International Communication Effectiveness of Traditional Chinese Textile Culture Based on a Transformer Multimodal Fusion Model

Haixia Xu
International Education College, Beijing Institute of Graphic Communication, Beijing 102600, China
Corresponding author: Haixia Xu (e-mail: 15801364102@163.com)

ABSTRACT With the rapid development of multimodal information transmission and intelligent semantic communication technologies, improving cross-modal consistency has become increasingly important for digital cultural dissemination and future communication systems. To address the problem of cross-modal semantic inconsistency in the international communication of traditional Chinese textile culture, this study proposes a culture-aware multimodal fusion framework based on Transformer architecture. Multilingual data collected from mainstream social media platforms are first utilized to construct a domain-specific knowledge graph covering craftsmanship, materials, patterns, and symbolic meanings. A cultural entity perception module is then introduced to enhance intra-modal feature representation, while a gated cross-modal attention mechanism dynamically aligns semantic information among text, images, and audio. Furthermore, a hierarchical fusion strategy generates unified communication representation vectors for communication effectiveness prediction. Experimental results demonstrate that the proposed framework achieves superior performance in cultural symbol recognition, cross-modal semantic alignment, and sentiment consistency, yielding an average Comprehensive Effectiveness Index (CEI) of 0.725 and significantly improving semantic fidelity and audience acceptance in international communication. By integrating knowledge-guided multimodal representation learning with adaptive semantic alignment, the proposed method provides an effective technical paradigm for intelligent cultural communication and offers valuable insights for multimodal semantic transmission, semantic communication architectures, and future wireless information dissemination systems.

INDEX TERMS multimodal fusion, transformer, semantic communication, knowledge graph, traditional Chinese textile culture

I. INTRODUCTION

WITH the deep integration of digital technology and global social media, traditional Chinese textile culture, as a vital component of intangible cultural heritage, is rapidly gaining international recognition through multimodal formats such as video, graphics, and live broadcasts [1-3]. However, in the process of cross- linguistic and cross-cultural communication, cultural symbols often suffer from semantic inconsistencies between text, images, and audio modalities, leading to simplification, misinterpretation, and even loss of deeper cultural connotations [4,5]. Existing AI (Artificial Intelligence)-driven communication analysis models focus primarily on information breadth. When dealing with international communication scenarios, they neglect the precise alignment of cultural semantics and multimodal consistency across language and cultural contexts [6-8]. This makes it difficult to effectively address the misinterpretation and distortion of symbolic connotations caused by language barriers and cultural differences, and prevents comprehensive international communication effectiveness evaluation and optimization. Therefore, there is an urgent need to build intelligent models that combine cultural understanding with multimodal fusion precision to enhance the international dissemination of traditional Chinese culture. Traditional Chinese textile culture encompasses multiple dimensions, including silk craftsmanship, ethnic costumes, hand-dyed printing, and symbolic patterns, possessing a high degree of visual expressiveness and cultural complexity. Zhang et al. [9] revealed the historical evolution of brocade patterns popular in China during the Tang and Ming dynasties and summarized their unique artistic style. Li et al. [10] found through analyzing Li Ziqi's YouTube channel that Li Ziqi presents Chinese culture through various symbols, and the audience interprets her content in a variety of

ways. Gao et al. [11] demonstrated the concept of totem in traditional Chinese clothing design, summarized the representative elements and meanings of totems, and revealed the cultural connotations of Chinese patterns widely used in modern clothing design. Zhang [12] used water ripples and cloud patterns as examples to explore the evolution and contemporary expression of traditional Chinese decorative patterns. The results showed that water ripples and cloud patterns represent harmony, perfection, and auspiciousness. Yuan et al's [13] research confirmed that the brocade textiles of the Yao ethnic minority in China embody ecological principles. Therefore, achieving a unified cross-modal semantic modeling of traditional Chinese textile cultural symbols is a key prerequisite for improving the quality of their international communication. To address the problem of multimodal semantic alignment, researchers have proposed a variety of fusion models. Gao et al. [14] proposed a visual and language model to enhance its ability to discern the authenticity of news content. Xia et al. [15] developed a clothing matching system based on knowledge graphs that integrates cultural background, clothing knowledge, and expert comments to ensure cultural consistency and historical accuracy of recommendations. Zhao et al. [16] proposed a framework for automotive interior texture design that integrates cultural semantic modeling and generative design technology, effectively improving the cultural expressiveness and emotional recognizability of automotive interior design. Al-Buraihy et al. [17] solved the cross-language image description problem by combining a neural network model with a novel approach of semantic matching technology. Bi et al. [18] used the bidirectional encoder representation BERT (Bidirectional Encoder Representations from Transformers) of Transformer to extract semantic features from text, and used Mel-frequency cepstral coefficients and residual neural networks to capture key features from speech and facial expressions to enhance multimodal emotion recognition capabilities. Existing models still have obvious shortcomings in cultural specificity modeling, multimodal dynamic alignment, and knowledge-guided fusion. To address these challenges, this study introduces a culturally aware multimodal fusion framework. This framework, with the Transformer at its core, deeply couples domain knowledge graphs to achieve the entire process from cultural feature enhancement and dynamic semantic alignment to the generation of unified communication representations. This method first systematically collects data from YouTube, Instagram, and TikTok platforms and constructs a high-quality annotation set and domain-specific knowledge graph. Secondly, it designs a culture perception multimodal encoder and introduces a knowledge graph-guided cultural entity fusion mechanism during the encoding phase. Subsequently, it achieves dynamic semantic alignment through a gated cross-modal attention mechanism, adaptively adjusting the alignment strength. Finally, a layered cross-modal fusion strategy was employed to generate a unified communication representation vector, which was then used to drive multi-task prediction of communication effectiveness. This innovative approach deeply couples cultural knowledge graphs into the entire multimodal fusion process, aiming to address the cross-modal semantic inconsistency in the international communication of traditional Chinese textile culture and achieve dynamic semantic alignment and representation generation driven by cultural context.

II. ALGORITHM DESIGN

A. MULTIMODAL DATA COLLECTION AND ANNOTATION

This paper systematically collects multimodal data from three major international social platforms: YouTube, Instagram, and TikTok. The paper constructs a query set using keywords such as "Chinese traditional textile", "yunjin", and "su embroidery", along with hashtags. For data retrieval, official platform APIs were used where available and permitted by the Terms of Service (TOS). In cases where API access was limited for specific metadata or content types, the Selenium automation framework was employed to simulate user behavior in a manner compliant with each platform's TOS, ensuring all collected data was publicly accessible. This approach allowed for the extraction of video URLs (Uniform Resource Locators), titles, descriptions, posting times, geotags, and interaction metadata, and extracting a keyframe image set (I={I1,I2,...,IT}). The audio stream is resampled to 16kHz and then segmented into 2-second sliding window segments (A={A1,A2,...,AS}). Text modalities include video subtitles, release notes, and user comments. To process multilingual comments and ensure the uniformity of model input and the operability of subsequent semantic analysis, non-English texts are uniformly translated into English while retaining the original language tags. A semantic cleaning process is performed on all collected multilingual text data, and metadata (such as video subtitles, release notes, user comments, titles, descriptions, posting times, and geotags) are extracted and standardized. The annotation phase employed a hierarchical annotation protocol, defining the entity set Evisual= {Yunjin Brocade, Su Embroidery, Blue Printed Cloth, Miao Embroidery, Shu Brocade, Chinese Knot Button} and further subdividing it into subcategories. All data was collected from publicly accessible content on social media platforms, ensuring compliance with the respective platforms' terms of service. This research focused solely on the semantic, sentimental, and cultural attributes of the content and did not involve tracking or identifying any individual users. Text and comment annotation encompassed two dimensions: sentiment polarity and cultural understanding. A double-blind annotation mechanism was employed, with labels ultimately determined by majority voting. Consistency was assessed using Krippendorff's α coefficient, defined as:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1}$$

In formula (1), D_o is the observed difference, D_e is the expected difference, and when $\alpha > 0.85$, it is considered to

be highly consistent [19]. Figure 1 illustrates the complete process for collecting and annotating multimodal data on traditional Chinese textile culture. First, a query set was constructed using Chinese and English keywords and hashtags from YouTube, Instagram, and TikTok. Publicly available content was collected using official APIs and compliant automation tools. Next, metadata was extracted from the videos, and text content, including subtitles, release notes, and comments, was translated into English while retaining the original language tags. Finally, bounding box annotations were performed for six categories of cultural symbols within the images. Double- blind annotation by five expert annotators was employed, with final labels determined by majority voting. Annotation consistency was rigorously assessed using Krippendorff's α coefficient, which achieved a value greater than 0.85 across all categories, indicating high inter-rater reliability. The annotation protocol covered two primary dimensions—sentiment polarity and cultural understanding—and included hierarchical labeling for six specific cultural symbol categories (Yunjin Brocade, Su Embroidery, Blue Printed Cloth, Miao Embroidery, Shu Brocade, and Chinese Knot Button).

B. CONSTRUCTION OF A KNOWLEDGE GRAPH OF TRADITIONAL CHINESE TEXTILE CULTURE

This paper constructs a knowledge graph of traditional Chinese textile culture, encompassing ancient texts, national intangible cultural heritage lists, museum collections, and authoritative academic research. By combining named entity recognition with relationship extraction, four core entities are extracted from unstructured text: craftsmanship (Ctech), materials (Cmat), patterns (Cpat), and symbolic meaning (Csym). Entity set E=Ctech∪Cmat∪Cpat∪Csym∪Cprod is defined, while Cprod represents the product class. Five types of ontology relationships are constructed based on semantic logic: r1:uses_material (eh, et) indicates that a craft uses a certain material; r2:embodies_pattern (eh, et) indicates that the product contains a certain pattern; r3:symbolizes (eh, et) indicates that the pattern or craft contains a certain symbol; r4:belongs_to (eh, et) indicates that the craft belongs to a certain school or region; r5:applied_in (eh, et) indicates that the craft is applied to a certain type of clothing. The TransE (Translating Embeddings) model is used to map entities and relations into a continuous vector space. For each triple (h,r,t), the goal is to minimize the following energy function:

$$L_{\mathrm{TransE}} = \sum_{(h,r,t)\in K} \Big[\gamma + \|e_h + r_r - e_t\|_2^2 - \min_{t'\in E} \|e_h + r_r - e_{t'}\|_2^2\Big]_+ \tag{2}$$

In formula (2), e_h, e_t are the head and tail entity embeddings, and K is the set of positive sample triplets. Negative sampling generates negative examples by uniformly replacing the head or tail entity. The optimization process uses stochastic gradient descent, and the loss function is supplemented with an L2 regularization term:

$$L = L_{\mathrm{TransE}} + \lambda \left(\|E\|_F^2 + \|R\|_F^2\right) \tag{3}$$

In formula (3), E, R are the entity and relation embedding matrices, respectively, and $\|\cdot\|_F$ is the Frobenius norm.

C. DESIGN OF A CULTURALLY AWARE MULTIMODAL ENCODER

This paper constructs a culturally aware multimodal encoder to perform feature extraction and cultural enhancement on image, text, and audio modalities. Each modality encoder is initialized based on a pre- trained architecture and introduces a knowledge graph-guided cultural entity fusion mechanism to ensure that the intramodal representations are culturally semantically sensitive [20]. For the image input $X_v \in \mathbb{R}^{H\times W\times 3}$, Vision Transformer (ViT) is used to split it into N non-overlapping image blocks $\{x_i^v\}_{i=1}^N$. After linear projection, position encoding is added to obtain the initial visual sequence:

$$Z_v^{(0)} = \{E_v x_i^v + p_i\}_{i=1}^N \tag{4}$$

The cultural entity perception module is introduced at the bottom layer of ViT. The knowledge graph entity embedding set is set to $G = \{g_k\}_{k=1}^K \subset \mathbb{R}^d$, and the cosine similarity between the current visual feature and g_k is calculated:

$$s_{ik} = \frac{\left(z_i^{(l_0)}\right)^\top g_k}{\left\|z_i^{(l_0)}\right\| \|g_k\|} \tag{5}$$

For each $z_i^{(l_0)}$, select the top 5 cultural entities with the highest similarity to form the association set $K_i = \mathrm{TopK}(s_{i\cdot}, 5)$. Perform weighted aggregation on the corresponding entity vectors:

$$c_i = \sum_{k\in K_i} \alpha_{ik} g_k, \quad \alpha_{ik} = \frac{\exp(s_{ik})}{\sum_{k'\in K_i} \exp(s_{ik'})} \tag{6}$$

The cultural vector c_i is concatenated with the original features and then fused through linear transformation:

$$\tilde{z}_i^{(l_0)} = W_f \left[z_i^{(l_0)}; c_i\right] + b_f \tag{7}$$

After replacing the original features, the subsequent Transformer layer continues to calculate and output the enhanced visual representation sequence $H_v \in \mathbb{R}^{N\times d}$. For the text input $X_t = [w_1, w_2, \ldots, w_M]$, the BERT model is used to generate the contextual representation $H_t^{(0)} = \mathrm{BERT}(X_t) \in \mathbb{R}^{M\times d}$. Cultural fusion is also introduced at the bottom layer: each word vector $h_j^{(l_0)}$ is matched with the knowledge graph entity for similarity to generate a cultural context vector c_j, and the representation is updated through a gated fusion mechanism:

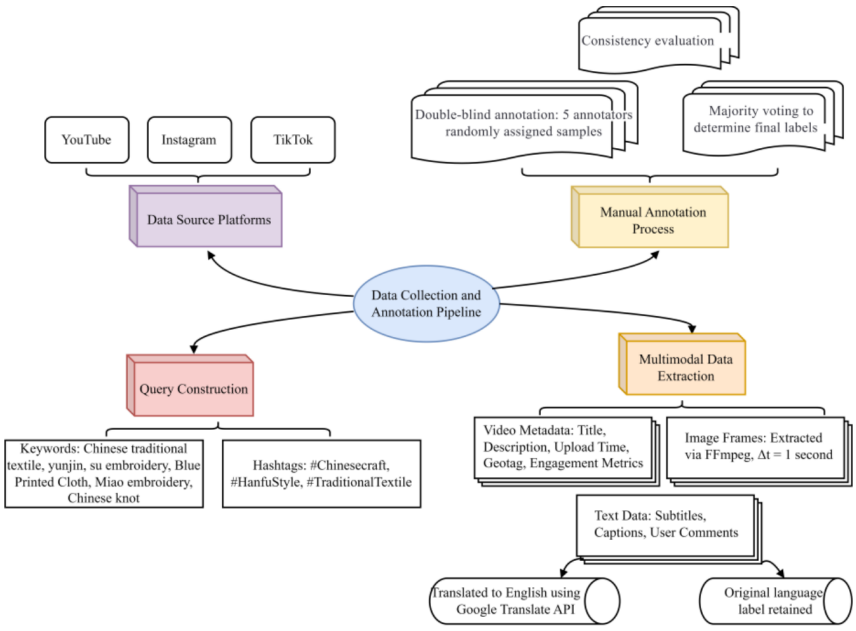

FIGURE 1. Data collection flow chart

$$g_j = \sigma\left(W_g\left[h_j^{(l_0)}; c_j\right]\right), \quad \bar{h}_j^{(l_0)} = g_j \odot h_j^{(l_0)} + (1-g_j)\odot c_j \tag{8}$$

For audio input X_a, a pre-trained Wav2Vec 2.0 model is used to extract the speech feature sequence $H_a^{(0)} \in \mathbb{R}^{T\times d}$. The same cultural entity fusion mechanism is introduced after the second Transformer layer to generate an enhanced audio representation H_a. Finally, the trimodal encoder outputs the culturally enhanced sequenced feature matrices $H_v \in \mathbb{R}^{N\times d}$, $H_t \in \mathbb{R}^{M\times d}$, $H_a \in \mathbb{R}^{T\times d}$.

D. IMPLEMENTATION OF THE DYNAMIC SEMANTIC ALIGNMENT MECHANISM

This paper designs a gated cross-modal attention mechanism to achieve culturally sensitive dynamic semantic alignment [21,22]. With text as the query, image as the key and value, cross-modal attention calculation is performed:

$$A_{tv} = \text{Softmax}\left(\frac{Q_t K_v^\top}{\sqrt{d}}\right), \quad Q_t = H_t W_q^v, \quad K_v = H_v W_k^v \tag{9}$$

In formula (9), W_q^v, W_k^v are projection matrices, which output the attention-weighted image-text fusion representation:

$$Z_{tv} = A_{tv} V_v, \quad V_v = H_v W_v^v \tag{10}$$

To control the alignment strength, a gating mechanism gtv∈[0, 1]M×N is introduced, whose value is determined by the semantic consistency between text and image modalities. Define the cross-modal similarity matrix:

$$S_{tv} = H_t H_v^\top \otimes \left(\|H_t\|\|H_v\|^\top\right) \tag{11}$$

Take the mean of each row as the global similarity:

$$\bar{s}_{tv} = \frac{1}{N}\sum_{j=1}^{N} S_{tv}(i,j), \quad i = 1,\ldots,M \tag{12}$$

The gate weights are generated by the Sigmoid function:

$$g_{tv} = \sigma\left(W_g[\bar{s}_{tv}; e_{\text{ctx}}]\right), \quad W_g \in \mathbb{R}^{1\times(1+d)} \tag{13}$$

In formula (13), e_{ctx} is the cultural context vector of the current sample. First, the similarity between the multimodal fusion representation of the current sample and the embeddings of all cultural entities in the knowledge graph is calculated. The top-K entities with the highest similarity are selected to form the core cultural entity set. Subsequently, an attention-based weighted aggregation mechanism is used to generate a more representative cultural context vector. The final fusion representation is:

$$\widetilde{Z}_{tv} = g_{tv} \odot Z_{tv} + (1 - g_{tv}) \odot H_t \tag{14}$$

Similarly, the paper constructs the text-audio and image-audio alignment modules, with all gating parameters sharing the same W_g for consistency. Finally, the paper outputs three sets of aligned representations: $\widetilde{Z}_{tv}$, $\widetilde{Z}_{ta}$, $\widetilde{Z}_{va}$.

E. CROSS-MODAL FUSION AND PROPAGATION REPRESENTATION GENERATION

This paper designs a hierarchical cross-modal fusion mechanism that further integrates the three sets of aligned representations: image-text, text-audio, and image-audio, based on dynamic alignment, to generate a unified propagation semantic vector [23]. The first layer of intermodal cross-fusion introduces a cross- attention module, taking text as the dominant modality and fusing image and audio

information separately. First, $\widetilde{Z}_{tv}$ and $\widetilde{Z}_{ta}$ are projected into a shared semantic space:

$$F_{tv} = \widetilde{Z}_{tv} W_{tv}, \quad F_{ta} = \widetilde{Z}_{ta} W_{ta}, \quad W_{tv}, W_{ta} \in \mathbb{R}^{d \times d_h} \tag{15}$$

Concatenate along the sequence dimension and input into the cross-attention layer:

$$C_1 = \mathrm{CrossAttn}(F_{tv}, F_{ta}) = \mathrm{Softmax}\left(\frac{Q_1 K_2^\top}{\sqrt{d_h}}\right) V_2 \tag{16}$$

The C_1 is fused with the image-audio alignment representation $\widetilde{Z}_{va}$ for a second time. Average pooling compression is used:

$$\bar{Z}_{va} = \frac{1}{N} \sum_{i=1}^{N} \widetilde{Z}_{va,i} \in \mathbb{R}^d \tag{17}$$

Perform final crisscross attention:

$$C_2 = \mathrm{CrossAttn}(C_1, B_{va}) \tag{18}$$

The second layer integrates global self-attention. C_2 is input into the Multi-head Self-Attention module (MHSA) to achieve cross-position semantic aggregation:

$$\mathrm{MHSA}(C_2) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h) W_o \tag{19}$$

Each attention head is defined as:

$$\begin{aligned} \mathrm{head}_i &= \mathrm{Softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_h}}\right) V_i, \\ Q_i &= C_2 W_i^Q, \quad K_i = C_2 W_i^K, \\ V_i &= C_2 W_i^V \end{aligned} \tag{20}$$

This paper introduces a parallel and specialized cultural attention head based on the standard multi-head self-attention (MHSA), aiming to calibrate and enhance global semantics at the cultural level from the perspective of knowledge graph. Assuming that the K_c core entities most relevant to the current sample in the knowledge graph are embedded as $G_c = \{g_k\}_{k=1}^{K_c} \subset \mathbb{R}^d$, calculate the attention distribution of the fusion representation and cultural entities:

$$\begin{aligned} \alpha_{ik} &= \frac{\exp(c_i^\top g_k / \tau)}{\sum_{k'=1}^{K_c} \exp(c_i^\top g_{k'} / \tau)}, \\ i &= 1, \ldots, M, \\ k &= 1, \ldots, K_c \end{aligned} \tag{21}$$

Weighted generation of cultural enhancement vectors:

$$e_i^c = \sum_{k=1}^{K_c} \alpha_{ik} g_k \tag{22}$$

The final propagated representation is fused through residual connections:

$$e_i = W_r[c_i; e_i^c] + b_r, \quad W_r \in \mathbb{R}^{d \times 2d}, \quad b_r \in \mathbb{R}^d \tag{23}$$

Output a unified propagation representation vector $E = [e_1, e_2, \ldots, e_M]^\top \in \mathbb{R}^{M \times d}$. Figure 2 illustrates the hierarchical process for cross-modal fusion and propagation representation generation: The first layer gradually fuses the text-image and text-audio aligned representations through cross-attention. It then introduces the image-audio representation after average pooling for secondary attention fusion, achieving progressive aggregation of multimodal information. The second layer uses MHSA to model the global context of the fusion results, capturing long-range dependencies between cross-modal elements. A cultural attention head is introduced to calculate the attention distribution using core entities in the knowledge graph and generate a culturally enhanced vector. Finally, a residual connection is used to fuse the global representation with the culturally enhanced vector, outputting a unified propagation representation vector.

F. CONVERSATION EFFECTIVENESS PREDICTION MODEL CONSTRUCTION

This paper constructs a multi-task prediction model based on the fused communication representation, jointly modeling four indicators: communication breadth, cognitive improvement value, emotional tendency, and cultural retention level. Assume that the global communication representation generates sentence vectors through average pooling:

$$e_{\mathrm{global}} = \frac{1}{M} \sum_{i=1}^{M} e_i \in \mathbb{R}^d \tag{24}$$

Predict two continuous variables: spread yreach and cognitive improvement ycog . Use a dual-output regression head:

$$\hat{y}_{\mathrm{reg}} = W_{\mathrm{reg}} e_{\mathrm{global}} + b_{\mathrm{reg}}, \quad \hat{y}_{\mathrm{reg}} = [\hat{y}_{\mathrm{reach}}, \hat{y}_{\mathrm{cog}}]^\top \in \mathbb{R}^2 \tag{25}$$

The regression loss uses smooth L1 loss:

$$\begin{aligned} L_{\mathrm{reg}} &= \sum_{j=1}^{2} l_\delta(y_j, \hat{y}_j), \\ l_\delta(y, \hat{y}) &= \begin{cases} \frac{1}{2}(y - \hat{y})^2, & |y - \hat{y}| \le \delta, \\ \delta |y - \hat{y}| - \frac{1}{2}\delta^2, & \text{otherwise.} \end{cases} \end{aligned} \tag{26}$$

Predict two discrete variables: sentiment y_{emo} and cultural retention y_{cul}. Suppose their true labels are one-hot encoded as:

$$y_{\mathrm{emo}} = [y_{\mathrm{emo}}^{(1)}, y_{\mathrm{emo}}^{(2)}, y_{\mathrm{emo}}^{(3)}]^\top, \quad y_{\mathrm{cul}} = [y_{\mathrm{cul}}^{(1)}, y_{\mathrm{cul}}^{(2)}, y_{\mathrm{cul}}^{(3)}]^\top \tag{27}$$

Construct a shared hidden layer:

$$h_{\mathrm{cls}} = \mathrm{ReLU}(W_1 e_{\mathrm{global}} + b_1) \tag{28}$$

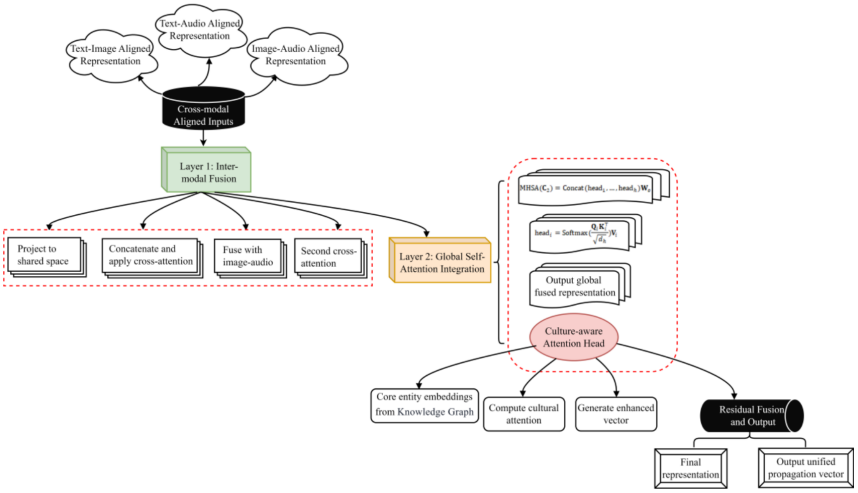

FIGURE 2. Fusion Module Diagram

Connect two classification heads separately:

$$\hat{p}_{\text{emo}} = \text{Softmax}(W_e h_{\text{cls}} + b_e), \quad \hat{p}_{\text{cul}} = \text{Softmax}(W_c h_{\text{cls}} + b_c) \tag{29}$$

The classification loss uses cross entropy:

$$L_{\text{emo}} = -\sum_{c=1}^{3} y_{\text{emo}}^{(c)} \log \hat{p}_{\text{emo}}^{(c)}, \quad L_{\text{cul}} = -\sum_{c=1}^{3} y_{\text{cul}}^{(c)} \log \hat{p}_{\text{cul}}^{(c)} \tag{30}$$

Use a weighted multi-task learning strategy to balance regression and classification losses:

$$L_{\text{total}} = \lambda_1 L_{\text{reg}} + \lambda_2 L_{\text{cls}} \tag{31}$$

The weights are dynamically adjusted through gradient normalization to avoid gradient scale imbalance between tasks. The model parameters are updated through the optimizer:

$$\theta_{t+1} = \theta_t - \eta \left(\beta_1 m_t + (1-\beta_1) g_t\right) - \eta \lambda_w \theta_t \tag{32}$$

The final communication effectiveness is output in a structured form:

$$\text{Output} = (\hat{y}_{\text{reach}}, \hat{y}_{\text{cog}}, s_{\text{emo}}, c_{\text{cul}}) \tag{33}$$

$\hat{y}_{\text{reach}}$, $\hat{y}_{\text{cog}}$ are the regression task outputs; s_{emo} is the probability of positive sentiment; $c_{\text{cul}} = \arg\max(P_{\text{cul}})$. This quantitative indicator group provides a calculable basis for subsequent communication strategy optimization. The model parameter settings are shown in Table 1.

TABLE 1. Model Parameters

Parameter	Value/Setting
Learning Rate (η)	2×10^{-5}
Momentum Parameters (β_1)	0.9
Momentum Parameters (β_2)	0.999
Weight Decay	0.01
Batch Size (B)	32
Training Epochs (T)	20
Early Stopping Strategy	Monitored validation loss with a patience of 5 epochs
Random Seed	42
Hardware Platform	NVIDIA A100 GPU (Graphics Processing Unit)

III. EXPERIMENT AND VALIDATION

A. EXPERIMENTAL DESIGN

This experiment used YouTube, Instagram, and TikTok, three major international social media platforms, as data sources to collect videos, images, text descriptions, and user comments related to traditional Chinese textile culture. The data covered six representative cultural themes: Yunjin, Suzhou embroidery, blue printed fabric, Miao embroidery, Shujin, and buttonholes. The experimental setup is shown in Table 2.

TABLE 2. Experimental Setup

Component	Configuration
Data Sources	YouTube, Instagram, TikTok
Cultural Categories	Yunjin Brocade, Su Embroidery, Blue Printed Cloth, Miao Embroidery, Shu Brocade, Chinese Knot Button
Content Types	Video, Image, Caption, User Comment
Original Sample Size	15,000
Sampling Strategy	Stratified sampling by platform, language, and cultural category
Dataset Split	Training: 70%, Validation: 15%, Test: 15%
Category Distribution	English, Japanese, Spanish, French, German, Italian
Annotation	Manual labeling of cultural symbols, sentiment polarity, and understanding level by 5 experts
Consistency Metric	Krippendorff's $\alpha > 0.85$

B. CULTURAL SYMBOL RECOGNITION ACCURACY EVALUATION

To assess the model's ability to recognize traditional Chinese textile cultural symbols, the paper used precision, recall, and F1 score as evaluation metrics. The calculation method is:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{34}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{35}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{36}$$

TP, FP, and FN are the number of correctly identified, falsely detected, and missed samples, respectively. Each baseline model is fine-tuned using the same training dataset as the proposed method and uses its standard, publicly available pre-trained weights. The hyperparameters of all models remain consistent with those in Table 1 and training is performed on the same hardware platform. Figure 3 compares the recognition performance of CLIP-Adapter, VisualBERT, BLIP-2 (Bootstrapped Language- Image Pre-training 2), and Proposed Method on six categories of traditional Chinese textile cultural symbols. Proposed Method significantly outperforms other methods in all categories, achieving an F1-score of 0.90 for Suzhou embroidery and 0.87 for Yunjin brocade. Even for the complex and densely detailed knot button patterns, it maintains a high F1 score of 0.81. Other methods generally exhibit lower F1 scores for similar tasks. Proposed Method achieves an F1 score of 0.83 on blue print cloth, 0.12 higher than BLIP-2's 0.71, demonstrating its greater robustness in low-contrast, repetitive pattern scenarios. Proposed Method achieves an overall average F1 score of 0.85, significantly exceeding CLIP-Adapter (0.73), VisualBERT (0.70), and BLIP-2 (0.74). This demonstrates Proposed Method's systematic advantage in fine-grained cultural symbol recognition and demonstrates the critical role of knowledge injection in improving the model's cultural understanding capabilities.

C. CROSS-MODAL SEMANTIC ALIGNMENT PERFORMANCE EVALUATION

To evaluate the model's ability to align semantics between image and text modalities, a bidirectional cross- modal retrieval task was designed. Recall@K (R@K) was used as the core evaluation metric, measuring the proportion of positive samples in the top K retrieval results. Figure 4 shows the performance differences among four methods on the cross-modal semantic alignment task. The vertical axis represents the four metrics: text-to-image R@1, text-to-image R@5, image-to-text R@1, and image-to-text R@5. The corresponding values for Proposed Method are 78.3%, 91.6%, 76.8%, and 90.2%, respectively, with an average of 84.2%, significantly higher than the other methods. In terms of text-to-image R@1, Proposed Method outperforms BLIP-2 (68.1%) by 10.2%. In terms of image-to-text R@5, proposed method achieves 90.2%. VisualBERT's performance is only 74.0%, a 16.2% gap. This demonstrates that proposed approach not only excels in high-precision matching (R@1), but also demonstrates greater robustness in retrieval coverage (R@5), comprehensively improving bidirectional alignment capabilities between image and text modalities. By integrating knowledge graph information during the feature extraction phase, it enhances the cultural semantic sensitivity of intra-modal representations, which is the foundation for achieving high-precision cross-modal semantic alignment.

D. ASSESSMENT OF ENHANCED CULTURAL AWARENESS AMONG INTERNATIONAL AUDIENCES

To quantitatively evaluate the impact of the model-optimized communication strategy on audience understanding, a controlled user study was conducted. A total of 450 volunteers were recruited from online participant panels based on their native language to form six distinct groups: English, French, Japanese, Spanish, German, and Italian speakers (75 participants per group). This study was approved by the Ethics Committee of Beijing Institute of Graphic Communication, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form. To ensure a baseline level of general cultural interest, all participants were screened to have no prior professional expertise in Chinese culture or textile arts. The assessment used a pre-test/post-test design with a single- blind protocol, where the participants were unaware of the specific research hypothesis being tested. Before watching the video, participants completed a pre-test questionnaire designed by cultural heritage experts to assess their baseline knowledge of traditional Chinese textile symbols. The questionnaire consisted of 10 multiple-choice questions covering core concepts such as craftsmanship names, pattern recognition, and symbolic meanings. After completing the pre-test, each participant watched a 5-minute video on traditional Chinese textile culture. This video's content and multimodal presentation were optimized using the analysis method proposed in this paper. The cognitive improvement rate was calculated as:

$$\Delta C = \frac{1}{n} \sum_{j=1}^{n} \frac{S_j^{\text{post}} - S_j^{\text{pre}}}{S_j^{\text{pre}}}, \quad S_j^{\text{pre}}, S_j^{\text{post}} \in [0, 10] \tag{37}$$

Figure 5 shows the cognitive changes in six language groups before and after watching the video on traditional Chinese textile culture optimized using this method. All groups achieved significantly higher scores after the post-test than before, with an overall average improvement of 81.4%. The French group saw a 92.3% improvement, from 3.9 to 7.5, the highest among all groups. The Spanish group saw a 90.0% improvement, from 4.0 to 7.6. The English and German groups also achieved improvements

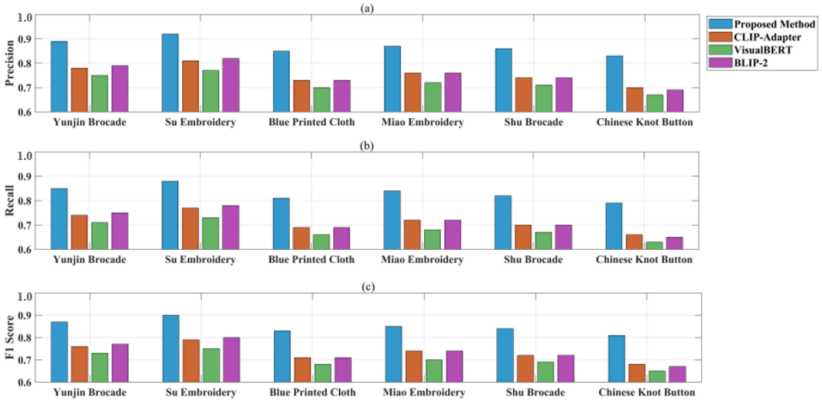

FIGURE 3. Cultural Symbol Recognition Accuracy: (a) Precision; (b) Recall; (c) F1-score

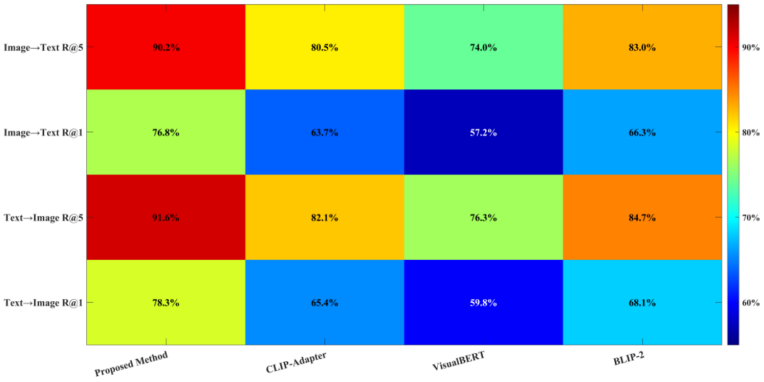

FIGURE 4. Cross-modal Semantic Alignment Performance Evaluation

of 85.7% and 86.5%, respectively. In contrast, although the Japanese group scored the highest in the post-test (8.6 points), their improvement rate was only 68.6% due to their higher pre-test base (5.1 points). The Italian group had the lowest improvement rate (65.1%), reflecting the differences in cognitive thresholds and acceptance efficiency among audiences from different language and cultural backgrounds.

E. CONSISTENCY ASSESSMENT OF COMMUNICATION SENTIMENT TENDENCY

To assess the consistency between the model's sentiment classification results and manual annotations, Cohen's Kappa coefficient was used to measure consistency:

$$\kappa = 1 - \frac{1 - p_o}{1 - p_e},$$
$$p_o = \frac{1}{N} \sum_{i=1}^{3} n_{ii}, \tag{38}$$
$$p_e = \sum_{i=1}^{3} \left(\frac{n_{i+}}{N} \cdot \frac{n_{+i}}{N} \right)$$

p_o is the observed agreement rate, p_e is the expected agreement rate, and N is the total number of samples. $\kappa > 0.8$ is considered extremely strong agreement. Table 3 shows the performance of Proposed Method in evaluating the consistency of communication sentiment with CLIP-Adapter, VisualBERT, and BLIP-2. The paper's method significantly outperforms the other methods with an overall accuracy of 91.6% and a Kappa value of 0.882. BLIP-2 achieves an accuracy of 82.9% and a Kappa value of 0.748, while CLIP-Adapter and VisualBERT both achieve accuracy values lower than 81% and 0.72, respectively. A Kappa value greater than 0.8 indicates a high level of agreement between the predictions of Proposed Method and manual annotations, demonstrating the high reliability of Proposed Method in assessing the sentiment of international audiences across languages and cultures. It excels in identifying implicit negative emotions or neutral attitudes driven by cultural differences. The gated cross- modal attention mechanism adaptively adjusts the alignment strength, enabling the model to focus more on culturally relevant modal information, improving the consistency of sentiment assessments.

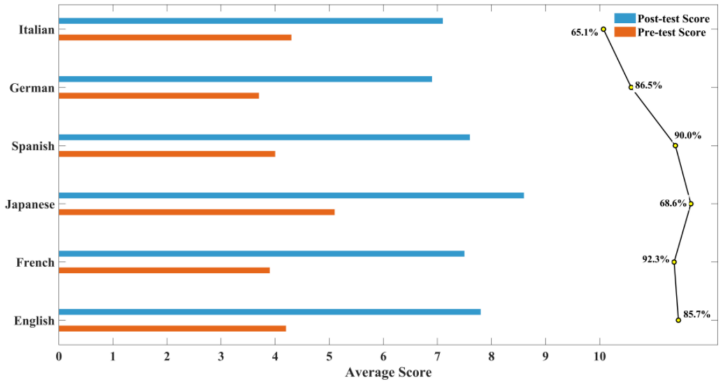

FIGURE 5. Assessment of Cultural Awareness Improvement among International Audiences

TABLE 3. Consistency Assessment of Communication Sentiment Tendency

Model	Overall Accuracy	Cohen's Kappa (κ)
Proposed Method	91.60%	0.882
CLIP-Adapter	80.60%	0.714
VisualBERT	77.90%	0.683
BLIP-2	82.90%	0.748

F. COMPREHENSIVE EVALUATION OF MULTI-PLATFORM COMMUNICATION EFFECTIVENESS

To comprehensively quantify the impact of Chinese traditional textile culture on international social media platforms, the Comprehensive Effectiveness Index (CEI) was constructed to measure the overall communication effect. For each sample, the following normalized metrics were calculated: interaction rate, cognitive improvement, positive sentiment, and cultural retention. To ensure a balanced and transparent evaluation, the CEI was computed as the arithmetic mean of these four equally weighted dimensions. Figure 6 compares the CEI values of Proposed Method, CLIP-Adapter, VisualBERT, and BLIP-2 across three major international platforms: YouTube, Instagram, and TikTok, in different language contexts. Proposed Method achieves the highest CEI across all platforms and languages, with an average of 0.725. On YouTube, among French users, Proposed Method achieves a CEI of 0.768, significantly higher than CLIP-Adapter (0.701), VisualBERT (0.675), and BLIP-2 (0.715). Among Italian-speaking users, Proposed Method achieved a CEI of 0.678, outperforming CLIP-Adapter (0.621), VisualBERT (0.588), and BLIP-2 (0.635). This advantage persisted across all language regions and across all three platforms. For Spanish-speaking users on Instagram, Proposed Method achieved a CEI of 0.746, surpassing all other methods. On TikTok, for English-speaking users, Proposed Method achieved a CEI of 0.701, also outperforming all other methods. These data strongly demonstrate that Proposed Method achieves superior overall communication effectiveness in diverse international cultural and linguistic contexts.

G. KNOWLEDGE GRAPH ABLATION EXPERIMENT

To clearly evaluate the independent contribution of knowledge graph (KG) to model performance, this study designs key ablation experiments. As shown in Table 4, ablation experiments clearly quantify the contribution of the knowledge graph. Compared to the w/o-KG model without knowledge graph integration, the proposed complete model achieves significant improvements in cultural symbol recognition F1 score, cross-modal retrieval R@1, and comprehensive effectiveness index (CEI). The CEI increases from 0.601 to 0.725, strongly demonstrating that the introduction of the knowledge graph is the core driver of the model's performance improvement. Furthermore, compared to the w/KG-Static model, which uses static knowledge fusion, the complete model maintains overall superior performance thanks to its dynamic and contextualized fusion mechanism. This demonstrates that the dynamic semantic alignment and cultural attention mechanisms designed in this paper can more effectively utilize the knowledge graph, achieving more accurate cultural perception and semantic alignment.

TABLE 4. Knowledge graph ablation experiment results

Model	F1	R@1 (Text→Img)	R@1 (Img→Text)	CEI
Proposed Method	0.85	78.30%	76.80%	0.725
w/KG-Static	0.80	73.10%	71.50%	0.692
w/o-KG	0.72	65.40%	63.70%	0.601

IV. CONCLUSION

This study addresses the problem of cross-modal semantic inconsistency in the international communication of traditional Chinese textile culture by proposing a Transformer-based culture perception multimodal fusion model. This method innovatively deeply couples a constructed textile cultural knowledge graph with the entire multimodal fusion process. It enhances the cultural semantic sensitivity of each modal representation through a culture perception encoder and designs a gated cross-modal attention mechanism for dynamic semantic alignment. Ultimately, it generates a unified communication representation vector for performance

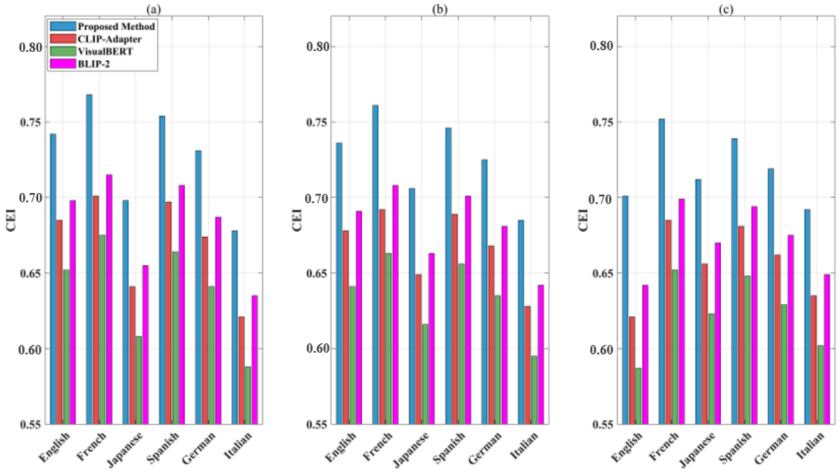

FIGURE 6. Comprehensive Score Evaluation of Multi-Platform Communication Effectiveness: (a) YouTube; (b) Instagram; (c) TikTok

prediction. Experiments show that this method significantly outperforms CLIP-Adapter, VisualBERT, and BLIP- 2 in cultural symbol recognition (average F1 of 0.85), cross-modal semantic alignment (average accuracy of 84.2%), and sentiment consistency accuracy (91.6%), effectively enhancing the dissemination and cultural fidelity of traditional Chinese textile culture on international social media platforms. This study focuses on traditional Chinese textile culture. The proposed framework has good transplantation potential and can be transferred to other fields that rely on deep cultural background knowledge, such as intangible cultural heritage handicrafts, traditional opera and other international communication tasks. For general multimodal tasks with vague knowledge systems or non-cultural fields, the specialized modules of this model may have limited advantages. Future work will focus on validating this transfer potential by applying the framework to additional cultural domains and expanding the evaluation to include a wider variety of content types and platforms.

AUTHOR CONTRIBUTIONS

All work in this study was independently completed by Haixia Xu.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was approved by the Ethics Committee of Beijing Institute of Graphic Communication, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] Z. Wang, R. Cui, T. Cong, and H. Liang, "Overseas Dissemination of Ancient Chinese Costume Culture from the Perspective of Cultural Confidence," FIBRES & TEXTILES in Eastern Europe, pp. 29, 3(147):111-116, 2021, doi: 10.5604/01.3001.0014.7796.

[2] C. Song, H. Zhao, A. Men, and X. Liang, "Design Expression of "Chinese-style" Costumes in the Context of Globalization," Fibres & Textiles in Eastern Europe, vol. 31, no. 2, pp. 82-91, 2023, doi: 10.2478/ftee-2023-0019.

[3] L. Yu, "Digital Sustainability of Intangible Cultural Heritage: The Example of the "Wu Leno" Weaving Technique in Suzhou, China," Sustainability, vol. 15, no. 12, pp. 9803, 2023, doi: 10.3390/su15129803.

[4] F. Chen, "Analysis of the Characteristics of Art Intangible Cultural Heritage in Cross-Cultural Communication," Art and Design Review, vol. 10, no. 3, pp. 389-396, 2022, doi: 10.4236/adr.2022.103030.

[5] S. Y. Pavlina, "A cross-cultural perspective on the comprehension of novel and conventional idiomatic expressions," Intercultural Pragmatics, vol. 21, no. 1, pp. 33-60, 2024, doi: 10.1515/ip-2024-0002.

[6] W. Xie, T. Zhang, M. Xiong, J. Zou, P. Zhu, and L. Yang, "Multimodal Semantic Communication: Research Progress, Key Challenges, and Future Trends," IEEE Communications Standards Magazine, vol. 9, no. 4, pp. 9-15, 2025, doi: 10.1109/MCOMSTD.2025.3579425.

[7] Y. Bai and S. Lei, "Cross-language dissemination of Chinese classical literature using multimodal deep learning and artificial intelligence," Scientific Reports, vol. 15, no. 1, Art. no. 21648, 2025, doi: 10.1038/s41598-025-05921-1.

[8] A. Li, X. Wei, D. Wu, and L. Zhou, "Cross-Modal Semantic Communications," IEEE Wireless Communications, vol. 29, no. 6, pp. 144-151, 2022, doi: 10.1109/MWC.008.2200180.

[9] F. Zhang and T. Krotova, "The Influence of Silk Road Culture on Modern Design: Artistic Features of Chinese Brocade Patterns," Art and Design, vol. 7, no. 1, pp. 56-67, 2024, doi: 10.30857/2617-0272.2024.1.5.

[10] J. Li, H. M. Adnan, and J. Gong, "Exploring Cultural Meaning Construction in Social Media: An Analysis of Liziqi's YouTube Channel," Journal of Intercultural Communication, vol. 23, no. 4, pp. 1-12, 2023, doi: 10.36923/jicc.v23i4.237.

[11] X. Gao and O. Yezhova, "Chinese Traditional Patterns and Totem Culture in Modern Clothing Design," Art and Design, vol. (2), pp. 20-30, 2023, doi: 10.30857/2617-0272.2023.2.2.

[12] Y. Zhang, "The Evolution and Contemporary Expression of Traditional Chinese Patterns: Water Patterns and Cloud Patterns as Examples," Mediterranean Archaeology and Archaeometry, vol. 24, no. 1, pp. 112-122, 2024, [Online]. Available: https://www.maajournal.com.

[13] X. Yuan and N. Chuprina, "Ecological Design Concept of Textile Products: A Study on Brocade Weaving Materials of the Chinese Ethnic Minority Yao," Art and Design, vol. (4), pp. 51-61, 2024, doi: 10.30857/2617-0272.2024.4.4.

[14] X. Gao, X. Wang, Z. Chen, W. Zhou, and S. C. H. Hoi, "Knowledge Enhanced Vision and Language Model for Multi-Modal Fake News De-

tection," IEEE Transactions on Multimedia, vol. 26, pp. 8312-8322, 2024, doi: 10.1109/TMM.2023.3330296.

[15] Y. Xia, X. Yao, J. Wang, and M. Hu, "Leveraging knowledge graphs for renaissance costume matching and cultural transmission," Npj Heritage Science, vol. 13, no. 1, pp. 219, 2025, doi: 10.1038/s40494-025-01742-7.

[16] S. Zhao and S. Zhu, "A Design Method for Automotive Interior Textures Based on Cultural Semantic Modeling and Generative Design," Design and Art Studies, vol. 1, no. 1, pp. 2, 2025, doi: 10.63623/8g8twq14.

[17] E. Al-Buraihy and D. Wang, "Enhancing Cross-Lingual Image Description: A Multimodal Approach for Semantic Relevance and Stylistic Alignment," Computers, Materials & Continua, vol. 79, no. 3, pp. 3913-3938, 2024, doi: 10.32604/cmc.2024.048104.

[18] X. Bi and T. Zhang, "Analysis of the fusion of multimodal sentiment perception and physiological signals in Chinese-English cross-cultural communication: Transformer approach incorporating self-attention enhancement," PeerJ Computer Science, vol. 11, pp. e2890, 2025, doi: 10.7717/peerj-cs.2890.

[19] F. M. Watts and S. A. Finkenstaedt-Quinn, "The current state of methods for establishing reliability in qualitative chemistry education research articles," Chemistry Education Research and Practice, vol. 22, no. 3, pp. 565-578, 2021, doi: 10.1039/D1RP00007A.

[20] Q. Yu, X. Tao, and J. Wang, "Sustainable Design on Intangible Cultural Heritage: Miao Embroidery Pattern Generation and Application Based on Diffusion Models," Sustainability, vol. 17, no. 17, pp. 7657, 2025, doi: 10.3390/su17177657.

[21] J. Xiang, N. Zhang, and R. Pan, "Cross-modal fabric image-text retrieval based on convolutional neural network and TinyBERT," Multimedia Tools and Applications, vol. 83, no. 21, pp. 59725-59746, 2024, doi: 10.1007/s11042-023-17903-4.

[22] E. Tian, Z. Zhu, F. Liu, Li, and Z, "Multimodal Neural Machine Translation Based on Knowledge Distillation and Anti-Noise Interaction," Computers, Materials & Continua, vol. 83, no. 2, pp. 2305-2322, 2025, doi: 10.32604/cmc.2025.061145.

[23] N. Zhang, Y. Liu, Z. Li, J. Xiang, and R. Pan, "Fabric image retrieval based on multi-modal feature fusion," Signal, Image and Video Processing, vol. 18, no. 3, pp. 2207-2217, 2024, doi: 10.1007/s11760-023-02889-1.