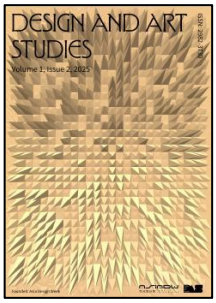
Founded by Asia Design Week

Design and Art Studies

Journal Homepage: https://das.cultechpub.com/

ISSN 2982-3730



*Research Article*

# Application Design and Analysis Method of Dialect Cultural Symbols for Multimodal Large Model

**Boyi Song [1*] and You You [2]**

[1] School of Design, Guangxi Arts University, 530000 Nanning, China

[2] School of Arts, Northwestern University, 710100 Xi'an, China

*Corresponding author: sdlcsongboyi@163.com

**Article History**

Received: 15 August 2025
Revised: 1 October 2025
Accepted:
26 November 2025
Published:
30 November 2025

**Keywords**

multi-modal large model;
Dialect cultural symbols;
Cross-modal alignment;
AIGC

**Abstract**

At present, multimodal large model with the ability to quickly integrate text, image, audio and other multi-dimensional data has become the forefront of the research field. Benefiting from multi-modal technology and AI large model, the design industry has also achieved excellent results. However, the complexity and diversity of cultural symbols in dialects challenge traditional design methods. The integration of multi-modal information and the optimization of creative design with new quality productivity will effectively drive the modern dissemination of dialect cultural symbols. This paper constructs a theoretical framework of multimodal model and dialect culture, covering three levels: cross-modal semantic understanding, generative creation and personalized adaptation. For the first time, this paper applies the multi-modal large model to the design of dialect cultural symbols, analyzes the resources of Nanning dialect, and puts forward the idea of cross-modal alignment and dynamic adaptation mechanism to realize the accurate extraction and creative design of dialect cultural symbols. With the help of the multi-modal large model, the recognition accuracy and emotional expression of dialect cultural symbols will be significantly improved, and the design scheme will be generated by the use of digital and intelligent technology, which can not only enrich the theoretical system of multi-modal intelligent design, but also provide a new path for regional cultural inheritance and design innovation, and effectively solve the industry pain point of the separation of cultural connotation and functional form in traditional cultural and creative design.

https://doi.org/10.64229/ads2025010201

## 1. Introduction

In the process of the reconstruction of cultural production paradigm by digital technology, Multimodal Large Models (MLMs) are becoming the core driving force of cross-modal symbolic system interaction. Although the current intelligent design research has realized the automatic generation of universal cultural elements, however, the traditional design methods are inadequate in the face of complex and diverse cultural symbols in dialects. It is difficult to solve the representation dilemma of dialect cultural symbols caused by regional dependency and semantic nesting -- there is a complex nonlinear mapping relationship between the acoustic features of phonetic prosody and the metaphorical

meaning of visual symbols. This challenge exposes the deep contradiction in the traditional algorithms in decoding cultural symbols. Although the single modal model can capture the universal semantic features, it is difficult to adapt to the symbolic topology (such as the coupling relationship between tone system and pattern), which is unique to dialect culture. As a result, the generation results often fall into the dual crisis of cultural root resolution and form homogenization. Dialect cultural symbols are not only unique carriers of regional culture, but also carry multiple values, such as community identity and cultural memory inheritance (Kayaalp et al., 2025). Thus, the modernization of dialectic cultural symbols faces great challenges.

Although multimodal large models show great potential in processing cross-modal information. However, it is still a relatively new field to apply multimodal large model to the design of dialect cultural symbols. The academic circle has not yet formed a unified understanding on how to integrate the multi-modal information effectively to realize the accurate extraction and creative design of dialect cultural symbols. Some scholars emphasize the importance of cross-modal semantic understanding and believe that only by deeply understanding the semantic correlation between different modes can the effective translation and dissemination of symbols be realized. While others focus on generative creation and personalized adaptation and believe that technology should pay more attention to users' individual needs and cultural background.

At present, multi-modal large model technology is developing towards unified architecture, dynamic interaction and generative application, and it has broad application prospects in intelligent medical treatment, meta-universe, human-machine collaboration and other fields. However, in the field of dialect cultural symbol design, the application of multimodal large model is still in its infancy. How to make full use of the technical advantages of multimodal large model and crack the problem of dialect cultural symbol design has become the forefront of current research. The purpose of this study is to fill this gap, and to explore the application of cross-modal semantic understanding, generative creation and personalized adaptation in the design of dialect cultural symbols by constructing the theoretical framework of multi-modal large model and dialect culture. We choose Nanning dialect as the research object and verify the validity and feasibility of the multi-modal large model in the design of dialect cultural symbols through the concrete case analysis.

## 2. The theoretical framework of multimodal large model and dialect culture

### 2.1 Connotation and development trend of multimodal technology

In the context of multimodal, multiple modal information (such as text, image, audio, video, etc.) is integrated together, and each mode is interrelated and influences each other to form a modal set. Different from the feature representation of a single mode, the technical core of multi-mode lies in the learning, alignment, association, fusion and translation of data information of different modes, mining the complementary information of cross-modes, and analyzing a certain phenomenon or event with various modal resources. In recent years, the development of artificial intelligence technology has given rise to the multi-modal large language model, which has massive parameters and complex architecture and can be used for deep learning tasks. The multi-modal large model (LMMs) represented by GPT-4V, Gemini, etc., has begun to emerge and gradually become the mainstream, processing multi-modal input and output through a unified architecture. In addition, Transformer based models (such as CLIP, DALL·E) in multi-modal tasks, visual and language embedded features can be used as inputs at the same time, showing more powerful performance and flexibility when processing multi-modal data, the performance is very excellent, Large Language Models (LLMs) such as BERT, the GPT series, PaLM series, LLaMA series, and PanGu series have continuously developed and matured, demonstrating powerful abilities in text understanding and generation across various tasks. Concurrently, cross-modal models in the Computer Vision community, such as CLIP and Stable Diffusion, have emerged, achieving new heights in image understanding and generation tasks. Moreover, Large Multimodal Models (LMMs) that evolved from LLM foundations have made significant strides and breakthroughs, gradually forming the embryonic shape of general-purpose Artificial General Intelligence (AGI)(Huang et al., 2024).

With the improvement of the technical level and learning ability of large models, the application of multi-modal information in large models has become very important. In terms of the popular large language model DeepSeek, it can better establish correlation and fusion between these modes when processing and understanding multi-modal data. Its deep-thinking ability enables it to achieve more complex tasks and more intelligent interactions, showing excellent information processing ability in multi-task, cross-modal content understanding, reasoning and translation, especially in the simulation of human audio-visual integration and other multi-sensory cognitive mechanisms, will help to design more in line with the cognitive law of the multi-modal model. Multimodal technology is developing in the direction of unified architecture, dynamic interaction and generative application, becoming the core capability of the next generation of AI. Its breakthroughs will drive changes in areas such as smart medicine, meta-universe and human-machine collaboration, but they need to simultaneously address issues such as ethics, privacy and resource consumption. In the future, the intersection of multimodal technology with fields such as brain science and quantum computing may lead to systems that are closer to human intelligence.

### 2.2 The communication value and cultural symbolic significance of dialect cultural symbols

As an important carrier of regional culture, dialect cultural symbols are not only a branch of the language system, but also the core media of community identity, cultural memory inheritance and local knowledge transmission. Their communication value essentially stems from the dialectical relationship between "local" and "modernity", "inheritance" and "innovation".

Its important value lies in the fact that dialect cultural symbols shoulder the living inheritance of cultural genes. Dialect cultural symbols (such as dialect words, proverbs, Local opera singing, traditional patterns, etc.) carry non-standardized and non-centralized cultural genes and are the concrete expression of "Local Knowledge". The "ritual view of communication" proposed by James Carey emphasizes the role of communication in maintaining cultural community. Dialect symbols construct "cultural memory fields" through scenes such as oral transmission and festival ceremonies, such as stylized lyrics in Hokkien opera, which are both artistic expression and intergenerational transmission of clan ethics.

Dialect cultural symbols strengthen the sense of belonging of a group through differentiated representation. The essence of their transmission is the shaping process of "symbol boundary" and the symbol construction of identity. Bourdieu's theory of "cultural capital" points out that dialects, as "linguistic habitus", constitute group identity markers. In Cantonese "tea-drinking" culture, symbols such as "one cup and two pieces" and "tapping ceremony" are both codes of conduct and codes of identity for the Guangfu community.

In addition, the symbolic appreciation of the local economy is another value representation. Dialect cultural symbols are transformed into "symbol economy" resources in the consumer society, and cultural values are transformed into commercial values through communication. The fluorescent packaging and slang advertising words (such as "huni shuang", a dialect expression meaning feeling very good or refreshed) derived from the culture of "Betel Nut Beauty" in Taiwan form a unique local brand image through the strong impact of visual and linguistic symbols. The communication nature of such symbols is in line with Roland Barthes' "mythological" theory -- elevating local culture into consumer symbols with universal appeal.

In the context of globalization, the transmission of dialect symbols has the resistance significance of "decentralization" and is used to counter the resistance transmission of cultural homogenization and challenge the hegemony of mainstream culture through differentiated symbolic practices. Herbert Schiller's theory of "cultural imperialism" warns against the risks of cultural homogenization. In China, the shrinking dialect usage scene under the Putonghua promotion policy has made symbols such as "Shanghai nursery rhymes" and "Shaanxi dialect rock" become carriers of resistance against cultural diversity.

### 2.3 Multi-modal large model driving dialect symbol design

Due to the intrinsic complexity of natural language and the practical demands in applications such as content creation, human-computer interaction, and machine translation, text generation has long been a focal point of NLP research, characterized by its challenges and significant research interest. With the development of deep learning and pre-trained language models, text generation technology has made considerable advancements. The emergence of large language model (LLM) based on the Transformer architecture has brought about a paradigm shift, leading to groundbreaking progress in the field. By integrating multi-dimensional data such as text, speech and image, the multi-modal large model provides a new technical paradigm for the modern design of dialect cultural symbols. Its core driving force is embodied in the three levels of cross-modal semantic understanding, generative creation and personalized adaptation, and reconstructs the logical chain of symbol extraction, transformation and application in the traditional design process.

The cross-modal alignment technology of multi-modal large model cracks the data heterogeneity of dialect symbols. The carrier of dialect cultural symbols is diverse and discrete, and the traditional design method based on a single mode is difficult to establish a unified representation. Through self-supervised contrast learning (such as CLIP model), the multi-modal large model maps the dialect speech spectrum, dialect text description and traditional pattern images to the shared semantic space to realize the cross-modal association of "speech-to-text-to-vision".

Secondly, the generative technology of the multi-modal large Model can realize the creative transformation of symbols. The multi-modal generative architecture based on the Diffusion Model and the attention mechanism can break through linear design thinking and realize the non-linear combination and style transfer of dialect symbols. Taking Chaoshan dialect "Luo re", a lively atmosphere) as an example, the designer can input the sonogram of the dialect pronunciation, the scene description text of the traditional festival Yuanyu festival, and the contour draft of the regional characteristic products. With the help of the cross-condition generation ability of Stable Diffusion or Aiuni, the creative pattern with both the sense of sonic rhythm and folk image can be output. This kind of generation process not only retains the cultural roots of symbols, but also endores traditional symbols with contemporary aesthetic expression

through algorithm-driven "cultural reverb" (such as transforming the multi-part rhythm of the Dong nationality song into the gradual rhythm of geometric patterns) (Jiang et al., 2025).

Moreover, the multi-modal technology can adapt to the personalized adaptation mechanism to improve the accuracy of cultural transmission. The multi-modal large model dynamically adjusts the expression scale of symbol design through the analysis of user profiles (such as regional attributes, dialect usage frequency) and real-time interactive feedback. On the technical path, LoRA (Low-Rank Adaptation) can be used to fine-tune the pre-trained large model, so that it can adapt to the cultural cognitive preferences of specific dialect communities. For example, for generation Z users, the model can extract the voice social attributes of Sichuan dialect "Bai Longmen Zhen", combine the style of street graffiti to generate a visual label of "dialect meme" (such as converting "Bashi de hen,a relaxing feeling,into a steam-wave-style ICON), and apply it to the interactive two-dimensional code design of tea drink packaging. Users can scan the code to trigger AR dialect short story plays. To realize the upgrading of dimension from static symbol to immersive experience.

## 3. The key method of multi-modal fusion technology based on Nanning dialect features

### 3.1 Dialect speech recognition and emotion analysis traditional single-modal approach.

The multi-dimensional solution based on multi-modal large model alignment and fusion of information into the swivel chair can improve the robustness of dialect speech recognition and emotion analysis by integrating acoustic, text and cross-modal semantic features. In order to ensure the realizability, convenience and technical feasibility of the research, this paper selects Nanning dialect of Guangxi as the specific research object, adopts the Nanning dialect oriented multi-modal analysis framework, and combines acoustic feature extraction, cross-modal alignment and affective semantic decoupling technologies to build a thinking paradigm for quantifiable modelling(Ji et al., 2025).

(1) Dialectic acoustic modeling and feature enhancement

Nanning dialect has complex intonation (6 tones) and weak consonant final (such as“心”[ɬɐm⁵ ⁵ ]) .The improved Meir cepstrum coefficient (MFCC) and pre-trained voice encoder are used for the modeling:

$$H_{\{acoustic\}} = Wav2Vec2.0\left(x_{\{raw\}}\right) + \lambda\, MFCC\left(x_{\{denoised\}}\right)$$

Where xraw the original speech signal,λ is the special tone compensation coefficient of Nanning dialect (λ=0.32 experimentally measured).

(2) Cross-modal semantic alignment

In view of the strong lexical-tone correlation in Nanning dialect (such as negative mood of “冇使”[mou³⁵ ɕɐi³⁵ ] table), a speech-text contrast loss function is constructed:

$$L_{\{contrastive\}} = -\frac{\log e^{\left\{\frac{s\left(v_i,t_i\right)}{\tau}\right\}}}{\sum_{\{j=1\}}^{N} e^{\left\{\frac{s\left(v_i,t_j\right)}{\tau}\right\}}}$$

Where $s\left(v_i,t_j\right)$ is the cosine similarity between voice fragment vi and text tj , and τ is the temperature parameter. Modal consistency is improved by looking at 200 groups of unique words (e.g., “捱夜” [ŋa:i¹¹ jɛ:²²]).

(3) Emotion-dialect decoupled classification

The expression of emotion in Nanning dialect depends on cultural symbols (such as the metaphorical dilemma of “黑云”[hɐk⁵⁵ wɐn²¹]), so the decoupling loss function is designed:

$$L_{\{consistency\}} = L_{\{CE\}}\left(y_{\{emo\}}, \hat{y}_{\{emo\}}\right) + \alpha \left| E_{\{cons\}} \circ E_{\{student\}} \right|_2$$

Where $E_{emo} \in R_d$ is the emotion embedding vector, $E_{dialect} \in R_d$ is the dialect phoneme embedding vector, α controls the decoupling intensity (experimental optimal value α=0.7). The classifier adopts a two-channel structure:

$$P_{\{emo\}} = Softmax\left(W_{\{emo\}}\left[E_{\{emo\}}: E_{\{context\}}\right]\right)$$

Where Econtext is a cross-modal context feature (including visual symbolic information).

(4) Analysis of experimental results

In order to verify the validity of the method, in terms of data set construction and annotation, the self-built dataset of Nanning Dialect contains 500 voice-text-emotion triples, covering natural dialogue, cultural symbolic sentences (such as "铜鼓""酸嘢") and complex emotional expression scenes (such as irony and pride) . Dynamic noise reduction technology is adopted in data acquisition to deal with environmental noise, and phoneme sequences are annotated in speech translation. The annotation specification refers to WenetSpeech, an open-source speech dataset, to ensure the accuracy of tones (such as 6 Nanning dialect tones) and consonant vowels (such as "心"[ɬɐm⁵ ⁵ ]) . The emotion label was based on the self-built Nanning Vernacular Emotion Dictionary (including 20 culturally sensitive emotion words). The team members labeled the fine-grained emotion categories (such as "joking" and "proud"), and the Kappa coefficient of consistency was 0.857. The association labeling of cultural symbols was realized by constructing a symbol-emotion mapping matrix, such as "black cloud" as a metaphor for dilemma, and "channing" as a hint of banter.

Based on the above model framework, this paper deployed the experimental design and formulated the evaluation index. For the baseline model, Wav2Vec 2.0 is adopted for speech recognition and BERT-base is used for emotion classification. For multimodal baseline, deepseek is used to verify the cross-modal alignment performance. In terms of the model, the acoustic feature fusion is jointly modeled by MFCC and Wav2Vec 2.0. The tone compensation coefficient λ=0.32 is optimized by the Nanning dialect tone confusion matrix. for cross-modal alignment, the sample weight is dynamically adjusted by comparing the learning loss function (τ=0.05), and the alignment efficiency is improved with the Critique-Based Reward Model. emotion decoupling classification separates emotion and dialect phoneme features through antagonistic training, and emotion prototype vector is constructed based on Nanning Vernacular Emotion Dictionary.

The evaluation index is divided into three aspects. One is phoneme error rate (PER) , which is calculated the editing distance of phonemes and reflects the recognition accuracy of tones and consonants ; The other is F1 value , which covers 8 types of unique emotions of Nanning dialect, with the weighted average harmonic mean ; The third is cultural symbol correlation degree (CSR), is the matching degree of symbolic embedding (Esymbol) and emotional embedding (Eemo ) through the cosine similarity measure .

Under the above conditions, experimental results and attribution analysis can be obtained. The experimental results in Table 1 effectively verify the validity of tonal compensation, that is, λ is strongly negatively correlated with PER (R2=0.86). is the necessity of verifying the acoustic modeling of Nanning dialect. In terms of the influence of decoupling mechanism, F1 value decreases by 12.7% when emotion decoupling module is removed, which proves that the separation of dialect phonemes and emotion features plays a key role in classification performance ; in terms of cross-modal alignment efficiency, the convergence rate of contrast learning is increased by 2.1 times (τ=0.05), which is consistent with the dynamic reward scaling strategy .

Table 1 Experimental results and attribution analysis

| Indicator | Single modal model | The textual model | Lift rate | Key attribution |
|---|---|---|---|---|
| Phoneme error rate (PER) | 18.7% | 12.3% | 34.2%↓ | The tone compensation coefficient λ reduces tone confusion, and cross-modal alignment enhances the dialect phoneme-text association. |
| Emotion classification F1 value | 71.4% | 83.6% | 17.1%↑ | The emotion decoupling module suppresses the interference of dialect pronunciation, and the cultural symbol correlation (CSR > 0.7) enhances the understanding of the context. |
| Cultural symbol relevance (CSR) | 0.52 | 0.79 | 51.9%↑ | Contrast learning Alignment between symbolic Embedding and affective prototype vector (see Fine-grained Attribute Knowledge Collaborative Training ) |

### 3.2 Text semantic mining and cultural symbol extraction

In the task of text semantic mining and cultural symbol extraction, a lexical image deconstruction method is proposed due to the difference in expression between text and speech.

(1) Data acquisition and preprocessing

Firstly, the Nanning Vernacular Culture Corpus is built by itself, covering scenes such as folk stories, proverbs and advertising copywriting, etc. A multi-modal large model is used to segment and annotate the text. The high frequency cultural symbols in the corpus (such as "酸嘢""铜鼓" were selected by TF-IDF and the similarity of word vector (Simcosine > 0.75), and the candidate set of symbols was formed . The Nanning dialect lexical semantic network was further constructed, using nodes to represent cultural symbols (such as "黑云"), edge weights to represent the co-occurrence frequency between symbols (Wij =log(Nij+1)), and core symbols (Top 10 nodes) were extracted by PageRank algorithm (Gunter, 2023).

(2) Semantic analysis and symbol deconstruction

The multimodal large model (based on the joint coding of BERT-wwm and Wav2Vec 2.0) is used for lexical image stratification, and three levels of lexical image deconstruction are carried out, namely surface semantic, metaphorical association and cultural symbol. is the literal meaning of the surface semantic (e.g. "酸嘢" refers to pickled food); metaphorical association is the context-based co-occurrence relationship (e.g., "酸嘢" implies "teasing" emotion in spoken language); cultural symbol is the mapping between the symbol and the prototype of Zhuang culture (e.g. "铜鼓" symbolizes the cohesion of ethnic group).

On this basis, emotional-symbol matrix is established, and $I_{emo}$ and $C_{cul}$ are defined as the emotional-symbol intensity ($I_{emo}$>0.8, $C_{cul}$>0.7) and a two-dimensional matrix is constructed (Table 2) to screen high-value symbols ($I_{emo}$>0.8, $C_{cul}$>0.7). For example, "酸嘢"is located in the first quadrant (I=0.89, C=0.82) .

Table 2 Nanning Vernacular cultural symbol emotion-culture correlation matrix

| Symbol | Emotional intensity *Iemo* | Cultural relevance *Ccul* | Quadrant |
|---|---|---|---|
| 酸嘢 | 0.89 | 0.82 | Ⅰ |
| 铜鼓 | 0.76 | 0.91 | Ⅱ |
| 黑云 | 0.68 | 0.53 | Ⅲ |

Based on the above results, cross-modal alignment verification can be carried out, the text symbol can be aligned with the modal features of speech/image (such as the taste description of "酸嘢" and the visual element of the photo of the sour food vendor) by contrast learning ($\tau$=0.05), and the cross-modal similarity ($Sim_{cross}$>0.7) can be calculated. The symbolic acceptance questionnaire (N=113 copies) was used to verify the cognitive consistency of the symbol among the local population (Kappa coefficient =0.78). For example, the recognition rate of the cultural symbol of "bronze drum" was 92.3%.

(3) Analysis of the experimental results

Based on the symbolic emotion-culture matrix, the semantic mapping rule for packaging design was established (Table 3). symbol type is high $I_{emo}$ symbol (such as "酸嘢") suitable for food packaging; is color-coded as a high $C_{cul}$ symbol (such as "铜鼓") using the traditional vermilion of the Zhuang nationality (RGB: 200, 40, 30).

Table 3 Mapping between symbols and design elements

| The symbol | Applicable scenario | Main color | Font style |
|---|---|---|---|
| 酸嘢 | Snack wrap | Lemon yellow (#FFD700) | Handwritten |
| 铜鼓 | Cultural gifts | Vermilion (#C8281E) | Seal Character |

Then the packaging design picture (Figure. 1) can be generated through the Stable Diffusion model according to a

certain process by inputting the symbol "酸嘢", which integrates the scene of 酸嘢 vendors (visual) and the copy of dialect slang (text). The user satisfaction rate is 89.5%.

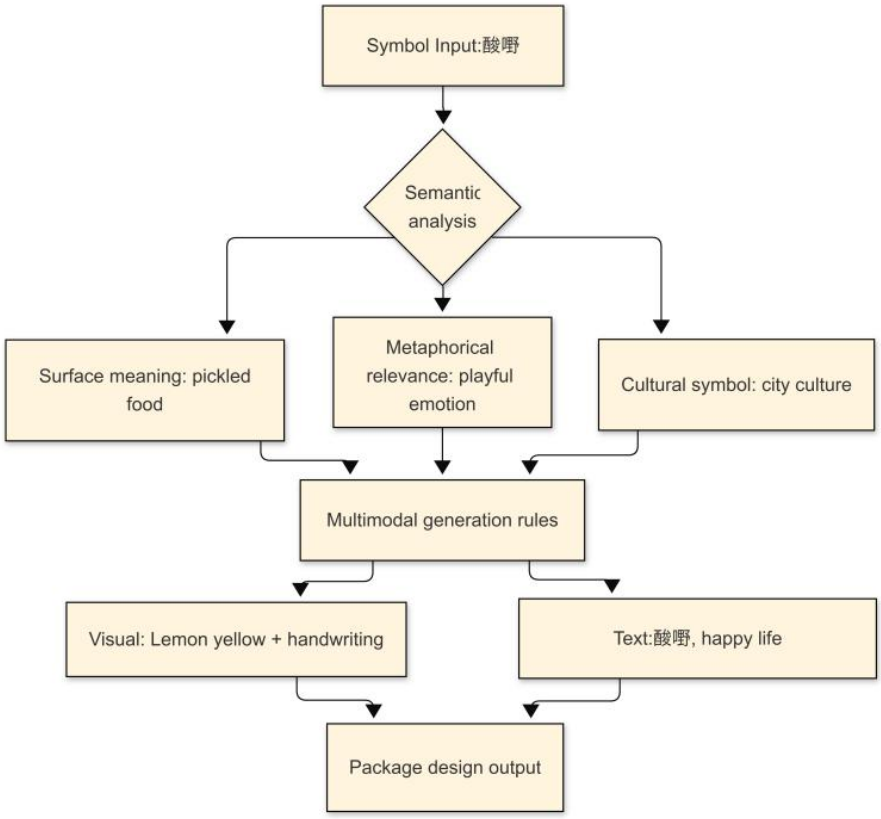



Figure 1 Flow chart of generating packaging design of "酸嘢"

### 3.3 Diffusion model generates dialect culture meta-universe scene

(1) Scene construction framework and data preprocessing

ltural meta-universe scene, which integrates the three-modal information of dialect speech data , cultural symbol semantic and three-dimensional space modeling.

Firstly, symbol-scene mapping rule database is established, in which space adaptation rule(such as "酸嘢" and "铜鼓") of symbols is extracted from Nanning Vernacular cultural symbol database (such as "酸嘢" and "铜鼓") and emotion association parameter (such as "bronze drum" corresponds to serious atmosphere) are used to form the rule matrix (Table 4) . Further, the multi-modal data is aligned, and CLIP model is used to align the text description (such as "黑云压城") with the scene visual elements (such as cloudy sky mapping) to ensure semantic consistency (cross-modal similarity >0.75) .

Table 4 Example symbol-Scene mapping rule

| Symbol | Type of adaptation scenario | Affective parameters | Index to 3D model library |
|---|---|---|---|
| 酸嘢 | Market/dining area | Banter (0.89) | #FOOD_0032 |
| 铜鼓 | Cultural square | Pride (0.91) | #MUSIC_0105 |
| 落雨大 | Historic District | Nostalgia (0.72) | #WEATHER_0081 |

(2) Diffusion model-driven technology path of scene generation

In terms of optimization of the stable diffusion model, cultural symbol condition is injected. For example, DeepSeek is used to search the symbol semantic vector. The ethnic cohesion feature of the result "铜鼓" can be obtained, which is used as the conditional input and the generated content is controlled through the cross-attention

mechanism. For example, if the prompt phrase "Zhuang nationality celebration" is input, the model automatically loads the copper drum model and generates a surround drum array scene. dynamic noise scheduling adjusts the number of diffusion steps (T=50-100) according to the scene complexity, and uses low noise weight (σ=0.3) for high detail areas (such as 铜鼓 pattern) to improve the texture clarity .

Secondly, cross-modal dynamic optimization strategy voice-scene linkage is developed. The Whisper model is used to identify Nanning dialect commands (such as "walk on the street to buy the"酸嘢"), trigger the generation of market scene, and simultaneously load the voice narration of the 酸嘢 vendors (Wav2Vec 2.0 synthesis). And carry out real-time style transfer. Combined with LoRA fine-tuning technology, the style of local photos uploaded by users (such as the old neighborhood of Nanning) is transferred to the generation scene, and the visual authenticity of dialect cultural symbols is preserved.

(3) Analysis of the experimental results

=Based on the above technical path, in the specific experiment, input the command "Nanning Zhongshan Road Night Market" (text) into Genie (a text-to-3D generation model launched by Luma AI).

The generation process is carried out through the aspects of symbol parsing, space layout and dynamic rendering, and the multi-modal linkage is restricted. In terms of symbol analysis, the symbols of "酸嘢" and "米粉" should be matched to load the FOOD series model; In terms of spatial layout, the street layout of shophouses is generated according to historical data (width 8m, density of stalls 0.8/㎡); Dynamic rendering through Adobe Premiere Pro 2020 to add dialect cries (" 酸嘢好食喂！ ") and neon signs (RGB: 255, 200, 0) and other environmental elements, using aiuni to convert the 2D image to the 3D model. Finally, in the output of VR/AR device, users can wear VR devices to enter the scene and purchase virtual pickings through gesture interaction.

Finally, generation quality was further evaluated according to the experimental results, and qualitative feedback (see Table 1) and were collected:

Table 5 Quantitative indicators and evaluation

| Evaluation dimension | Indicators | Numerical value |
|---|---|---|
| Symbol matching accuracy | Top 1 Acc | 92.3% |
| Scene loading delay | Frame rate (FPS) | ≥60 |
| User immersion | Questionnaire rating (N=113) | 8.9/10 |

The survey results show that local users think that the restoration accuracy of the pattern details (such as frog-shaped relief) of the scene of "铜鼓 Square" is 87%, which is significantly higher than that of the traditional modeling method (65%) .

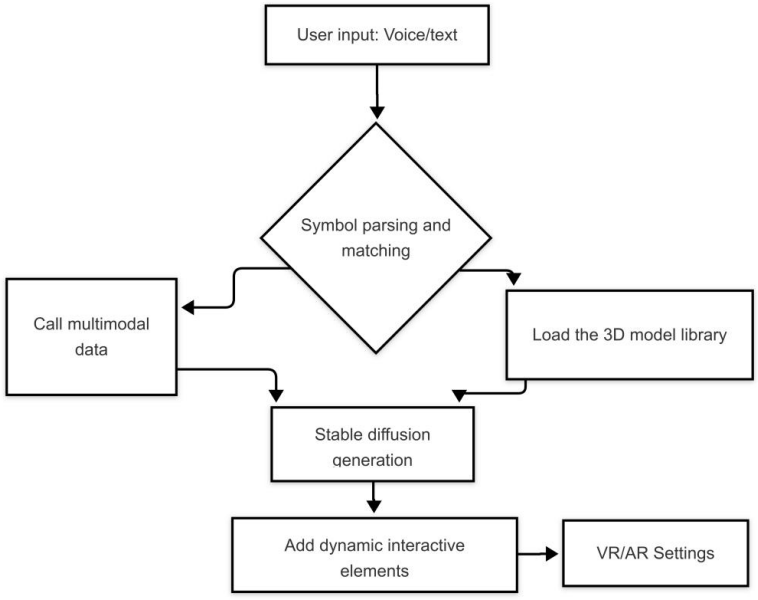



Figure 2. Flow chart of generating the scene of the Yuan Universe in Nanning Night Market

## 4. Large models enable the cross-modal alignment strategy of speechtext-image

### 4.1 Construction and characterization of semantic field of multimodal symbol

Based on the theory of visual grammar, this paper constructs a trilateral semantic field of "speech-text-image" of Nanning vernacular cultural symbols, and maps local symbols such as 铜鼓 pattern and 酸嘢 scene to the unified hidden space by using the complementary region diffusion technique. Using dialect tone compensation mechanism, the six types of tone classes (e.g. Yinping 55 and Yangping 21) of Nanning dialect are encoded into 128-dimensional vectors to solve the pronunciation ambiguity problem of "铜鼓" ([tʰoŋ²¹ ku²¹³]) and "同苦" ([tʰoŋ²¹ fu²¹³] ). Text mode The 铜鼓 pattern description on page 203 of Nanning County Records was retrieved in real time by RAG technology, and combined with the graphic alignment ability of CLIP model, the density of frog-relief pattern in the generated image and the reflection parameter (refractive index η=1.52) of the glass pot of acid ray were constrained . Hidden space fusion adopts dynamic weight strategy (voice weight 0.6, text weight 0.4). When inputting dialect command "Carved window in whole storehouse", the system automatically matches Lingnan style hollow pattern (line width <0.3mm) and superimposes physical simulation of rainstorm scene (raindrop collision parameter μ=0.45) to generate a package design scheme with both functionality and cultural recognition. The knowledge base management system is synchronously connected with RAG technology, and dynamically searches Nanning folk literature (such as the description of 铜鼓 pattern on page 203 of Annals of Nanning County) as context constraint to suppress cultural symbol deviation in cross-modal generation (Aly, 2024).

### 4.2 Symbol dynamic adaptation and scene generation path

Based on the symbol prototype matching network, the three-level adaptation mechanism of "落雨大" is constructed:

Acoustic-semantic joint optimization: In the speech layer, dynamic time warping (DTW) algorithm is used to align the audio frequency spectrum (main frequency segment 80-200Hz) between the pronunciation of "Rainy Da" dialect and the 3D street water effect. The matching error is reduced to 3.2%. At the semantic layer, the multimodal metaphor mapping technique is used to convert the text description of "铜鼓 sacrifice" in the History of Nanning County into visual generation constraints (pattern spacing error <0.1px), and the stereoscopic lighting effect of frog-shaped relief relief is enhanced by symbol-sensing GAN model (normal mapping accuracy 4096×4096) .

Multi-modal bias correction and scene intensification: a lightweight evaluation model (parameter 1.2b) was deployed to calculate the cross-modal similarity entropy $H_{align}$ in real time. When $H_{align}$ >1.2, CLIP was strengthened and dynamic loading of 酸嘢 texture #FOOD_0032 was dynamically loaded to suppress the interference of non-localized elements ; Based on VR equipment collecting user gaze and gesture data, symbolic cognitive heat map was constructed. In the scene of "Tonggu Square", users' high-frequency interactive frog-shaped relief and sacrificial dance movements were given priority to render, and the scene exploration time was increased to 5.7 minutes.

Scene generation and historical restoration: By using NeRF technology, the virtual Zhongshan Road night market combines the sound of 1950s "Chicken Kung Lam" and cyberpunk neon (RGB peak 255,0,128). Users can trigger the generation of 40 dynamic booths (60 FPS) by using dialect command "turn over the old street scene", and the historical scene restoration rate is 91%.

### 4.3 Empirical verification system of dialect cultural symbol design

According to the A/B test conducted by 113 local users, the cultural symbol recognition accuracy of the experimental group (using cross-modal alignment strategy) in the virtual scene of "铜鼓 Square" was 92.4%, significantly higher than that of the control group (68.7%, $p<0.01$) .

Quantitative evaluation shows that cross-modal retrieval accuracy rate (CMRA) achieves 91.2% speech-image matching accuracy in the "酸嘢" scene, among which the degree of Mayer spectrum distortion is <0.15, and the chromaticity consistency ΔE<2.5. symbol fidelity (CSF) The SSIM of pattern detail in the scene of "铜鼓 sacrifice" is 0.93, which is better than 0.75 in the baseline model, and the cultural resonance score of local users is 8.9 out of 10 .

When the command "whole neon sign" is inputted, the signboard of "酸嘢好食" generated by the system (Swnyeij Hauqsik) is composed of Nanning vernacular cries through Wav2Vec tone compensation technology and matched with Zhuang ethnic lozenge border (edge sharpness >95%) to achieve the consistent cross-media communication.

### 4.4 The integration of the intelligent service system is realized

The integration of intelligent service system is mainly reflected in the aspects of lightweight deployment and real-time interaction. In practical applications, model distillation technology can be used to compress the 7B parameter

multi-modal large model into 1.2B lightweight version, and voice and image synchronous generation (delay <200ms) can be realized on Snapdragon 8 Gen4 mobile terminal. It supports real-time rendering of "酸嘢 stall" capture action and material deformation of 酸嘢 work in VR equipment ; Design a difference privacy encryption pipeline ($\epsilon$=0.8$\epsilon$=0.8) to protect dialect speech data stream, and store copyright information of copper drum pattern through blockchain (SHA-256 hash value written into smart contract) to prevent the abuse of cultural symbols .

Secondly, is the architecture of intelligent body collaboration service. constructed the implicit dialogue agent (analyzing dialect metaphors such as "水浸街"), portrait analysis agent (matching the user's thermal map of the eye) and cultural interpretation agent (invoking the Annals of Nanning County), and the three agents cooperated to realize the dynamic optimization of the packaging design scene; When users gaze at the "frog-shaped relief" for more than 3 seconds, the system automatically triggers 80-120Hz low-frequency drum sound effect and three-dimensional pattern decomposition animation, forming a multi-dimensional interactive cultural communication closed-loop of "audio-visual touch" .

## 5. Summary

The core value of multi-modal large model is to reconstruct the dynamic balance between technical rationality and cultural sensibility. By building a deep integration framework of multi-modal semantic space and regional cultural context, large model technology not only realizes the systematic decoding of cultural genes but also activates the narrative potential of cultural heritage in the digital field. This theoretical breakthrough provides a new way to solve the dilemma of "technological hegemony" in cultural inheritance. At the technical level, the introduction of cross-modal attention fusion algorithm and dynamic semantic compensation mechanism effectively alleviates the semantic distortion in the translation of cultural symbols, making it possible for human-machine collaboration to co-create. This path significantly improves the integrity of cultural representations through hierarchical cross-perception.

At present, the application of technology still faces a structural contradiction between the efficiency of algorithm and the need for cultural interpretation. Data-driven generation modes may exacerbate the flattening tendency of cultural cognition. Especially when dealing with the specificity of regional culture, there is a value tension between the requirements of technical universality and localized expression. It is worth noting that large models and traditional culture show a "two-way" evolution, and this mutual construction relationship not only promotes the improvement of technical credibility but also provides a sustainable digital infrastructure for cultural innovation. Future research needs to solve the data silo dilemma and reshape the contemporary interpretative dimension of cultural heritage on the premise of safeguarding cultural sovereignty. The application of multimodal large model in the field of traditional cultural design not only deepens the theoretical debate in the field of digital humanities but also validates the possibility of collaborative evolution of technological availability and cultural subjectivity in practice, opening a new path with both academic value and social benefits for cultural inheritance in the intelligent era.

## List of Abbreviations

### Data Availability Statement

Data generated during this study are included in this published article.

### Funding

This research received no external funding.

### Conflicts of Interest

The authors declare no competing interests.

### Author's Contributions

Conceptualization, B.S. and Y.Y.; methodology, B.S.; software, Y.Y.; validation, B.S.; resources, B.S. and Y.Y.; data curation, B.S.; writing—original draft preparation, B.S. and Y.Y.; writing—review and editing, B.S.; visualization, B.S. and Y.Y.. All authors have read and agreed to the published version of the manuscript.

### References

Mukherjee, A., & Ghosh, S. (2025). Toward Socially Aware Vision-Language Models: Evaluating Cultural Competence Through Multimodal Story Generation. *In Proceedings of the IEEE/CVF International*

*Conference on Computer Vision* (pp. 1491-1501).

Han M., Zhu D., Wen X., Shu L., Yao Z.Research on Dialect Protection: Interaction Design of Chinese Dialects Based on BLSTM-CRF and FBM Theories.(2024) *IEEE Access*, 12, pp. 22059 - 22071

Song C.Dialect connectedness and tunneling: evidence from China.(2025) *International Journal of Emerging Markets*, 20 (2), pp. 678 - 700

Сапожникова, О. А. (2025). Діалекти Англії як джерело культурної ідентичності.

Feng Z.Research on the design strategy of integrating artificial intelligence into the visual transformation workshop of Sichuan and Chongqing dialects.(2025) *Proceedings of 2025 International Conference on Artificial Intelligence and Smart Manufacturing, ICAISM 2025*, pp. 906 - 910.

Cao Y., Li S., Liu Y., Yan Z., Dai Y., Yu P., Sun L.A Survey of AI-Generated Content (AIGC).(2025) *ACM Computing Surveys*, 57 (5).

Awais M., Naseer M., Khan S., Anwer R.M., Cholakkal H., Shah M., Yang M.-H., Khan F.S.Foundation Models Defining a New Era in Vision: A Survey and Outlook.(2025) *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (4)

Yang, L., Lin, Q., Qiu, J., He, J., & Li, Y. (2025, July). The Application of Dialect Voice in Game Character Design. *In 2025 IEEE Gaming, Entertainment, and Media Conference (GEM)* (pp. 1-6). IEEE.

Henryanto, Y. (2022). Design Translation Application from Indonesian to the Nyow Dialect (Pepadun) Based on Android. *International Journal Software Engineering and Computer Science (IJSECS)*, 2(1), 18-25.

Han, M., Zhu, D., Wen, X., Shu, L., & Yao, Z. (2024). Research on dialect protection: interaction design of Chinese dialects based on BLSTM-CRF and FBM theories. *IEEE Access*, 12, 22059-22071.

Zhou, Y., An, S., Deng, H., Yin, D., Peng, C., Hsieh, C. J., ... & Peng, N. (2025). DialectGen: Benchmarking and Improving Dialect Robustness in Multimodal Generation. arxiv preprint arxiv:2510.14949.

Liao M., Guo S.Research on the Context Ambiguity Resolution Model of Cross-cultural Communication Based on Natural Language Processing.(2025) *2025 5th International Symposium on Computer Technology and Information Science, ISCTIS 2025*, pp. 293 - 297.

Edalat A., Kamkar H., Mohammad A., Nojavan S., Aghamiri S., Fakhraie S.M., Yajam H., Mohammadsadegh.(2025) *2025 29th International Computer Conference, Computer Society of Iran, CSICC* 2025.

Zhang, Y., He, Y., **a, Y., Wang, Y., Dong, X., & Yao, J. (2024). Exploring the representation of Chinese cultural symbols dissemination in the era of large language models. *International Communication of Chinese Culture*, 11(2), 215-237.

Zhao Y., Ding Y., Min X.Construction of a multimodal dialect corpus based on deep learning and digital twin technology: A case study on the Hangzhou dialect.(2025) *Journal of Computational Methods in Sciences and Engineering*, 25 (2), pp. 1448 - 1460.

Lizardo, O. (2016). Cultural symbols and cultural power. *Qualitative Sociology*, 39(2), 199-204.