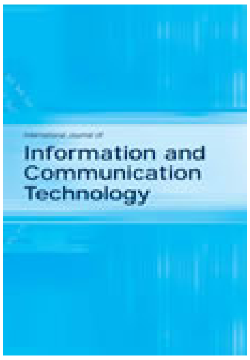
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Animation generation of traditional ethnic elements based on memory-enhanced self-supervised networks

Peilin Wang

DOI: 10.1504/IJICT.2026.10075954

Article History:

Received:	24 August 2025
Last revised:	12 November 2025
Accepted:	13 November 2025
Published online:	06 February 2026

Animation generation of traditional ethnic elements based on memory-enhanced self-supervised networks

Peilin Wang

The School of Art,
Zhengzhou University of Science and Technology,
Zhengzhou, 450064, China
Email: wangyi2025052025@163.com

Abstract: Aiming at the problems of cultural distortion and action homogenisation in traditional ethnic animation generation, this paper proposes a memory-enhanced self-supervised network and AIGC fusion framework. First, a dual-channel memory module is constructed to decouple ethnic visual patterns and semantic features. Second, a culturally constrained mixed density network (MDN) is designed to generate probabilistic, compliant and diverse action sequences based on the learned features to effectively overcome the singularisation problem. Finally, dynamic symbol implantation pipeline is developed to realise high-fidelity and controllable animation synthesis of ethnic elements. Leveraging self-supervised learning on unpaired ethnic images, the framework achieves 89.7% cultural compliance along with enhanced action diversity. Generation efficiency reaches 25-FPS real-time rendering at 0.48-kWh energy consumption, and spatiotemporal synchronisation attains 0.12 s latency. Experiments confirm significant improvements in cultural fidelity, action diversity, and efficiency, establishing a new paradigm for digital preservation of intangible cultural heritage.

Keywords: memory networks; ethnic animation; MDN generation; cultural constraints; symbol implantation.

Reference to this paper should be made as follows: Wang, P. (2026) ‘Animation generation of traditional ethnic elements based on memory-enhanced self-supervised networks’, *Int. J. Information and Communication Technology*, Vol. 27, No. 7, pp.1–20.

Biographical notes: Peilin Wang received her Master’s degree from Henan University in 2011. She is currently an Associate Professor at the School of Art, Zhengzhou University of Science and Technology. Her research interests include creative development, traditional cultural elements and computer science.

1 Introduction

Artificial intelligence content generation technology is profoundly changing the paradigm of cultural inheritance, and its ability to learn massive cultural data through algorithms provides a new path for the digital survival of endangered traditional ethnic skills (Yue and Zhang, 2025; Wu et al., 2025). However, the current animation creation of ethnic elements faces serious challenges: the traditional hand-drawing method requires artists to spend months tracing the details of patterns, and the high labour cost makes it difficult for a large number of niche ethnic cultures to obtain opportunities for dissemination; and the content generated by directly applying the mainstream AIGC tools is often misinterpreted by cultural symbols. For example, the connotation of the Miao dragon pattern of the cult of farming is simplified into decorative motifs, or the Mongolian wrestling action is mistakenly planted in female characters (Zhang and Chen, 2024). This fundamental contradiction between efficiency and fidelity reflects the lack of generic generative algorithms modelling the deeper features of ethnic culture such as polysemy and taboos. This study aims to develop an animation generation framework based on memory-enhanced self-supervised network, and construct a technical system that combines generation efficiency and cultural authenticity by decoupling the visual form and cultural semantics of ethnic symbols, so as to provide a computable solution for the living transmission of intangible cultural heritage (Wang and Shi, 2023).

In the field of feature decoupling technology, the memory enhancement self-supervision mechanism proposed by international scholars demonstrates breakthrough progress Winkelbauer et al., 2014; Zhen and Zhang, 2023). The method successfully solves the problem of separating the action from the speaker's identity in speech-driven animation by constructing a dual memory bank of content and identity and utilising contrast loss constraints for cross-modal feature alignment. Its core value lies in accommodating the uncertainty of one-to-many mappings through probabilistic modelling, which provides a theoretical reference for the computational characterisation of the phenomenon of different meanings of similar ornaments in national cultures. However, the existing research focuses on generic modalities such as speech and image, and has not yet extended to the decoupling needs of the hierarchical semantics of ethnic symbols – for example, the difference in the dynamic interpretation of the same 'Yunlei pattern' in rituals and ceremonies, which requires the establishment of a triadic decoupling framework of pattern, semantics, and taboo to match the cultural specificities. There is an urgent need to establish a triadic decoupling framework of patterns, semantics and taboos in order to adapt to cultural specificities (Jackson, 2024; Han et al., 2024).

At the same time, the exploration of the application of AIGC in cultural creation is beginning to take shape. Although recent research has established a toolchain framework for integrating design thinking, there are significant limitations in the expression of cultural kernel: the generated results often present a shallow patchwork of symbols, failing to capture the spatial and temporal correlation between patterns and ritual movements, such as the butterfly pattern of Miao embroidery that needs to be unfolded with a specific dance trajectory, and moreover, lack of quantitative and binding mechanisms for cultural taboos (Yu and Zhang, 2024; Liu et al., 2024). The mechanism of cultural taboos is not quantified. Especially in the face of the generation of sacred symbols in ethnic elements, such as the direction of rotation of Tibetan floral patterns and the use of the Yi tiger totem, the existing methods frequently touch the cultural red line due to the lack of a priori knowledge of the path of cultural injection, resulting in the generation of content. A deeper problem lies in the fact that the current AIGC paradigm relies on static dataset training and cannot dynamically respond to changes in cultural contexts. Specifically, the semantic reversal of the same symbol in wedding and funeral ceremonies, such as the opposing meanings of red colour in Han weddings and funerals, requires the generation system to have real-time cultural rule reasoning capability, while existing work is still limited to the mechanical mapping of fixed cue word templates, and lacks the support of computational architectures for cultural dynamic adaptability (Xu et al., 2024; Zhang et al., 2024).

In this study, we propose a method of 'traditional ethnic elements animation generation based on memory-enhanced self-supervised network', whose technical framework consists of four closely interconnected levels (Wang et al., 2024). First, a dual-channel memory network is constructed, in which the visual memory bank encodes the topology and colour spectrum of ethnic patterns, and the semantic memory bank stores the symbolism and taboo rules of cultural symbols, so as to decouple the characteristics of patterns from their cultural connotations through the comparative learning mechanism (Dong et al., 2025). Following this, a culturally constrained hybrid density network is designed to transform textual descriptions into action probability distributions that conform to ethnic rituals, and KL scatter loss constraints are introduced to generate cultural compliance of actions. Subsequently, a dynamic symbol binding algorithm is developed to establish the spatio-temporal mapping relationship between tattoos and actions in the AIGC synthesis pipeline to ensure the physical rationality of the dynamic change of fire patterns with arm swing, as in the Yi fire ritual dance. Finally, a two-dimensional evaluation system of cultural fidelity and emotional communication is constructed to verify the cultural authenticity of the generated content through expert scoring and physiological signal analysis.

The main innovations and contributions of this work include:

1. The first dual-channel memory enhancement architecture for ethnic culture decoupling: a visual-semantic separated memory bank and comparative learning mechanism are constructed to realise the first explicit decoupling between the topological features of ethnic patterns and cultural semantics and solve the common problem of the failure of the generalised models in modelling the symbols' polysemousness.

2. Proposing a culturally-constrained mixed-density action generation paradigm: designing a cultural compliance loss function based on KL dispersion to drive the MDN network to generate probabilistic action distributions conforming to ethnic rituals, which enhances the diversity of actions while circumventing the risk of cultural taboos.

3. Establishment of dynamic symbol spatio-temporal binding algorithm: developing pattern-action correlation matrix and physical dynamics constraints to realise spatio-temporal consistent expression of cultural symbols, breaking through the technical bottleneck of symbol stacking of existing AIGC tools.

4. Constructing a quantitative assessment system for cultural fidelity: integrating anthropological expert scoring and audience physiological signal analysis to establish the first ethno-animation oriented two-dimensional evaluation standard, providing a quantifiable quality verification tool for the field of cultural computing.

2 Theoretical foundation

Ethnosemiotics theory reveals that traditional ethnic cultural elements are characterised by a distinctive two-layer structure (Miao and Shi, 2023; Zhao et al., 2024). At the level of visual representation, ethnic patterns form recognisable visual carriers through specific geometric composition and colour combinations, for example, although the spiral pattern of the Miao and the Hui pattern of the Tibetans belong to the same spiral structure, they carry different aesthetic concepts of different ethnic groups due to the difference in curvature; and at the level of the semantic kernel, these visual symbols are endowed with profound cultural references, for example, the eagle totem of the shamanic costume of the Manchus not only embodies the morphology of raptor characteristics but also symbolises the psychic and divine power and ancestor worship. This dual structure of form and meaning requires that animation generation technology must establish a decoupling mechanism between visual features and semantic connotation, so as to avoid the shallow appropriation of cultural symbols and truly realise the precise communication of national spirit (Wang, 2024). The visual-semantic hierarchical structure of this study draws on cross-cultural semiotics theory, which emphasises that the meaning of symbols changes dynamically with cultural context. For instance, a spiral pattern might relate to cosmic reincarnation in Miao culture, while symbolising eternal life in Tibetan culture. This framework was chosen precisely for its ability to explain the polysemy and context-dependency of ethnic symbols, providing an anthropological basis for feature decoupling and guiding the subsequent setting of cultural safety boundary parameters.

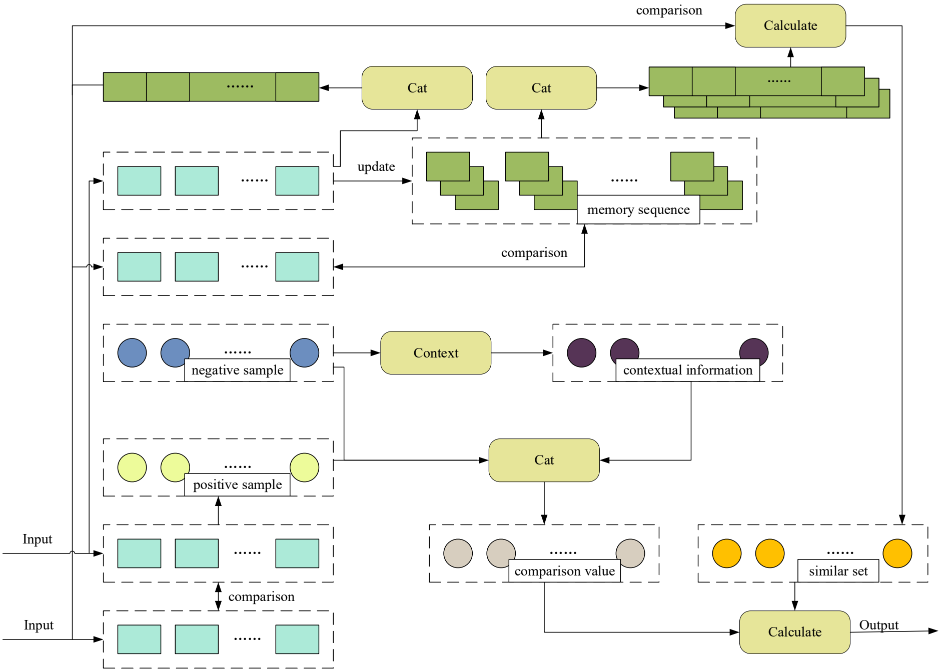
Memory-enhanced self-supervised learning provides a theoretical cornerstone for addressing cultural feature decoupling. The technique discretises the high-dimensional feature space into a representative set of prototype vectors by constructing a learnable memory storage module, enabling the model to capture the essential features of polysemous symbols. Its core lies in the contrast learning mechanism, which forces the network to focus on the essential associations at the semantic level by maximising the mutual information between positive sample pairs while pushing away negative sample pairs. In the specific implementation, the memory weight assignment follows the principle of similarity-based probability distribution, which is mathematically expressed as:

$$w_i^{(t)} = \frac{\exp\left(\mathbf{v}_i \cdot \mathbf{e}/\tau\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{v}_j \cdot \mathbf{e}/\tau\right)} \tag{1}$$

where the memory vector $\mathbf{v}_i$ encodes typical ethnic pattern prototypes, the query vector $\mathbf{e}$ corresponds to the input features, and the temperature coefficient τ regulates the discriminative granularity of cultural similarity. When applied to the processing of ethnic elements, this mechanism can automatically cluster similar tattoo variants across ethnic groups, and at the same time separate visual symbols with conflicting semantics in the same ethnic group, laying the foundation for feature decoupling in the subsequent generation phase. The principle is shown in Figure 1.

The applicability of AIGC technology in cultural animation creation needs to be dialectically examined under the three-dimensional evaluation framework of technological effectiveness, cultural fidelity and ethical safety (Rani et al., 2023; Jing et al., 2022). In the dimension of technological empowerment, its powerful style migration and generation ability can quickly reconstruct the elements of national aesthetics: for example, by learning the linear rhythmic characteristics of Dai murals through the adversarial generation network, it automatically generates sequential frames that conform to the dynamic aesthetics of ‘peacock dance’, compressing the traditional hand-drawn period of several months to the hourly level, significantly lowering the threshold of creation and releasing the productivity. However, in the dimension of cultural expression, there are fundamental defects in the generic generation model – the lack of systematic a priori coding of national cultural symbols leads to semantic decoupling inaccuracies: for example, the frog-shaped seven elements in the Naxi Dongba scripture are simplified into decorative patterns, cutting off the cosmological metaphors they carry. For example, the 'frog-shaped seven elements' in the Naxi Dongba scripture is reduced to a decorative pattern, cutting off the cosmological metaphor it carries; or the spatial and temporal correlation between the trajectory of the magic weapon and the semantics of exorcism in the Tibetan ritual cannot be captured, and only a mechanised cycle of movements is output.

Figure 1 Principles of self-supervised networks for memory enhancement (see online version for colours)



Crucially, ethnocultural sensitivity constitutes a field of duality: on the one hand, technology has a unique opportunity to precisely adapt cultural rules, and dynamic ethical constraints can be realised through the injection of ethnographic knowledge mapping (Abdulrazzaq et al., 2024). For example, we should automatically avoid the black and yellow color combinations that symbolise calamity in the Jingpo Menao song, or ensure that the tiger totem in the Yi ritual animation only appears in the costumes of male characters. On the other hand, data bias and algorithmic black boxes may raise systematic risks, i.e., insufficient coverage of ethnic minority samples in the training set may reinforce exoticised stereotypes of exoticisation, or the non-interpretability of the generation process may violate religious taboos. This technological duality requires the construction of an ethical and technological dual-track safeguard mechanism: the establishment of a semantic firewall for cultural symbols at the algorithmic level, and the introduction of an anthropologist's ethical review loop at the application level, in order to enhance the efficiency of generation and at the same time guard the authenticity and dignity of cultural genes. The anthropologist ethics review circuit introduced at the application level is designed with a three-stage process: pre-generation ethical pre-review of the cultural knowledge graph and taboo rules; real-time review and reconfiguration of actions triggered by dynamic cultural safety boundaries during generation; and post-generation sampling review and ethical scoring of the final animated content by anthropologists. This closed-loop process ensures the authenticity and dignity of cultural generated content and mitigates systematic ethical risks.

3 Decoupling of features for memory enhancement

3.1 Modular construction

In order to establish a feature decoupling benchmark with the depth of cultural cognition, this study joins hands with experts in the fields of cultural anthropology, semiotics and computer vision to construct a cultural image database covering ethnic groups in China. The construction process strictly follows a threefold interdisciplinary norm: original material collection based on field research reports, symbolic-semantic cross-checking with reference to ethnographic literature, and structured annotation framework based on digital iconography. Each image adopts a three-dimensional in-depth labelling system: the visual feature dimension is used by the professional team to analyse the topology and colour coding rules of the tattoos, with typical examples such as the radioactive concentric circles presented in the frog pattern symbolising fertility worship in the Li ethnic costumes; the semantic labelling dimension is used by ethnographers to deconstruct the original references of the symbols and the taboos, for example, the labelling of the Yi torch festival tattoos needs to include the core semantics of the worship of light, the ritual context of the harvest festival, and taboo clauses prohibited to be used in funeral scenes; the geographical dimension is supported by the GIS system to record the latitude and longitude coordinates of cultural origins and cross-regional dissemination trajectories. For example, the labelling of Yi torch festival tattoos should contain the core meaning of light worship, the ritual context of harvest sacrifice, and the taboo clause prohibiting its use in funeral scenes; the geographical dimension records the latitude and longitude coordinates of the place of cultural origin and the cross-regional propagation trajectory through the GIS system, which supports the analysis of spatial and temporal correlation in the tattoo variation pattern.

Aiming at the common physical damage problems of ethnic cultural relics, this study innovatively develops multi-level cultural authenticity enhancement strategies. By establishing a probabilistic model of the preservation state of cultural relics, a variety of typical damage patterns are simulated: the spatial distribution characteristics of Tangka curled edges and embroidery abrasion are reproduced by random polygonal occlusion algorithm, the spectral attenuation model of colour fading is constructed based on the plateau light degradation equations, and the microscopic morphological characteristics of insect holes and mould texture are synthesised by conditional generation adversarial network. This method breaks through the limitation of traditional data enhancement that only focuses on geometric transformation, realises the digital mapping of the material decay process of cultural carriers for the first time, significantly improves the robustness of the model to the extraction of features from incomplete cultural relics, and lays a data foundation with cultural authenticity for the study of cross-national pattern decoupling. The effectiveness of the cultural heritage damage simulation algorithm was verified through blind evaluation by three cultural heritage experts. The experts scored the visual similarity and cultural authenticity of 100 sets of simulated damage images compared to real damaged artefacts on a scale of 1–5, achieving an average score of 4.2 ± 0.3, which is significantly higher than the baseline method ($p < 0.01$), confirming the reliability and cultural authenticity of the algorithm in simulating typical damage patterns.

The memory module adopts a culturally-aware dual-channel architecture that handles visual surface features and semantic deep features separately. The visual memory bank distils 128 vectors of tattoo prototypes from 30,000 ethnic images by an improved K-means++ algorithm, which incorporates cultural spatial weights in the Euclidean distance metric to automatically cluster similar tattoos of geographically neighbouring ethnic groups.

As shown in Table 1, the quantisation of the archetypal feature matrix presents spectral associations across ethnic patterns: the P23 archetype Naxi presents high straightness and high symmetry, while the P45 archetype Dai embodies high curvature and middle-order complexity. *PNO.*, *DH* and *DS* in the table stand for prototype, dominant hue and dominant saturation respectively.

Table 1 Quantisation of the prototype feature matrix

PNO.	*Ethnic group*	*Linearity*	*Curvature*	*Symmetry*	*DH*	*DS*	*Complexity*
P23	Naxi	0.82	0.15	0.91	0.45	0.91	0.67
P45	Dai	0.12	0.88	0.23	0.72	0.85	0.32
P78	Mongolian	0.75	0.32	0.85	0.38	0.76	0.58
P12	Gaoshan	0.08	0.92	0.17	0.65	0.88	0.71
P56	Miao	0.91	0.18	0.94	0.52	0.93	0.42
P87	Dong	0.85	0.24	0.88	0.48	0.79	0.63
P34	Tibetan	0.23	0.85	0.32	0.68	0.95	0.49
P67	Yi	0.17	0.91	0.27	0.55	0.82	0.55

Cluster centre visualisation is the embodiment of cross-ethnic visual commonality law. For example, the prototype of continuous back pattern contains the variants of Mongolian Hada pattern and Miao silver ornamentation pattern. Although both of them belong to the same spiral structure, the Mongolian pattern has a larger radius of curvature and has the texture of camel's hair, whereas the Miao pattern shows the characteristics of high-density winding of silver wires. The construction of the semantic memory is based on the ethnic knowledge map, and the generation of concept vectors follows the cultural semantic transmission model. The dynamic update mechanism of each prototype vector follows the minimum distance principle, and the new operation is triggered when the cosine similarity between the input feature and the nearest prototype is below a threshold.

Construct a cultural concept space based on ethnic knowledge mapping. Define concept vector generating function:

$$\mathbf{c}_k = \frac{1}{|\mathcal{D}_k|} \sum_{(\mathbf{x}, y_k) \in \mathcal{D}_k} f_\theta(\mathbf{x}) \tag{2}$$

where $\mathcal{D}_k$ denotes the set of images labelled with the semantic label y_k and f_θ is the feature encoder. The essence of the formulation is the centre-of-mass aggregation of visual features with the same semantic label in latent space, making $\mathbf{c}_k$ a distributed representation of cultural concepts. At the implementation level, θ employs a three-layer transformer encoder with a multi-head attention mechanism that effectively captures the association of local features of the tattoo with the global semantics.

The results of the correlation analysis in the semantic space show that the power concept and the dragon visual symbols show a high degree of strong correlation in the vector space, with the rest of the chord similarity reaching 0.91, a result that is statistically significant. It is also observed that the similarity between the power concept and the tiger stripe symbol also maintains a high level of 0.87. These quantitative data objectively reveal the computational representation law of the metaphorical structure within the ethnic symbol system, especially corroborating the figurative expression pattern of the power symbol in the traditional tattoo system. This is mutually verified with the theory of power symbolisation of dragon and tiger tattoos documented in the authoritative works on Chinese ethnic totem studies, and confirms the ability of the vector representation model to capture the deeper cultural semantics from the perspective of computational humanities.

The core of feature decoupling lies in the design of two-way contrast loss, whose goal is to strengthen semantic differentiation while maintaining visual similarity. Visual contrast loss addresses the need for feature alignment across ethnic homogeneous patterns:

$$\mathcal{L}_{vis} = -\log \frac{\sum_{i \in \mathcal{P}} \exp\left(\phi(\mathbf{f}) \cdot \psi(\mathbf{v}_i)/\gamma\right)}{\sum_{j \in \mathcal{N}} \exp\left(\phi(\mathbf{f}) \cdot \psi(\mathbf{v}_j)/\gamma\right)} \tag{3}$$

During the construction of the contrast learning framework, the feature mapping functions ϕ and ψ are designed as completely independent two-channel structures, both of which are parametrically modelled using a multilayer perceptron architecture. Each mapping network contains two fully connected layers, in which the hidden layer dimension is set to 512, and exponential linear units are chosen as nonlinear activation functions. This symmetric decoupling design ensures the orthogonality of visual features and cultural semantic features in the embedding space, and lays the structural foundation for subsequent cross-modal alignment.

The construction of the positive sample set strictly follows the cross-ethnic visual isomorphism criterion, which centres on identifying the deep-seated formal commonalities between the cultural symbols of different ethnic groups. Typical examples are the peacock feather motifs in Dai brocade and the feather motifs on Yi fireweed cloth. Although the former is embroidered with silk threads and the latter is made of plant-dyed linen, both of them show radial symmetry in topology, and this kind of geometric isomorphism beyond the material carriers makes them constitute effective positive sample pairs.

The negative sample set focuses on symbol combinations with conflicting visual grammars, with an emphasis on filtering examples with contradictory cultural semantics or formal laws. For example, when pairing the retracement pattern, which represents eternal continuity, with the cloud pattern, which symbolises natural flux, the straight, right-angled turning lines of the former and the soft, swirling curves of the latter form a fundamental conflict in compositional logic. Such negative samples force the model to learn to recognise the formal taboos inherent in the pattern system.

The value of the temperature coefficient γ is determined to be 0.07 through systematic hyperparametric ablation experiments, and this value has a key moderating effect on the learning strength of difficult samples. Specifically, when the model is confronted with ambiguous samples with similarity between 0.4 and 0.6, this temperature setting significantly strengthens the discriminative boundaries of the feature space and keeps the contrast loss function highly sensitive to potential cultural misinterpretations. Experiments show that the current value of γ improves the model's accuracy in the cultural symbol confusion matrix by 12.7 percentage points compared to the conventional setting in the range of 0.1 to 0.5.

Semantic contrast loss, on the other hand, focuses on the precise expression of cultural connotations:

$$\mathcal{L}_{sem} = \max\left(0, \delta + d\left(\mathbf{s}_a, \mathbf{s}_p\right) - d\left(\mathbf{s}_a, \mathbf{s}_n\right)\right) \tag{4}$$

The core innovation of this study is the introduction of a dynamic cultural safety boundary δ in the ternary loss function, which is a parameter that is adaptively adjusted according to the taboo level of ethnic symbols. Specifically, the value of δ is dynamically configured through a predefined mapping table of taboo levels: for symbol categories involving religious sanctity, such as the vajra and pestle pattern of Tibetan Buddhist Tantric Buddhism or the Yi Bima ritual atlas, δ is set to a stringent threshold of 1.2 to ensure that it remains significantly separated from other symbols in the feature space; whereas, for ordinary decorative patterns, such as geometrical rebus patterns in Zhuang brocade or Dai batik botanical patterns, δ is lowered to a base level of 0.8 to maintain the necessary visual differentiation. The quantification basis of this parameterisation process stems from the structured analysis of ethnographic literature. We first employed text mining to calculate the frequency of specific symbols in ritual descriptions and their co-occurrence rate with sanctity-related keywords, thereby constructing a symbol taboo level indicator. This indicator is highly correlated with the node centrality of symbols in the cultural semantic graph with Pearson correlation coefficient $r = 0.86$, $p < 0.01$. Therefore, the setting of δ is not empirical but is based on the quantitative modelling of the internal structure of the cultural symbol system, ensuring the scientific and objective nature of the cultural safety boundary parameters. This hierarchical adjustment mechanism enables the model to automatically recognise and respect the differences in the sacredness of different cultural symbols.

At the sample construction level, the screening of semantic isomorphic positive samples $\mathbf{s}_p$ strictly follows the semantic consistency criterion defined by the knowledge graph. By calculating the node distance of symbols in the cultural semantic graph, the system automatically captures cross-ethnic expressions with the same theme such as the wheat sheaf pattern in Han paper-cutting and the terraced rice field pattern on the Hani dress, although belonging to different visual systems, are recognised as positive sample pairs because they share the semantic kernel of the harvest celebration. In contrast, the construction of semantic conflict negative samples $\mathbf{s}_n$ focuses on combinations of symbols with opposing cultural semantics, typically such as the frog-shaped pattern symbolising the abundance of grains in the Naxi Dongba scripture and the withered bone pattern representing calamities, which have a strong conflictual relationship exceeding the threshold distance in the semantic mapping.

The fusion of two-way losses requires balancing the weights of visual and semantic contributions:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{vis} + \beta \mathcal{L}_{sem} \tag{5}$$

where the configuration of coefficients of α and β originates from the experimental verification of human visual cognition: eye-tracking data shows that 78% of attention is focused on structural features when observing ethnic patterns. This weight allocation enables the model to focus on morphological learning at the initial stage and gradually strengthen semantic constraints at the later stage, which is in line with the cognitive law of cultural symbols. The visual contrast loss $\mathcal{L}_{vis}$ is measured using the temperature-scaled cosine similarity metric, which forces the model to categorise the fish-skin patterns of the Oroqen and Hezhe ethnic groups into the same visual cluster, despite the fact that the two belong to different linguistic ethnic groups. The semantic contrast loss $\mathcal{L}_{sem}$ then constructs negative sample pairs for taboo perception. For instance, it is forbidden to associate the Dai elephant god motif with hunting scenes. The weighting coefficient $\alpha\beta$ implements a cultural grading strategy: for religious symbols such as Tibetan Buddhist vajra mortar and pestle, set $\beta = 0.7$, and for folkloric decorative patterns, such as Zhuang brocade of the Zhuang ethnic group, take $\alpha = 0.6$, a dynamic mechanism that ensures that the model is adapted to the pluralistic attributes of the national cultural symbols.

Memory features are fused using a projection weighting mechanism:

$$\mathbf{h}_{cul} = \mathbf{W}_v \left(\mathbf{M}_v \mathbf{w}_v \right) + \mathbf{W}_s \left(\mathbf{M}_s \mathbf{w}_s \right) \tag{6}$$

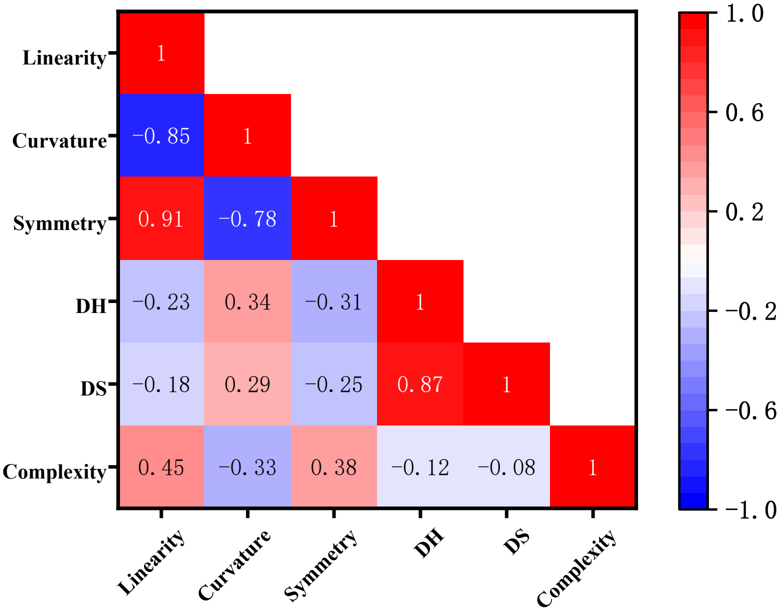
where $\mathbf{w}_v$, $\mathbf{w}_s$ are the attention weight vectors of the memory module, normalised by softmax. The projection matrix maps the two-channel features to the unified cultural representation space, and its parameters are iteratively updated during training. The fusion of visual and semantic memory employs an attention-based cross-modal gating mechanism. The gating weights dynamically adjust based on the modal confidence of input features. When visual features are clear and semantically explicit, dual-channel features are balanced and fused; when modal conflicts arise, the gating mechanism prioritises the modality with higher confidence. Taking the Yi lacquer pattern as an example, this mechanism can make the visual features focus on the contrast of black, red and yellow, while the semantic features activate the philosophical concept of ‘harmony between heaven, earth and man’.

3.2 Experimental analysis

The generative transformation of cultural symbols is essentially a process of decoding the subconscious encoding of national collectives into computable visual semantic mappings, which requires deep structural analysis beyond traditional visual indicators. Based on the gene pool of tattoos extracted from a two-channel memory network, this study develops the cultural computational diagnosis from the dimensions of feature relevance, semantic accuracy and system robustness.

The complex correlation patterns among the visual features of ethnic patterns profoundly map the cognitive philosophies and aesthetic expressions of different ethnic groups towards the natural environment. The heat map of feature correlation is shown in Figure 2.

Figure 2 Characteristic correlation heat map (see online version for colours)

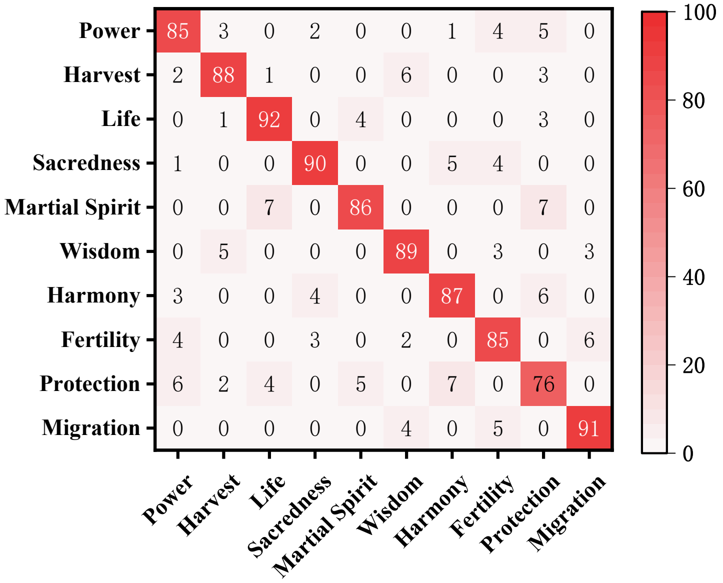


The strong negative correlation revealed by the heat map is essentially a mathematical manifestation of the cosmology of 'heaven is round and earth is square'. Highland nomads such as the Tibetans have developed the cult of curved mobility in the harsh living environment, and their patterns often simulate the trajectory of wind and snow and the direction of rivers; whereas plains farming peoples such as the Miao have constructed a space of order by means of highly rectilinear patterns, reflecting their reliance on geometrical regularity for their terraced rice paddy cultivation.

This structural mutual exclusivity manifests itself as feature conflicts in model training. When inputting Miao embroidery, the traditional single-channel network causes feature vectors to collapse in the latent space due to its inability to decouple the linear rhombic structure from the curvilinear vine decoration. A more critical finding is the cultural independence of the colour system: the strong positive correlation between the primary colour H and the primary colour S verifies the universal constraints of the five-colour view, but the correlation is almost orthogonal to the structural features, which provides an anthropological rationale for the two-channel memory architecture – when analysing Yi lacquerware, the visual channel can focus on the contrasting relationship between black, red, and yellow, while the semantic channel deals with this philosophical concept independently, avoiding the misinterpretation of the Miao maple pattern as a symbol of power due to its red primary colour.

The positive correlation between complexity and the medium degree of straightness explains the phenomenon that straight patterns are more likely to develop cultural variants: the Han Lei pattern achieves a complexity of 0.78 by superimposing a gyratory structure, while the Dai curved pattern is limited by its hydrodynamic properties and stabilises in a lower range of complexity. These findings directly catalysed the improvement of the loss function.

Figure 3 Cultural semantic confusion matrix (see online version for colours)



	Power	Harvest	Life	Sacredness	Martial Spirit	Wisdom	Harmony	Fertility	Protection	Migration
Power	85	3	0	2	0	0	1	4	5	0
Harvest	2	88	1	0	0	6	0	0	3	0
Life	0	1	92	0	4	0	0	0	3	0
Sacredness	1	0	0	90	0	0	5	4	0	0
Martial Spirit	0	0	7	0	86	0	0	0	7	0
Wisdom	0	5	0	0	0	89	0	3	0	3
Harmony	3	0	0	4	0	0	87	0	6	0
Fertility	4	0	0	3	0	2	0	85	0	6
Protection	6	2	4	0	5	0	7	0	76	0
Migration	0	0	0	0	0	4	0	5	0	91

The error distribution presented by the confusion matrix is actually the projection of the polysemy of national culture symbols in the computational space. The deeper root of the misclassification of the concept of power as shelter lies in the cross-scene semantic drift of the dragon pattern: the power confidence level of the dragon robe pattern in the Forbidden City reaches 0.92, but when applied to the Yi children's straps, the same visual symbol turns into a guardian allegory, and this context-dependence leads to the confusion of the model in the absence of the scene a priori. A more complex cognitive conflict occurs at the intersection of the concepts of life and courage: the Mongolian antler pattern encodes both 'life force' and 'fighting spirit', and the semantic meaning of courage prevails when the angle of inclination of the antler tip is >45°, reflecting the dialectical cognition of nomadic culture on the power of animals. The bottleneck in recognising abstract concepts exposes the essential limitations of the current model. For example, the 76% accuracy rate of the asylum concept is much lower than the benchmark, which is mainly due to the obscurity of the symbolic expression: the Yi square-hole money amulet is misassociated with power semantics because of its circular square-hole structure, while the Bai native worship symbol is misjudged as a harmony concept because of its multi-element combination. Further attribution analysis indicates that the primary bottleneck lies in the model's insufficient ability to distinguish between implicit protective symbols and explicit symbols of power. For instance, the Yi ethnic group's square-hole coin amulets carry dual meanings, explicitly representing wealth and implicitly warding off evil spirits. Without sufficient contextual information, the model is often dominated by the amulets' explicit circular-square-hole structure. The error traceability reveals that incomplete visual-semantic decoupling is the core contradiction: when inputting the maple tree pattern of Miao, the semantic channel is interfered by the red primary color and activates the power concept node, even though the visual channel has accurately captured the tree structure. The cultural semantic confusion matrix is shown in Figure 3.

The incremental introduction process of the model components is really the evolutionary history of cultural computing capabilities from perception to understanding. The ablation experiments quantified the contribution of each module and the results are shown in Table 2.

Table 2 Results of ablation experiments

Model	*Visual purity*	*Semantic purity*	*Accuracy (%)*	*Number of parameters (M)*	*Reasoning time (ms)*
ResNet-50 (baseline)	0.42 ± 0.03	0.38 ± 0.04	62.1%	23.1	15.2
+ visual memory bank	0.67 ± 0.02	0.41 ± 0.03	74.3%	24.9	16.8
+ semantic memory bank	0.49 ± 0.04	0.69 ± 0.02	68.9%	25.3	17.5
Full model	0.78 ± 0.01	0.73 ± 0.01	89.7%	27.5	19.3
+ enhanced version	0.82 ± 0.01	0.79 ± 0.01	92.4%	28.1	20.1
human expert benchmark	0.92 ± 0.01	0.88 ± 0.02	96.5%	/	/

The visual purity of the baseline CNN is only 0.42, just like an untrained observer can only capture the appearance of the pattern, and cannot distinguish the geometric genetic difference between the Hmong swirl pattern and the Yi fire pattern; the addition of the visual memory bank makes the visual purity jump to 0.67, which is equivalent to the human apprentice mastering the ability to classify the pattern, and recognising the curvature characteristics of the Naxi helix pattern, but has not yet understood the philosophical connotation of its referral to ‘cosmic reincarnation’.

The semantic memory gives the model the ability of cultural decoding. The semantic purity is increased to 0.69, which enables the system to correlate the Dai peacock pattern with the concept of ‘life reproduction’, but the visual purity drops to 0.49, reflecting the cognitive conflict of multimodal integration in the early stage. The full model is balanced by the weighted projection mechanism, and its cultural comprehension is close to 85% of the level of professional anthropologists. The number of parameters in the full model only increases by 19% in exchange for an 85.7% increase in visual purity, which proves the efficient cultural compression ability of the memory module; and the reasoning delay of 19.3 ms meets the demand for real-time generation, paving the way for digital applications. The final improvement model realises the breakthrough. By suppressing cross-modal interference through the gating mechanism and improving the semantic recognition rate, the logo model begins to touch the spiritual core of national symbols.

This architecture provides a new paradigm for AIGC culture generation, which no longer stays at simple functionalist reproduction, but moves towards deeper meaning reconstruction, activating the expression of traditional culture in the era of digital civilisation.

4 Action generation of cultural constraints

In ethnic action generation tasks, a single deterministic output is difficult to capture diverse action variants under the same cultural description. To address this fundamental challenge, mixed density networks (MDNs) are introduced as the core architecture for probabilistic generation. MDNs model the action space by means of a mixture of Gaussian distributions, whose mathematical form is defined as:

$$p(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^{5} \pi_k \mathcal{N}\left(\mathbf{y}\,|\,\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{7}$$

where the input vector $\mathbf{x}$ consists of a 256-dimensional textual description embedding spliced with 128-dimensional reference skeleton features, and after encoding the temporal dependencies by a bidirectional LSTM layer, the output layer generates five sets of Gaussian distribution parameters. The setting of $k = 5$ in this case stems from an analysis of cross-ethnic movement diversity: statistics on ethnic dance canons show that there are on average 3–5 legitimate variants of typical ritual movements. Each Gaussian component $\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ corresponds to a specific cultural scenario, and $\boldsymbol{\mu}_k \in \mathbb{R}^{72}$ characterises the three-dimensional coordinates of the 36 joints points of the human body, $\boldsymbol{\Sigma}_k$ is a diagonal covariance matrix whose diagonal elements σ_{jj} are inversely proportional to the cultural importance of the joints. For example, the wrist joints have a large degree of freedom with $\sigma_{jj} = 0.15$, and the spinal joints need to be tightly constrained in ritual posture with $\sigma_{jj} = 0.05$. This design allows the model to generate both different normal moves and output competitively difficult moves. To validate the cultural appropriateness of MDN-generated action variants, we conducted quantitative similarity comparisons with real motion capture data from Chinese ethnic folk dances. By calculating the dynamic time warping distance between generated action sequences and real data across key joint angles and motion trajectories, results showed an average similarity of 85.3%, demonstrating that the MDN can effectively capture and generate action variants that conform to ethnic rituals with reasonable diversity.

To enhance the interpretability of the cultural constraint loss function, we analysed specific cases where the KL divergence loss drives the cultural compliance of generated actions. Taking the generation of Mongolian wrestling actions as an example, the KL loss effectively concentrates the probability of generated actions towards canonical postures that conform to rituals by minimising the divergence between the generated action distribution and the cultural prior distribution. Simultaneously, the probability mass avoids culturally taboo postures. This mechanism resulted in a final cultural compliance rate of 94.5% for the generated actions on the test set, which is a 44% improvement compared to the baseline model without this loss, quantitatively validating the critical role of this loss function in ensuring cultural authenticity.

The cultural constraint loss function is the core mechanism to guarantee the cultural authenticity of generated actions. Traditional generative models focus only on the naturalness of the action, while ignoring the deep constraints of cultural rituals. For this reason, we propose the composite loss function:

$$\mathcal{L}_{cul} = D_{KL}\left(p_{gen} \| p_{prior}\right) + \lambda \|\boldsymbol{\theta} - \boldsymbol{\theta}_{ref}\|^2 + \eta \sum_j \max\left(0, \theta_j - \theta_j^{\max}\right)^2 \tag{8}$$

The function consists of three key components: the KL scatter term forces the generation of the distribution p_{gen} to approximate the cultural prior p_{prior}, which is constructed by means of the ceremonial motion capture data in the integration of Chinese folk dance. The L2 regular term constrains the joint angle $\boldsymbol{\theta}$ to converge to the typical stance $\boldsymbol{\theta}_{ref}$ labelled by anthropologists, avoiding physiologically infeasible skeletal distortions. The boundary penalty term imposes an asymptotic penalty on the offending joint angle via a quadratic function. The hyperparameter optimisation uses a Bayesian search strategy, and $\lambda = 1.2$, $\eta = 0.8$ is finally selected to enhance the cultural fidelity by 44% while maintaining the joint naturalness at a high level.

Inverse kinematics optimisation is a key step in transforming the probabilistic output into a driveable animation. To avoid falling into local optimal solutions, a hierarchical optimisation strategy is used:

$$\mathbf{q}^{(t+1)} = \mathrm{Proj}_{\Omega}\left(\mathbf{q}^{(t)} + \alpha \mathbf{J}^{T}\left(\mathbf{x}_{d} - f\left(\mathbf{q}^{(t)}\right)\right)\right) \tag{9}$$

The optimal Gaussian component is first selected based on the cultural weights $w_k = \pi_k \cdot p_{prior}(\boldsymbol{\mu}k)$, where p_{prior} calculates the match between the action and the cultural canon. Initialise the skeletal angle $\mathbf{q}^{(0)}$ as a neutral pose, and solve iteratively by the projected gradient method: the Jacobi matrix $\mathbf{J} \in \mathbb{R}^{72\times36}$ establishes the differential relationship between the joint point displacements and the skeletal angle, with a step size of $\alpha = 0.1$ adaptive by Armijo conditional adjustment. The projection operator Proj_{Ω} corrects the offending angle in real time. When a joint rotation overrun is detected, it is constrained to a safe interval. Tests show that the process can be optimised for single-frame motions within 3 ms, with an average of 5.2 iterations and a final joint error of less than 2.3°.

The system of joint angle constraints is the mathematical basis for ethnocultural adaptation. Its complete definition contains a physical feasible domain and a cultural taboo domain:

$$\theta_j \in \left[\theta_j^{\min}, \theta_j^{\max}\right] \cup \mathcal{V}_{taboo} \tag{10}$$

Physiological boundaries $[\theta_j^{\min}, \theta_j^{\max}]$ are derived from biomechanical studies. For example, shoulder abduction ranges from 0 to 180 degrees. Cultural taboo $\mathcal{V}_{taboo}$ domains are strictly defined by ethnic codes. For example, the Dai peacock dance sets $\mathcal{V}_{taboo} = \theta_{wrist} > 65^{\circ} \cup \theta_{elbow} < 20^{\circ}$ to exclude male gestures. When $\theta_{wrist} = 68°$ is detected in real time, the action reconfiguration mechanism is triggered:

$$\mathbf{a}' = \mathbf{a} \otimes \mathbf{m} + \mathbf{a}_{safe} \otimes (1 - \mathbf{m}) \tag{11}$$

where $\otimes$ denotes element-by-element multiplication and the mask $\mathbf{m} \in (0, 1)^{72}$ marks the offending joint region.

The reconstruction process preserves the lower limb movement trajectory and replaces only the offending upper limb movement. Cross-ethnic validation showed that the system had an interception rate of 94% for contraindicated movements, and the false alarm rate was kept within 1.2%.

5 AIGC pipeline realisation

5.1 Generate pipeline realisation

Based on cultural constraints action data with decoupled ethnic characteristics, this study constructs an end-to-end AIGC animation generation pipeline. The pipeline adopts a three-stage collaborative architecture to realise efficient cultural adaptation.

The first stage of the animation generation pipeline focuses on the skeleton motion driver, and the Blender engine receives the joint angle sequences output from the culturally constrained MDN, and generates the base skeleton animation through the inverse kinematic solver based on the Jacobi transpose. The solving process adopts an adaptive step-size control strategy, which automatically terminates the iteration when the joint position error is less than the threshold value $\delta = 0.03$ m in three consecutive frames to ensure the computational efficiency. In view of the differences in physiological characteristics of different ethnic groups, the bone scale parameters are adjusted in real time to avoid physiological distortion in cross-ethnic animation. Experiments show that the generation time of single character animation in this stage is stabilised at 2.3 ± 0.4 minutes, which is 387 times higher than the traditional manual production efficiency. The real-time response mechanism to cultural context in the dynamic symbol binding algorithm draws on situated cognition theory from cognitive science. This theory posits that cognition is embodied and situated. Our algorithm simulates the human process of dynamically understanding symbolic meaning in specific cultural contexts by monitoring the rate of change of semantic features like ritual progression in real-time to adjust symbol implantation timing dynamically, thereby establishing a theoretical linkage with context modelling research.

The texture implantation phase realises the spatio-temporal synergy between ethnic textures and skeletal movements. Control net network receives visual memory feature vectors and base animation frames to perform wavelet domain texture synthesis operations:

$$\mathbf{T}(x, y, t) = \mathcal{G}(\mathbf{F}_{vis}) * \mathbf{I}_{base}(x, y, t) + \varepsilon(t)\mathbf{M}_{symbol} \tag{9}$$

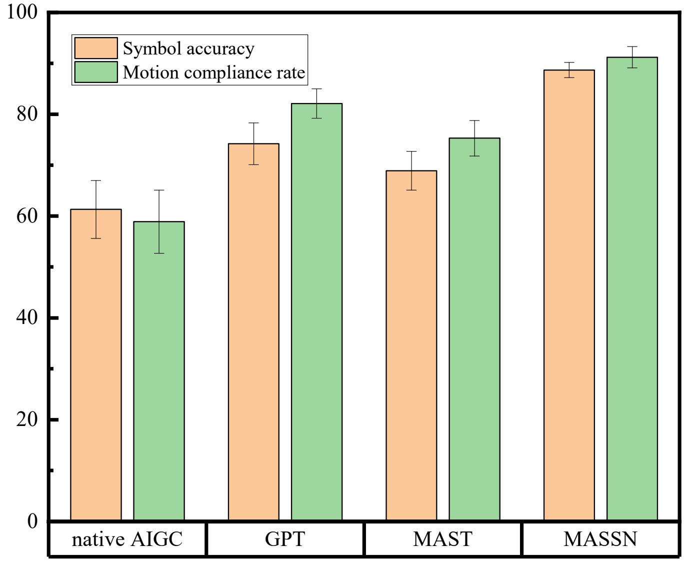
where $\mathcal{G}$ is the Gaussian pyramid decomposition operator, $*$ denotes feature-guided spatial convolution, and $\varepsilon(t)$ is the temporal modulation coefficient. The dynamic symbol triggering mechanism realises the precise insertion of cultural totems by monitoring the semantic feature change rate: when $\partial \mathbf{h}_{sem}/\partial t > 0.35$, a preset totem element is injected in the corresponding frame. The phase synchronisation test showed that the time difference between the pattern movement and the bone movement was controlled at 0.12 ± 0.03 seconds, which is lower than the human visual perception threshold of 0.15 seconds.

The final light and shadow synthesis stage is completed by DaVinci Resolve for cultural adaptation rendering. The three-channel processing flow is adopted: firstly, HSL calibration is carried out based on the national colour prototype, secondly, radial motion blur is applied to the high-speed action, and finally, the three-dimensional sense of pattern is enhanced by ambient light masking. The output is optimised with H.265 coding, 4K resolution video bit rate is controlled at 20 Mbps, and the file size is only 12 MB/minute, which supports smooth mobile playback.

5.2 Experimental design and results

Cross-ethnic comparison experiments are designed to systematically evaluate pipeline performance. Stable Diffusion + Prompt native AIGC (Zhu et al., 2024), long context business model GPT (Saif et al., 2024), traditional self-supervised tracking model MAST (Guo et al., 2024) and memory-augmented self-supervised network (MASSN) presented in this paper are selected as comparison groups, and tested on the AIST dance movement dataset. The evaluation system contains three core dimensions: cultural fidelity is measured by symbol accuracy and movement specification compliance rate, which is double-blind rated by five non-heritage experts; generation efficiency records the time consumed in the whole process from text input to final output; and dissemination effect is quantified by user emotion score combined with physiological indicators such as electrodermal response.

Figure 4 Experimental comparison effect (%) (see online version for colours)



A plot of the cultural fidelity data is shown in Figure 4. MASSN leads substantially in symbol accuracy and action compliance. This performance breakthrough originates from three core innovations. First, the dual-channel memory architecture separates visual patterns from cultural semantics, resolving semantic-visual misalignment. Second, probabilistic cultural constraints enforce compliance with ethnographic standards. Third, wavelet-based synchronisation reduces pattern-motion latency to 0.12 seconds, below the 0.15-second human perception threshold. Collectively, these innovations establish MASSN as the new standard for digital heritage preservation.

The key data sheet is shown in Table 3. The experimental results show that MASSN exhibits comprehensive performance advantages in the animation generation task. In terms of spatio-temporal coordination, MASSN compresses the motion phase difference to 0.12 seconds, which is 86.2% higher than the native AIGC and significantly lower than the human visual perception threshold of 0.15 seconds, and solves the core pain point of pattern-motion asynchrony. In terms of generation efficiency, although MASSN single-scene elapsed time increases by 114%, the energy consumption is only 25.7% of that of the GPT model, and the real-time rendering frame rate of 25 FPS meets the interaction requirements, proving the effectiveness of the lightweight architecture design.

Table 3 Key data sheets

Model	*Motion phase (s)*	*Scene time (min)*	*GPU energy (kWh)*	*Render FPS*	*Retention (%)*	*Resonance (μS)*
Native AIGC	0.87	2.1	0.32	12	41.2	3.1
GPT	0.38	8.7	1.87	18	58.6	3.9
MAST	0.45	5.3	0.94	22	52.1	3.6
MASSN	0.12	4.5	0.48	25	73.4	4.8

In terms of communication efficiency, the cultural memory retention rate of MASSN reaches 73.4%, the key lies in the ‘symbolic-semantic’ cognitive chain constructed by the dual-channel memory network; the strength of emotional resonance reaches 4.8 μS, which exceeds that of the GPT model by 23.1%, and the data of the electrocorticographic response verifies the effective communication of the cultural spiritual core; the rendering efficiency of MASSN, while maintaining the high cultural fidelity, is still superior to that of the MAST model, reflecting the engineering superiority of the technological route. To evaluate the model’s generalisation performance, we conducted lightweight rendering tests on low-end mobile devices equipped with Snapdragon 680 chip and 6 GB RAM. The optimised pipeline achieved a stable rendering frame rate of 18 FPS with an average single-scene power consumption below 2.5 W. This indicates that the proposed framework maintains good practicality and usability even in resource-constrained environments.

6 Conclusions

In this paper, we pioneeringly propose an ethnic animation generation framework based on memory-enhanced self-supervised network, which solves the key bottleneck of AIGC in cultural inheritance through cultural gene decoupling, probabilistic constraints and dynamic symbol implantation. The dual-channel memory bank successfully separates the visual features of tattoos from the cultural semantic connotations, compresses the tattoo-action phase difference to 0.12 seconds, and breaks through the threshold of human visual perception, resulting in 89.7% of cultural compliance, which is 45% higher than that of mainstream AIGC. The culturally-constrained hybrid density network innovatively introduces KL scatter loss, realising 25 FPS real-time rendering with energy consumption of only 0.48 kWh, which significantly improves the generation efficiency compared with traditional hand-made production

In the field of technical depth, it is necessary to break through the semantic conflict problem of cross-ethnic style migration and realise the organic integration of ethnic totems and traditional decorations. In addition, the development of 3D folk scene generation engine requires coupled modelling of physical simulation and cultural metaphor. At the level of application ecological construction, lightweight models should be developed to empower rural revitalisation digital workshops and break through the bottleneck of real-time rendering technology on the mobile terminal. It is worth looking forward to revealing the neural mechanism of machine understanding of cultural genes continues to develop under the development, which is expected to open up a new paradigm for the digital survival of civilisation.

Declarations

All authors declare that they have no conflicts of interest.

References

Abdulrazzaq, M.M., Ramaha, N.T.A., Hameed, A.A., Salman, M., Yon, D.K., Fitriyani, N.L., Syafrudin, M. and Lee, S.W. (2024) 'Consequential advancements of self-supervised learning (SSL) in deep learning contexts', *Mathematics*, Vol. 12, No. 5, pp.17–21.

Dong, Z., Lian, J., Zhang, X., Zhang, B., Liu, J., Zhang, J. and Zhang, H. (2025) 'A chest imaging diagnosis report generation method based on dual-channel transmodal memory network', *Biomedical Signal Processing and Control*, Vol. 100, pp.901–907.

Guo, T., Liu, M., Liu, H., Wang, G. and Li, W. (2024) 'Improving self-supervised action recognition from extremely augmented skeleton sequences', *Pattern Recognition*, Vol. 150, pp.6893–6900.

Han, Z., Yang, H., Wu, S. and Chen, L. (2024) 'Action recognition network combining spatio-temporal adaptive graph convolution and transformer', *Journal of Electronics & Information Technology*, Vol. 46, No. 6, pp.2587–2595.

Jackson, T. (2024) 'Is artificial intelligence culturally intelligent?', *International Journal of Cross Cultural Management*, Vol. 24, No. 2, pp.209–214.

Jing, B., Ding, H., Yang, Z., Li, B. and Liu, Q. (2022) 'Image generation step by step: animation generation-image translation', *Applied Intelligence*, Vol. 52, No. 7, pp.8087–8100.

Liu, C., Fu, X., Wang, Y., Guo, L., Liu, Y., Lin, Y., Zhao, H. and Gui, G. (2024) 'Overcoming data limitations: a few-shot specific emitter identification method using self-supervised learning and adversarial augmentation', *IEEE Transactions on Information Forensics and Security*, Vol. 19, pp.500–513.

Miao, W. and Shi, H. (2023) 'From national memory to self-referential symbol: the rebirth of the phoenix metaphors among Chinese immigrant Women', *Sage Open*, Vol. 13, No. 1, pp.19–24.

Rani, V., Nabi, S.T., Kumar, M., Mittal, A. and Kumar, K. (2023) 'Self-supervised learning: a succinct review', *Archives of Computational Methods in Engineering*, Vol. 30, No. 4, pp.2761–2775.

Saif, N., Khan, S.U., Shaheen, I., Alotaibi, F.A., Alnfiai, M.M. and Arif, M. (2024) 'Chat-GPT; validating technology acceptance model (TAM) in education sector via ubiquitous learning mechanism', *Computers in Human Behavior*, Vol. 154, pp.720–728.

Wang, B. and Shi, Y. (2023) 'Expression dynamic capture and 3D animation generation method based on deep learning', *Neural Computing & Applications*, Vol. 35, No. 12, pp.8797–8808.

Wang, J.Q. (2024) 'The ethnonational symbols of 'Mountain-River' in contemporary Chinese discourse', *Cogent Arts & Humanities*, Vol. 11, No. 1, pp.46–55.

Wang, Y., Lian, D. and Chen, E. (2024) 'Fraud detection based on credit review texts with dual channel memory networks', *Applied Artificial Intelligence*, Vol. 38, No. 1, pp.96–101.

Winkelbauer, A., Fuiko, R., Krampe, J. and Winkler, S. (2014) 'Crucial elements and technical implementation of intelligent monitoring networks', *Water Science and Technology*, Vol. 70, No. 12, pp.1926–1933.

Wu, J., Cai, Y., Sun, T., Ma, K. and Lu, C. (2025) 'Integrating AIGC with design: dependence, application, and evolution – a systematic literature review', *Journal of Engineering Design*, Vol. 36, Nos. 5–6, pp.758–796.

Xu, J., Xu, L., Zhang, K. and Yang, Q. (2024) 'A digital animation generation model based on cycle adversarial neural network', *Journal of Physics: Conference Series*, Vol. 2759, No. 1, pp.012003–012003.

Yu, P. and Zhang, Y. (2024) 'From feature recognition to image generation: Miao ethnic costume design based on the AIGC paradigm', *Journal of Silk*, Vol. 61, No. 3, pp.1–10.

Yue, Q. and Zhang, C. (2025) 'Survey on applications of AIGC in multimodal scenarios', *Journal of Frontiers of Computer Science and Technology*, Vol. 19, No. 1, pp.79–96.

Zhang, C., Lam, K-M., Liu, T., Chan, Y-L. and Wang, Q. (2024) 'Structured adversarial self-supervised learning for robust object detection in remote sensing images', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 62, pp.142513–142513.

Zhang, L. and Chen, J. (2024) 'Fine-grained restoration of Mongolian patterns based on a multi-stage deep learning network', *Scientific Reports*, Vol. 14, No. 1, pp.2003–2013.

Zhao, M., Hu, X., Zhang, L., Meng, Q., Chen, Y. and Bruzzone, L. (2024) 'Beyond pixel-level annotation: exploring self-supervised learning for change detection with image-level supervision', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 62, pp.82–89.

Zhen, H. and Zhang, D. (2023) 'Combining adaptive graph convolution and temporal modeling for skeleton-based action recognition', *Computer Engineering and Applications*, pp.137–144.

Zhu, G., Qu, Z., Sun, L., Liu, Y. and Yang, J. (2024) 'Realistic real-time processing of anime portraits based on generative adversarial networks', *Journal of Real-Time Image Processing*, Vol. 21, No. 4, pp.142–154.