while minimizing the similarity between the image and irrelevant keywords can compress semantic ambiguity and enhance the focus of cultural orientation.

The entire process does not introduce additional network structures and relies entirely on the graphic and text alignment capability of CLIP pre-training architecture. Only minor adjustments are made to the projection layer to avoid damaging its original semantic priors. The image is standardized before input, and the text is segmented and template encapsulated before input to fit the original CLIP training format. This strategy effectively alleviates the problem of semantic dilution of ethnic clothing visual symbols in the general model, enabling the extracted visual embeddings to have clear cultural semantic anchoring.

B. SEMANTIC SEQUENCE MODELING AND CULTURAL ANNOTATION ALIGNMENT OF DANCE MOVEMENTS

The raw motion data is collected by the Kinect V2 sensor to obtain a three-dimensional coordinate sequence of the actor's body joints. Each frame contains the spatial positions of 25 joint points, forming a spatiotemporal graph structure: nodes are joints, and edges are predefined based on human skeletal connections. The sequence first undergoes denoising and coordinate normalization, aligning all frames with pelvic joint points as the origin and scaling them to the unit scale to eliminate individual body shape and absolute position differences. It should be specifically noted that the unit scale scaling here only applies to the 3D joint coordinate sequence; the clothing image maintains its original resolution and proportions to preserve the cultural semantic relationship between clothing and limbs.

The processed sequence is input into the ST-GCN network, which aggregates spatiotemporal neighborhood information layer by layer. The spatial dimension transmits inter joint dependencies within each frame through graph convolution, while the temporal dimension models dynamic evolution along frame order through causal convolution. The network is stacked in 9 layers, and each layer is followed by batch normalization and ReLU (Rectified Linear Unit) activation [33-35]. Finally, action embedding vectors are generated through global average pooling, with dimensions aligned with visual modalities.

To establish a mapping between action and cultural semantics, a pre-built action culture label mapping library is constructed. This database is jointly annotated by dance experts based on movement form, rhythm characteristics, and ethnic context, with each type of movement corresponding to one or more cultural keywords. These keywords are processed by CLIP text encoder to generate a unified dimensional cultural semantic vector. Align action embeddings and cultural vectors in a shared semantic space, and achieve semantic binding by minimizing the distance between the two:

$$L_{\text{align}} = 1 - \frac{a^T c}{\|a\| \cdot \|c\|} \tag{2}$$

a is the action embedding output by ST-GCN, and c is the semantic vector obtained by CLIP encoding the corresponding cultural keywords. This loss term guides action features to converge towards cultural semantic anchors during training, enabling the model to not only recognize action forms but also capture their cultural connotations.

The entire process transforms the original 3D joint sequence into semantic representations with cultural orientation, avoiding action modeling from being disconnected from ethnic contexts and providing a structured semantic foundation for cross-modal associations.

C. CROSS-MODAL CONTRASTIVE LEARNING DRIVEN EMBEDDING SPACE ALIGNMENT

The construction of positive and negative sample pairs is based on the time synchronization information of dance video clips. The clothing visual embeddings and corresponding action embeddings extracted within the same segment form positive sample pairs, ensuring consistency between the two in cultural context and performance timing; Negative samples are generated through random pairing across fragments, disrupting the original semantic associations and forming interference terms. All sample pairs enter a unified embedding space defined by the CLIP pre-trained model, where visual and action modalities are mapped to their respective encoders.

To achieve cross-modal alignment, the cross-modal variant of the InfoNCE loss function is used as the optimization

objective. Assuming there are N positive sample pairs in the batch, for the i-th clothing embedding v_i, its corresponding action positive sample is a_i, and the remaining $N-1$ action embeddings form a negative sample set. The similarity is measured by cosine similarity, scaled by temperature coefficient τ and input into softmax. The loss is defined as:

$$L_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s(v_i,a_i)/\tau)}{\sum_{j=1}^{N}\exp(s(v_i,a_j)/\tau)} \tag{3}$$

s represents cosine similarity. This loss driven model brings together semantically consistent clothing action pairs in the embedding space, while pushing away irrelevant pairings to enhance cross-modal semantic consistency.

In terms of training strategy, freeze the CLIP backbone network parameters and only fine tune the linear projection head behind it, adapting the original 512 dimensional output to the alignment space. To maintain semantic stability, a momentum update mechanism is introduced: the CLIP encoder weights are not directly optimized, but slowly synchronized from the teacher model through exponential moving average, with a momentum coefficient of 0.995. This effectively suppresses training oscillations and prevents cultural semantics from deviating from pre-training priors due to end-to-end fine-tuning.

The entire alignment process does not rely on strictly paired fine-grained annotations, and only requires fragment level synchronization to construct weakly supervised signals. By comparing learning mechanisms, an implicit association path between clothing and action is established in the shared semantic space, achieving cross-modal semantic coupling without explicit alignment labels.

D. EXPLANATORY ATTENTION GUIDED FINE-GRAINED ASSOCIATION GENERATION

After completing the multimodal embedding alignment, this paper directly reuses the trained clothing and action encoder outputs and integrates a lightweight cross-modal attention module. This module does not add any learnable parameters and only calculates the correlation weight between clothing region features and action keyframe embeddings through dot product interaction. Specifically, the clothing image is divided into spatial patches corresponding to ViT, with each patch corresponding to a visual feature vector; The action sequence extracts ST-GCN embeddings of several keyframes through time pooling. The two are multiplied pairwise in a 512 dimensional space, and then normalized by softmax to generate a normalized attention weight matrix that reflects the response strength of each clothing region to specific action segments. To map the association back to the original image space, Grad CAM technology is used for visualization. Firstly, select the cultural semantic category as the target output node, and backpropagate the gradient of the last layer feature map of the clothing encoder to this node. Perform global average pooling on the channel dimension to obtain the importance weight α_k of each channel, where k represents the channel index of the feature map. The heatmap H is generated by weighted summation:

$$H = \text{ReLU}\left(\sum_{k=1}^{C}\alpha_k \cdot A_k\right) \tag{4}$$

A_k is the feature map of the kth channel, C is the total number of channels, and ReLU ensures that only the positive contribution region is retained. This heatmap is aligned with the original clothing image through upsampling, highlighting the patterns, contours, or structural regions that play a decisive role in the current action semantics.

The entire process is entirely based on the internal response embedded after alignment, without the need for additional annotation or training. The attention mechanism captures fine-grained patterns of cross-modal interactions, while Grad CAM maps abstract embeddings to pixel level interpretations, forming a traceable path from cultural semantics to visual elements. The output heatmap is directly used for aesthetic feedback, allowing learners to intuitively understand how specific dance movements activate cultural symbols in clothing, enhancing the transparency of cultural cognition and teaching credibility. This design achieves visual decoupling of semantic associations while keeping the model concise.

III. MULTIDIMENSIONAL QUANTITATIVE EVALUATION OF AESTHETIC EDUCATION EFFECTIVENESS

A. EXPERIMENTAL DATA

The experimental data collection covers typical dance types of major ethnic minorities, including Mongolian bowl dance, Tibetan string dance, Uyghur Sainaim, Miao reverse wood drum dance, and Dai peacock dance, totaling 127 high-definition performance videos (1080p/30fps) with a total duration of about 9.6 hours. After manual verification, 2840 high-quality samples were extracted from the clothing images, and the motion data was synchronously captured by Kinect V2 to generate a 3D joint sequence of 38760 frames, with a unified frame rate of 25Hz. All motion sequences are manually verified, severely occluded frames are removed, and median filtering and interpolation are used to repair minor noise to ensure the reliability of the data input to ST-GCN. It should be noted that the selected dance types are all highly stylized representative dances of various ethnic groups, with highly standardized costume-movement combinations and limited intra-class diversity;

moreover, the test set is completely isolated by ethnicity. The cultural keyword system was jointly constructed by three ethnic dance scholars, including 5 categories, 23 subcategories, and 112 fine-grained semantic labels, covering clothing pattern symbolism, action ritual function, and ethnic aesthetic principles. The test set is divided into ethnic categories to ensure that there is no overlap between training and evaluation. Although training and testing share

the same set of cultural keywords (standardized by experts), no text input is provided during the testing phase. The model calculates cross-modal similarity only based on unseen clothing images and action sequences; therefore, high consistency reflects semantic generalization ability, not label memorization. All comparison models are run on the same data slice, and the input is uniformly preprocessed: clothing images are cropped to clothing areas and normalized, action sequences are aligned to the center of the human body, and skeletal scales are standardized. This data provides a reliable foundation for cross-modal alignment and aesthetic education efficacy evaluation.

B. EVALUATION OF CULTURAL SEMANTIC CONSISTENCY

This article calculates the cosine similarity and semantic alignment error (SAE) of clothing action pairs in a shared embedding space. SAE is defined as the mean square error between the embedding distance and the manually annotated semantic level. Several typical ethnic dance groups (Mongolian, Tibetan, Uyghur, Miao, and Dai) were selected, and the proposed method was compared with five mainstream cross-modal models: ViLBERT (Visual-and-Language Bidirectional Encoder Representations from Transformers), UNITER (Universal Image-Text Representation Learning), ALBEF (Align Before Fuse), BLIP (Bootstrapped Language-Image Pretraining), and Flamingo. Evaluation was based on a unified test set to ensure consistent input format.

The horizontal axis in Figure 2 represents the cross-modal methods compared, and the vertical axis represents five typical ethnic dance categories. Figure 2a reflects the semantic affinity of clothing action pairs in the shared space, while Figure 2b measures the deviation between their embedded representations and artificial semantic levels. The proposed method has a significant advantage in consistency, with cosine similarity ranging from 0.82 to 0.86 and SAE ranging from 0.12 to 0.15. This is due to its alignment mechanism guided by cultural keywords and weakly supervised contrastive learning strategy, effectively bridging the semantic gap of general models in ethnic cultural contexts. Although mainstream models such as ViLBERT or BLIP perform well in general graphic and text tasks, their pre-training corpora lack the unique symbol system and action costume coupling logic of ethnic dance, resulting in cultural semantic dilution in the embedding space. This article uses CLIP to anchor cultural terms, ST-GCN to model action context, and explicitly constrains cross-modal distance during training, making the model more discriminative of fine-grained differences in ethnic culture. The systematic reduction of SAE further confirms that when visual and action features are co optimized under the guidance of the same cultural semantic vector, the embedding distance can more accurately reflect the semantic hierarchy annotated by experts, thereby supporting the quantitative judgment of cultural expression accuracy in aesthetic education scenarios.

C. AESTHETIC PERCEPTION INTERPRETABILITY ASSESSMENT

Attention Coverage Ratio (ACR) and Key Region Hit Rate (KRHR) were used to measure whether the model focused on the core cultural region. ACR calculated the proportion of the attention heatmap covering the clothing pattern area, and KRHR counted whether the motion keyframes hit the culturally labeled motion nodes. The model was compared with the five models mentioned above under different dance style complexities (low: repetitive single movements; medium: combined movements; high: complex rhythms and multi-limb coordination).

The horizontal axis of Figure 3 represents different cross-modal methods, and the vertical axis corresponds to two types of interpretability indicators: Figure 3a is ACR, and Figure 3b is KRHR, which sequentially reflect the model performance under low, medium, and high complexity dance movements. The proposed method is significantly superior in all types of complexity, with slow performance degradation. Under high complexity dance style conditions, ACR and KRHR are 0.87 ± 0.04 and 0.88 ± 0.04, respectively. This advantage stems from its alignment mechanism guided by cultural semantics and non parametric attention design. In a single repetitive action, the model can accurately focus on the core pattern of clothing; With the increase of action combinations and the enhancement of limb coordination, the general cross-modal model lacks ethnic and cultural context constraints, and attention is easily dispersed in irrelevant areas, resulting in action activation deviating from cultural annotation nodes. The proposed method anchors cultural keywords through CLIP and employs cross-modal contrastive learning to constrain visual and motion features within a fine-grained embedding space constrained by cultural semantics. Even in high complexity scenarios, it maintains high sensitivity and structural consistency to cultural symbols, ensuring interpretability does not drastically decrease with the dynamic complexity of actions. This indicates that the embedding of cultural priors is not only the key to performance improvement, but also the fundamental source of interpretability robustness.

D. USER COGNITIVE FEEDBACK CONSISTENCY EVALUATION

This article conducts controlled user experiments to collect participants' ratings (1-5 points) for clothing action matching, and calculates the Spearman rank correlation coefficient (ρ) and root mean square error (RMSE) between the model's predicted similarity and user ratings. The experiment was divided into three groups based on users' knowledge level of ethnic dance (no foundation, enthusiasts, professional learners), and a double-blind test was conducted on a fixed video set to compare the consistency between the model output and human judgment.

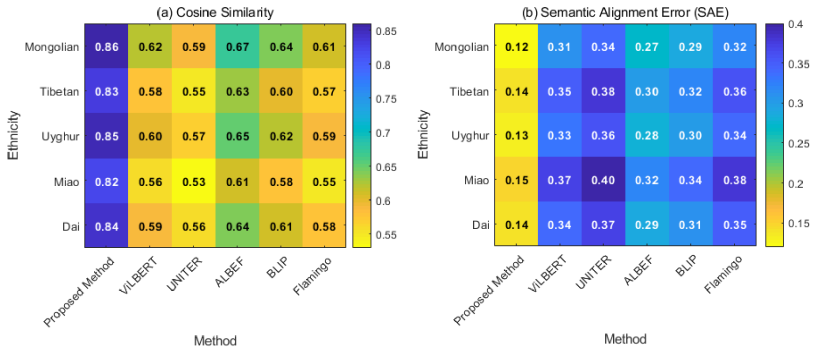

FIGURE 2. Cultural Semantic Consistency Evaluation of Cross Ethnic Dance Categories

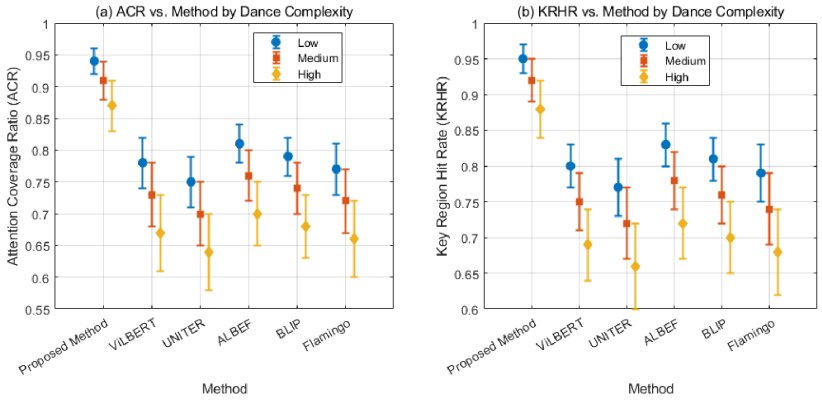

FIGURE 3. Aesthetic Interpretability Evaluation of Dance Complexity

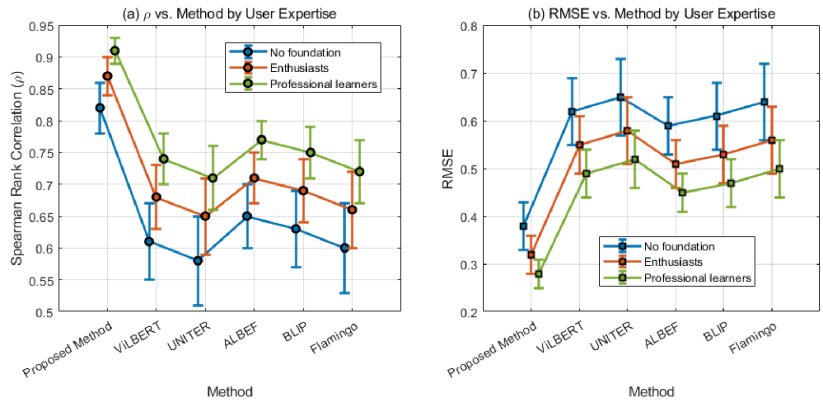

FIGURE 4. Consistency Evaluation of User Cognitive Feedback Across Professional Levels

Figure 4 depicts the performance of different cross-modal models in terms of user cognitive feedback consistency: the horizontal axis represents six models including the method proposed in this paper, and the vertical axis in Figure 4a represents the Spearman rank correlation coefficient (ρ), reflecting the ranking consistency between model predictions and human ratings; The vertical axis in Figure 4b represents the root mean square error (RMSE), which measures the absolute deviation between predicted values and user ratings. The three curves with error lines correspond to three types of user groups, distinguished by their knowledge level of ethnic dance. This method achieves optimal alignment performance in all groups, with a Spearman rank correlation coefficient mean of no less than 0.82 and a root mean square error mean of no more than 0.38, and the advantage expands with the improvement of user professionalism. This phenomenon arises from the model explicitly aligning cultural semantics in the shared embedding space, rather than relying on surface statistical associations. Professional learners are more sensitive to the metaphorical logic of clothing action, and general models lack cultural semantic anchoring, which can lead to their predictions deviating from deep cognitive structures; The keyword guided alignment and interpretable heatmap mechanism introduced in this article make the

model decision logic more closely aligned with the path of human cultural understanding. Although the criteria for enthusiasts and non basic users are relatively rough, they can still perceive semantic coordination. The robust performance of the model in this group confirms that it not only adapts to expert knowledge systems, but also has universal explanatory power for mass aesthetic education. The error range narrows with increasing specialization, further indicating that cultural semantic modeling effectively reduces cognitive noise and improves the consistent boundary of human-machine judgment.

E. EFFECTIVENESS EVALUATION OF AESTHETIC EDUCATION INTERVENTION

To verify whether the model truly supports the educational goals of digital aesthetics, this paper designed a pre- and post-test educational intervention experiment to directly assess learners' improvements in cultural knowledge acquisition and the stability of aesthetic judgment. To measure the improvement in cultural knowledge mastery (ΔKnowledge) and aesthetic judgment stability (ΔStability) of users after viewing model generated associated content. ΔKnowledge calculates the score difference of standardized cultural knowledge test questions, while ΔStability measures the variance change of scores in repeated aesthetic tasks. This article conducts parallel control experiments with content generated by five comparison models.

TABLE 1. Effectiveness of Aesthetic Education Intervention

Method	ΔKnowledge (mean $\pm$ standard deviation)	ΔStability (mean $\pm$ standard deviation)	p-value (vs Proposed Method)
Proposed Method	23.6 $\pm$ 3.1	0.41 $\pm$ 0.05	—
ViLBERT	14.2 $\pm$ 3.8	0.27 $\pm$ 0.07	1.32e-05
UNITER	13.5 $\pm$ 4.0	0.25 $\pm$ 0.08	7.84e-06
ALBEF	15.8 $\pm$ 3.6	0.29 $\pm$ 0.06	2.17e-04
BLIP	15.1 $\pm$ 3.7	0.28 $\pm$ 0.07	1.05e-04
Flamingo	12.9 $\pm$ 4.1	0.24 $\pm$ 0.08	4.91e-06

Table 1 presents the effectiveness differences of different cross-modal models in aesthetic education intervention, focusing on two core indicators: cultural knowledge mastery and aesthetic judgment stability. The method presented in this article performs well, with a ΔKnowledge score of (23.6 $\pm$ 3.1) and a ΔStability score of (0.41 $\pm$ 0.05). All other models are significantly inferior, and statistical tests confirm that the differences are highly significant. The data shows that not all models with image text alignment capabilities can effectively support the goals of ethnic cultural education. The reason is that although the universal crossmodal architecture can capture surface semantic associations, it lacks a modeling mechanism for the deep symbolic system of clothing and movements in ethnic dance. This article uses cultural keyword anchoring, weakly supervised fine-grained alignment, and interpretable heatmap feedback to enable learners to not only recognize "what" but also understand "why" - such as how specific patterns respond to the rhythmic tension or ritual function of actions. The association generated by the embedding of this cultural context strengthens the structural consistency between cognition and aesthetic judgment. In contrast, the contrastive model relies on a universal semantic space, which can easily lead to semantic drift on ethnic specific symbols, resulting in shallow knowledge transmission and a lack of cultural basis for aesthetic feedback, thereby weakening the learning transfer effect. Therefore, the effectiveness of aesthetic education depends not only on the accuracy of modal alignment, but also on the explicit modeling and interpretable guidance of cultural logic, which is the key to surpassing mainstream architectures in this method.

IV. CONCLUSION

This study breaks through the bottleneck of visual and motor modality separation in the digital aesthetic education of ethnic dance by integrating the structural logic of textile crafts into a cultural semantic-driven multimodal correlation modeling framework. Unlike existing methods that rely on universal alignment targets, this paper integrates CLIP's semantic priors with ethnic knowledge systems, using cultural keywords as anchor points to guide clothing and action to achieve fine-grained coupling in a shared embedding space. Crossmodal contrastive learning enhances semantic consistency under weakly supervised conditions without the need for dense labeling. It is particularly crucial that through the collaboration of non parametric attention and Grad CAM, the model not only establishes associations, but also traces the activation path of cultural symbols, making "why does clothing respond to this action" a visible, solvable, and trainable question. User experiments and multidimensional evaluations jointly verify that only by embedding technology alignment into cultural logic can aesthetic education intervention go beyond surface matching and touch upon the core of meaning generation. This path provides a technical reference combining the structural depth of textile art with educational transparency for the intelligent inheritance of intangible cultural heritage dance, offering a methodological framework for applying cross-modal learning to culturally sensitive scenarios where material craft and motion intersect.

AUTHOR CONTRIBUTIONS

Conceptualization –Chen Lei and Zheng Meng; methodology – Chen Lei and Zheng Meng; investigation – Chen Lei and Zheng Meng; writing-original draft preparation – Chen Lei and Zheng Meng. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] C. Pan and F. Alizadeh, "Costume and Body Language of Ethnic Dance Drama under Semiotics: An Analytical Study," Journal of Ecohumanism, vol. 3, no. 7, pp. 741-754, 2024, doi: 10.62754/joe.v3i7.4242.

[2] P. Patel-Grosz, G. Grosz P, T. Kelkar, et al., "Steps towards a semantics of dance," Journal of Semantics, vol. 39, no. 4, pp. 693-748, 2022, doi: 10.1093/jos/ffac009.

[3] T. Utoh-Ezeajugh and A. Ume J, "Dance Costumes as Expressions of Cultural Identity: A Study of Selected Cultural Dances," Frontiers in Art and Design, vol. 1, no. 1, pp. 34, 2025, doi: 10.30560/fad.v1n1p34.

[4] J. Burelle, "All That Moves Us: The Semantic Density of Clothing and Objects in El Buen Vestir-Tlakentli's Choreographies of Indigenous Movements," Theatre Research in Canada, vol. 45, no. 1, pp. 30-55, 2024, doi: 10.3138/tric-2023-0011.

[5] M. Moghanipour, B. Shamshiri, A. Rahmani, et al., "Memorable Dances from the Days of War: A Study on the Form and Meaning of Kurdish Dances in North-Eastern Iran," Folklore, vol. 135, no. 2, pp. 201-228, 2024, doi: 10.1080/0015587X.2024.2325253.

[6] J. LIANG, M. VASINAROM, and D. A. N. C. E. THE DEVELOPMENT OF DAUR, "Procedia of Multidisciplinary Research," 2025; 3(5): 90-90.

[7] M. Zhanguzhinova, "The phenomenon of creative images of Dimash Qudaibergen in the context of sustainable development of the cultural values of Kazakhstan," Creativity Studies, vol. 18, no. 1, pp. 319-338, 2025, doi: 10.3846/cs.2025.20366.

[8] E. Monroy, T. Imada, N. Sagiv, et al., "Dance across cultures: Joint action aesthetics in Japan and the UK," Empirical Studies of the Arts, vol. 40, no. 2, pp. 209-227, 2022, doi: 10.1177/02762374211001800.

[9] S. Jahbrob, "Ite Kakang Aring: Symbolism and Kinship in the Traditional Lego-Lego Dance," Artistic Studies, vol. 1, no. 1, pp. 1-7, 2025, doi: 10.58920/art0101406.

[10] S. Qazvini P, "Sufi Sema and Hindu Kathak Dance: The Rotations' Semantic Layers through History: Sufi Sema ve Hindu Kathak Dansı: Tarih Boyunca Rotasyonların Semantik Katmanları," Near East University Journal of Scientific Mysticism and Literature, vol. 1, no. 2, pp. 35-49, 2025, doi: 10.32955/neujsml202512975.

[11] K. Wygnaniec, "Dance Floors of Polish Traditional Dances: Sensory Anthropology as a Research Tool in Dance Studies," Ethnologia Fennica, vol. 51, no. 2, pp. 112-140, 2024, doi: 10.23991/ef.v51i2.141697.

[12] S. Akbarian, "An analytical study of ethnic identity components in the works of contemporary Kurdish painters in Iranian Kurdistan from 1981-2019," Middle eastern studies, vol. 59, no. 6, pp. 1008-1027, 2023, doi: 10.1080/00263206.2023.2164925.

[13] Z. Shaygozova and L. Nehviadovich, "Clothes as a marker of otherness in traditional Kazakh culture," Bulletin of LN Gumilyov Eurasian National University. Historical Sciences. Philosophy. Religious Studies, vol. 148, no. 3, pp. 135-156, 2024, doi: 10.32523/2616-7255-2024-148-3-135-156.

[14] A. Chiselev, "The Role of "Folk" Costume in the Sustainable Development of Ethnic Communities from Tulcea County, Romania," Case Study: Ukrainians and Russian Lipovans. Culture. Society. Economy. Politics, vol. 2, no. 1, pp. 60-83, 2022, doi: 10.2478/csep-2022-0006.

[15] A. Maiorani and C. Liu, "The functional grammar of dance applied to ELAN annotation: meaning beyond the naked eye," Journal of World Languages, vol. 10, no. 1, pp. 221-249, 2024, doi: 10.1515/jwl-2023-0050.

[16] C. Lin and C. Liu, "Intercultural aesthetics in traditional Chinese dance performance," International Review of the Aesthetics and Sociology of Music, vol. 54, no. 1, pp. 129-146, 2023, doi: 10.21857/mwo1vc37wy.

[17] M. Dangaura, "The memory of performance: from contents to contexts of selected Tharu folk dances," SCHOLARS: Journal of Arts & Humanities, vol. 4, no. 1, pp. 11-28, 2022, doi: 10.3126/sjah.v4i1.43050.

[18] K. Liu, H. Wu, Y. Gao, et al., "Archaeology and virtual simulation restoration of costumes in the Han Xizai banquet painting," Autex Research Journal, vol. 23, no. 2, pp. 238-252, 2023, doi: 10.2478/aut-2022-0001.

[19] K. Lonyangapuo M, M. Obuchi S, S. Nganga, et al., "Reinvention of Lyre Music and Dance for Knowledge Preservation and Political Mitigation: The Case of The Bukusu of Kenya," Mwanga wa Lugha, vol. 7, no. 2, pp. 67-84, 2022.

[20] N. Nirwan and M. Fauzi, "Symbolic Meaning of Kecak Dance Performance in Balinese Culture," JELL (Journal of English Language and Literature) STIBA-IEC Jakarta, vol. 9, no. 02, pp. 389-396, 2024, doi: 10.37110/jell.v9i02.254.

[21] R. Cisneros and V. Lingham, "Costume as bridge: The hoodie's potential to connect audiences with Gypsy, Roma and Traveller communities," Studies in Costume & Performance, vol. 9, no. 1, pp. 9-27, 2024, doi: 10.1386/scp_00106_1.

[22] J. Napoli D, "Stimuli for initiation: a comparison of dance and (sign) language," Journal of Cultural Cognitive Science, vol. 6, no. 3, pp. 287-303, 2022, doi: 10.1007/s41809-022-00095-y.

[23] T. Li, X. Lyu Y, L. Ma, et al., "Research on garment flat multi-component recognition based on Mask R-CNN," Industria Textila, vol. 74, no. 1, pp. 49-56, 2023, doi: 10.35530/IT.074.01.202199.

[24] X. Bi, J. Hu, B. Xiao, et al., "Iemask r-cnn: Information-enhanced mask r-cnn," IEEE Transactions on Big Data, vol. 9, no. 2, pp. 688-700, 2022, doi: 10.1109/TBDATA.2022.3187413.

[25] D. Song, X. Zhang, J. Zhou, et al., "Image-based virtual try-on: A survey," International Journal of Computer Vision, vol. 133, no. 5, pp. 2692-2720, 2025, doi: 10.1007/s11263-024-02305-2.

[26] M. Lovanshi and V. Tiwari, "Human skeleton pose and spatio-temporal feature-based activity recognition using ST-GCN," Multimedia Tools and Applications, vol. 83, no. 5, pp. 12705-12730, 2024, doi: 10.1007/s11042-023-16001-9.

[27] W. Wu, F. Tu, M. Niu, et al., "STAR: An STGCN architecture for skeleton-based human action recognition," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 70, no. 6, pp. 2370-2383, 2023, doi: 10.1109/TCSI.2023.3254610.

[28] L. Benhamida and S. Larabi, "Human action recognition using ST-GCNs for blind accessible theatre performances," Signal, Image and Video Processing, vol. 18, no. 12, pp. 8829-8845, 2024, doi: 10.1007/s11760-024-03510-9.

[29] X. Xie, P. Zhou, H. Li, et al., "Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 12, pp. 9508-9520, 2024, doi: 10.1109/TPAMI.2024.3423382.

[30] Y. Higuchi, N. Moritz, J. Le Roux, et al., "Momentum pseudo-labeling: Semi-supervised asr with continuously improving pseudo-labels," IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1424-1438, 2022, doi: 10.1109/JSTSP.2022.3195367.

[31] T. Yao, Y. Li, Y. Pan, et al., "Hiri-vit: Scaling vision transformer with high resolution inputs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 9, pp. 6431-6442, 2024, doi: 10.1109/TPAMI.2024.3379457.

[32] A. Khan, Z. Rauf, A. Sohail, et al., "A survey of the vision transformers and their CNN-transformer based variants," Artificial Intelligence Review, vol. 56, no. Suppl 3, pp. 2917-2970, 2023, doi: 10.1007/s10462-023-10595-0.

[33] S. Yoon Shin, G. Jo, and G. Wang, "A novel method for fashion clothing image classification based on deep learning," Journal of Information and Communication Technology, vol. 22, no. 1, pp. 127-148, 2023, doi: 10.32890/jict2023.22.1.6.

[34] Z. Zhou, M. Liu, W. Deng, et al., "Clothing image classification algorithm based on convolutional neural network and optimized regularized extreme learning machine," Textile Research Journal, vol. 92, no. 23-24, pp. 5106-5124, 2022, doi: 10.1177/00405175221115472.

[35] J. Xiang, R. Pan, and W. Gao, "Clothing recognition based on deep sparse convolutional neural network," International Journal of Clothing Science and Technology, vol. 34, no. 1, pp. 119-133, 2022, doi: 10.1108/IJCST-06-2018-0081.