ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Mechanism Analysis of Multimodal Deep Learning in Enabling the Integrated Development of Agriculture, Commerce, Culture, and Tourism through the Creative Transformation and Innovative Development of Qilu Cu

Congcong Zhao[1] *and Guanghui Liu*[2,*]

[1] Grassroots Two Committees Education College, Shandong Open University, Jinan 250014, Shandong, China
[2] School of Sport and Leisure, Shandong Sport University, Jinan 250102, Shandong, China

SUMMARY: *To promote the creative transformation and innovative development of excellent traditional culture (the two–culture concept), to facilitate the deep integration of agriculture, commerce, tourism, and culture is a major path for regional high–quality development. However, the traditional methods have difficulty in dealing with cultural data that are heterogeneous, massive and multi–source industrial data such as semantic gaps and low integration efficiency. This study constructs an enabling mechanism analysis framework based on a multimodal deep learning model, it systematically studies how this model can digitally activate cultural resources through cross–modal semantics, knowledge graph construction, generative recommendation to achieve intelligent matching and collaborative value–added of agricultural, commercial and tourism elements. Empirical Analysis Research uses typical Cultural Eco–Protection Zone as Case Area, collect five type multimodal Datasets(Sample Size N=12847).Text,image,voice,video,Spacetime economy Data.By building a contrastive language image pre training(CLIP) + Transformer fused Model,the effects of the model on the enabling of three dimensions,cultural symbol extraction,consumer preference prediction, industrial chain optimization,were quantitatively analyzed. Results show: ① The accuracy of the recognition of cultural elements by the multimodal model reaches 89.3%, which increases by 22.6% compared with the traditional single–mode method. ②The cooperative index of Agriculture and Commerce driven by the model increases from 0.324 during the baseline period to 0.687 in the intervention period, which grows by 112%. ③The mechanism analysis shows a four–level transmission path according to attention:"cultural genes–product innovation–scene experience–value return". This paper is the first to explore the quantitative mechanism of enabling industrial integration between cultural two–culture through the perspective of deep learning, providing computable and transferable methodology support for regional smart tourism and rural revitalization.*

KEYWORDS: *multimodal deep learning; two–culture concept; agriculture, commerce and tourism integration; Mechanism analysis; cross–modal alignment; Enabling effect*

1 Introduction

Digital Economy and Cultural Heritage Protection are two strategies to be used for transforming

*zhaoccky@126.com
https://doi.org/10.65102/is2026810
Published: 30 April, 2026

traditional cultural resources from different regions into endogenous driving forces for the integration of agriculture, commerce and tourism, it has become a major issue that both academia and policy community have been paying attention[1]. Chinese excellent traditional culture has many spiritual symbols, aesthetic genes, and craft wisdom, but its value has long been constrained by the "resource dormancy" and "industrial islets" predicaments[2]. one aspect, there are numerous intangible heritages such as ancient books, ancient houses, etc., which do not have structured information or digitization level at all, thus cannot be recognized and assimilated by modern industrial chain effectively; on another hand, when developing cultural elements using Agriculture, Business, Tourism industries they usually fall into just picking up superficial symbols of cultures without doing any thorough investigation about their profound meanings and emotions behind those symbols leading to samey type products with no feelings attached hence quick results only last briefly[3].

Multimodal deep learning has made a great achievement in recent years. This is achieved by the field of computer vision and natural language processing as well as that of speech recognition[4]. Its basic ability – cross–modal semantic alignment, exactly addresses the most difficult problem for cultural innovation, and integrates agriculture, commerce, tourism, culture. Cultural resources exist in various heterogeneous modes such as images (Antiquities photos), texts (ancient records), audios (performances), videos (traditional craft processes), etc., while industrial data is expressed in structured form like transaction record, tourist trajectories, agricultural traceability information[5]. Traditional methods are unable to establish semantic mapping relationships between "paper–cut pattern image" and "agricultural product packaging design commercial asset", it also struggles with quantifying the causal effect of the "emotional polarity of local opera (audio)" on "tourist destination loyalty (consumer behavior)"[6].

Multimodal deep learning model (CLIP,ViLBERT,ImageBind, etc.), they can be trained through sharing a embedding space and mapping data from different modalities into a unified semantic representation vector to achieve the consistency of "image–text–audio–number"[7]. Furthermore, by combining attention mechanisms with graph neural networks, it is possible for the model to find out some latent relations among cultural elements and industrial ones in order to produce some brand–new results which carry on tradition but have their own identity as well as being very marketable. Therefore, this paper proposes that multimodal deep learning models form a four–level enabling mechanism "cultural decoding – creative generation – precise matching – value loop" by cross–modal feature extraction and association reasoning, greatly enhancing the system efficiency of agricultural–industrial–cultural tourism integration[8].

Current related research can be divided into three branches. The first one is digital protection of cultural resources, using computer vision to identify diseases in cultural relics or NLP to automatically index ancient books, it only reaches 'preservation', not reaching industrialization and application. 2. Second branch, exploring influencing factors of culture–tourism integration, through questionnaires, panel regression, we find out that there are positive effects on variables like cultural atmosphere, brand construction etc., but cannot explain which culture element brings about how much economic gain through what way[9]. The third part tries to apply machine learning to agricultural–business cooperation at the beginning, such as predicting the sale volume of agriculture with random forests, but the inputs are structured indicators and unstructured cultural semantic features are not included yet.

Summary: The theoretical gap is: There is no quantitative model that combines the "content production process" of cultural innovation and the "value realization process" of agriculture, commerce, tourism, and culture into one analytical framework. Methodological Gap : there is no research on whether to systematically evaluate the enabling efficiency and mechanism path of multimodal deep learning for cross–domain (culture → agriculture,commerce,tourism) knowledge transfer. To bridge these gaps. Research Objectives: (1) Constructing a theoretical

mechanism framework to realize agricultural, commercial, and cultural tourism integration via Multimodal Deep Learning[10]; (2) Quantify the enabling effect of the model in three aspects – Cultural Recognition, Preference Prediction, and Collaborative Optimization using Empirical Data; (3) Reveal the Multi–level Transmission Path and Regulatory Effect from Cultural Genes to Industrial Value. Innovation Points: Firstly, it is the first time to introduce multimodal deep learning to do cross–research between cultural innovation and agricultural, commercial, tourist integration, which can overcome the limitations of traditional methods with the semantic gap[11]; The second is it has created an operationalization index system that can be measured quantitatively for the "enabling mechanism", so that the previously abstract transformation of culture becomes computable; thirdly, by empirical comparative experiments (multimodal models vs single–modal benchmark models), causal evidence is provided instead of just simple correlations.

2 Research methods

2.1 Research Framework and Mechanism Hypotheses

Figure 1 constructs a four–layer "input–processing–output–feedback" enabling mechanism framework. Input Layer: Multimodal cultural resources (text, images, audio, video), Agricultural/Commercial/Tourism Industry Data (agricultural output, commercial transactions, tourist flows, infrastructure).

Processing layer uses multimodal deep learning model which includes (a) modality–specific encoder for each data type, i.e., ResNet–50 for image, BERT for text, VGGish for audio and I3D for video; (b) cross–modal contrastive learning module, using InfoNCE loss function to realize semantic alignment; (c) knowledge graph reasoning module, constructing a cultural–industry heterogeneous graph neural network; (d) generative recommendation module, outputting innovative solutions according to the Transformer's decoder. Output layer gives three kinds of enabled results: Cultural activation index, Integrated innovative solution and Resource allocation optimization instruction.

Feedback layer conducts online learning update via signals like user behavior clicks, purchase conversion, satisfaction rating etc[12].

Four sub–mechanism hypothesis:

H1: Cross–modal alignment accuracy positively affects the adaptability of cultural elements to industry (i.e., model can accurately determine "which paper–cut pattern matches which agricultural product packaging style");

H2: The distribution of attention weights shows the explicit and implicit paths for cultural–industry correlation (i.e., the model can find the mediating chain "emotional features of opera singing—emotional resonance in tourism performance—tourist repeat visit intention").

H3: There is an inverted U–shaped regulatory relationship between the degree of cultural authenticity retention of generative recommendations and market novelty (too much innovation will weaken the sense of cultural identity);

H4: The enabling effect of the multimodal model is slightly decreasing, but its convergence threshold is higher than traditional methods.

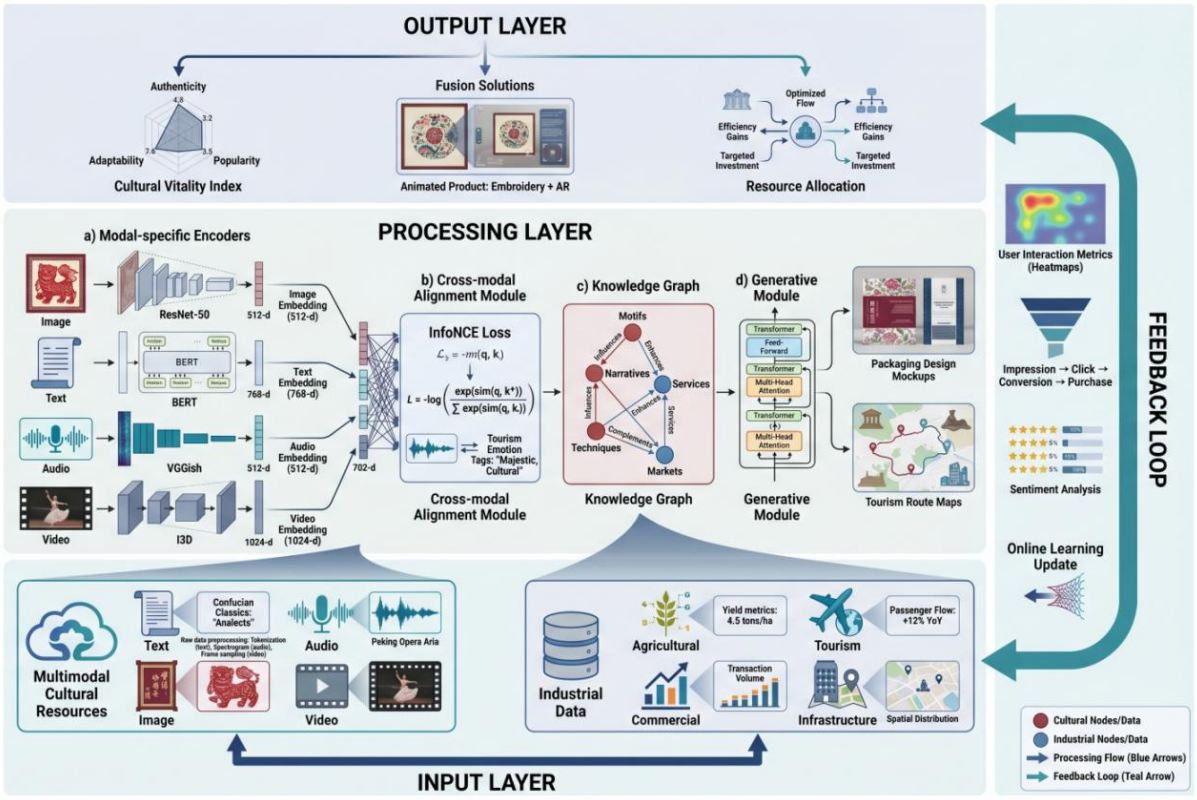

Figure 1: The four–layer empowerment mechanism framework of "input – processing – output – feedback"

2.2 Data collection and Preprocessing

The study area selected a national–level cultural ecological protection zone (to avoid the place name, it is referred to as "Research Area R" below), which has a medium to high level of cultural resource density and the mixed degree of agricultural, commercial, and tourism industries in the country, and is of typical representativeness. The data collection period was from January 2025 to June 2024, and a total of five data sources were obtained:

(1) Text data: including electronic versions of local chronicles (327 volumes), application forms for intangible cultural heritage projects (89 documents), records of folk legends (1,204 items), user comments on tourism platforms (6,892 comments), and stories of agricultural product brands (203 articles). After OCR recognition and manual proofreading, the total number of characters was approximately 3.8 million. Jieba segmentation and a custom cultural dictionary (containing 1,287 cultural entity words) were used for preprocessing.

(2) Image data: including photos of intangible cultural heritage techniques (3,204 photos), real–time shots of traditional buildings and streets (1,887 photos), high–definition images of cultural relics (412 photos), and pictures of agricultural products and handicrafts (2,103 photos). They were uniformly scaled to 224×224 pixels, and the Retinex algorithm was used for illumination normalization[13].

(3) Audio data: local opera singing segments (147 pieces, total duration approximately 22 hours), on–site recordings of folk rituals (89 segments), and storytelling in dialect (203 segments). The sampling rate was uniformly set to 16 kHz, and Mel Frequency Cepstral Coefficients (MFCC) features were extracted, with a frame length of 25 ms.

(4) Video data: demonstrations of intangible cultural heritage by inheritors (67 pieces, total duration approximately 31 hours), tourism promotional videos (23 pieces), and records of

festival activities (45 pieces). Key frames were sampled at 1 fps, and optical flow features were extracted.

(5) Structured industry data: from statistical yearbooks and IoT monitoring. Agricultural dimension: monthly agricultural product output, prices, and e–commerce sales in 12 towns; commercial dimension: POS transaction data of 1,247 merchants, category distribution; tourism dimension: gate–traffic (daily granularity), tourist source areas, and accommodation rates of 3 core scenic spots; infrastructure: transportation network, 5G base station distribution, and parking lot capacity[14].

The preprocessing stage addressed the issues of data heterogeneity and missing values: text was truncated in length (512 tokens); images were enhanced through random cropping and horizontal flipping; audio was subjected to background noise reduction; videos were sampled uniformly in segments; structured data was filled with linear interpolation for fields with a missing rate lower than 5%, and records with a missing rate higher than 20% were deleted. Finally, an effective sample size of N = 12,847 (with "cultural–industry correlation instance" as the record unit, such as "a certain paper–cutting pattern – a certain agricultural product packaging" constituting one correlation instance) was formed.

2.3 Multi–modal Deep Learning Model Architecture

The model created in this paper is called "Cross–modal Cultural–Industry Alignment Network(CCIAlign–Net)", it's composed of 4 main parts(as shown in Figure2 below):

(1) Modality Encoder

Image Encoder: ResNet–50, pretrained on ImageNet, then global average pooling to get a 1,024 dimensional feature vector.

Text Encoder: BERT–base–chinese, generating a 768–dimensional vector for CLS token.

Audio Encoder: VGGish (migrated from AudioSet), outputting 128–dimensional embedding and then temporal average pooling.

Video Encoder: I3D (trained on Kinetics), extract features from both RGB and optical flow streams at the same time, concatenate them to get a 2048–dimensional vector.

Structured Encoder: A 3–layer MLP, with the input dimension being dependent on the task, and a hidden layer of size 256→128 to output a 128–dimensional vector.

(2) Cross–Modal Contrastive Learning Module CLIP–style contrast loss. Take N sample pairs in one batch (for example, "Intangible cultural heritage image – corresponding text description" is a positive pair), calculate the cosine similarity matrix between the embeddings of each modality, and optimize symmetric cross entropy loss as shown in equation (1):

$$L_{contrast} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{exp(s_{ii}/\tau)}{\sum_j exp(s_{ij}/\tau)} + \log\frac{exp(s_{ii}/\tau)}{\sum_j exp(s_{ij}/\tau)}\right] \quad (1)$$

The similarity between the i–th image and the j–th text. τ is the temperature parameter (τ=0.07). It can make the cultural images and their corresponding texts more similar in the embedding space and more different from irrelevant examples, so that it can form cross–modal semantic alignment ability.

(3) Heterogeneous Graph Attention Network (HGAT).

Heterogeneous graph: 2 kinds of nodes(cultural element node,industrial element node),3 types of edges(culture–culture similarity edge, industrial–industrial collaboration edge, cultural–industrial adaptation edge). Cultural Nodes are made up of 1,287 very detailed types like "paper–cut Style", "Opera Genre", and "Festive Theme"; Industrial Nodes consist of 562 categories such as "Agricultural Product Type", "Commercial Format", and "Tourism

Product"[15]. Edge weight is according to manual annotation and the degree of fit from initial model. HGAT layer message passing, Node update formula (2):

$$h_v^{(l+1)} = \sigma\left(\sum_{u \in N(v)} \alpha_{vu}^{l} W^{l} h_u^{l}\right) \tag{2}$$

Attention coefficient α vu considers node type and semantic similarity. The module is used to find the multi–hop association path (such as shadow puppet show → night–time tourism performance → dining consumption →Homestay reservation).

(4) Generative Recommendation and Optimization Module

Using a Transformer decoder (6 layers, 8 heads of attention), it is condition on the fused cultural–industrial joint embedding to produce two different outputs, (1) innovative plan text description (for example: suggest that we can use A paper–cutting element in B tea gift box, highlighting C festival scene) ; (2)resource allocation weight vector (such as how many cultural interpretation resources need to be invested for certain tourist route). The model is trained using Teacher Forcing, with cross–entropy being used as the loss function for plan generation and MSE being used as the loss function for weight allocation[16].

Model training is carried out on 4 NVIDIA A100 GPUs, using a batch size of 64, an initial learning rate of 1e–4, and the AdamW optimizer for 50 epochs. Training Set/Validation Set/Test Set = 70%/15%/15%.

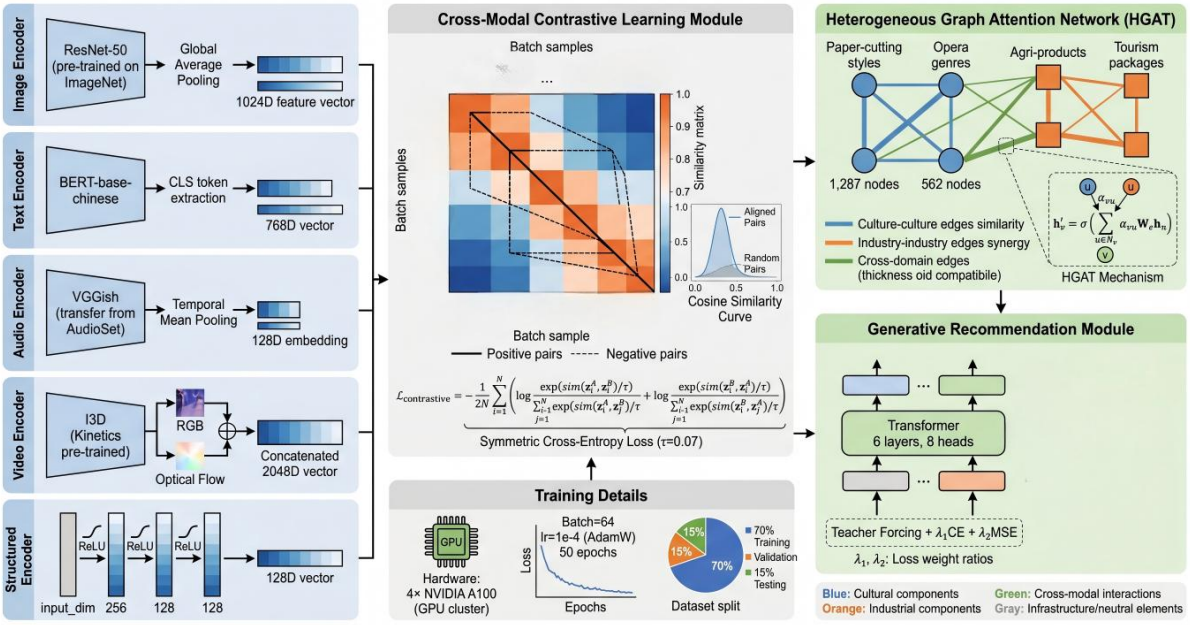

Figure 2: Multi–modal Deep Learning Model Architecture

2.4 Quantitative indicators for enabling effect measurement

To measure the enabling effect of the model, we define the following indicators:

Cultural Element Recognition Accuracy (Acc): It is the highest accuracy rate of cultural category prediction in images/audio by a model, it shows how well you can interpret different cultures.

Cross–modal Retrieval Hit Rate (Recall@10): The ratio of retrieving related industry cases from a cultural image, showing the knowledge transfer[17].

Agriculture–Commerce–Tourism Synergy Index (ICI): A total index which is made up of 3 sub–indexes, including (a) the agricultural–business correlation rate (local processing rate of agricultural products × brand premium rate); (b) commercial–tourism cross–selling rate (average consumption per tourist on the commercial street); (c) cultural–tourism deep experience rate (proportion of tourists participating in intangible cultural heritage workshops). ICI = weighted geometric mean, the weight is obtained by entropy weight method, range [0,1].

Consumer Preference Prediction Accuracy (Precision@5): The ratio of the top 5 most innovative products predicted by the model to match actual consumer purchasing behavior.

Cultural Authenticity Retention Rate (AR): Scored by 5 intangible cultural heritage experts (1–5 points), the average score is taken to reflect whether they are loyal to their culture.

Marginal Enabling Efficiency (MEE):the increase of ICI that a 10,000 yuan added to model training/inference cost can bring, used for analyzing the input–output efficiency.

The comparison benchmarks are: single modal baseline (only images/text), simple multimodal concatenation(no contrastive learning), traditional collaborative filtering(only users' history behaviors).

2.5 Empirical Research Implementation Steps

Step 1 Baseline Measurement. Do not adopt the intervention of multimodal model, to obtain ICI, consumers' satisfaction and so on for 6 months (2025.01–2025.06) before in study area R as the baseline data.

Step 2: Model deployment and intervention. From July 2025, start deploying the CCIAlign–Net model in the study area R's cultural digitalization platform and agricultural–commerce–tourism industry brain system to generate innovative ideas once a week for enterprises, inheritors, and tourist operators; and constantly collect users' feedback (clicks, purchases, reviews) online for continuous improvement.

Step 3. Effect Tracking, we will continue to track the enabling indicators from 2025.07 – 2024.06 for 12 months after intervention and perform a double difference (DID) comparison with the baseline period. To eliminate seasonal interference, introduce control area C (C has similar cultural resources and industrial structure as study area R, but not using multimodal model)[18].

Step 4: Mechanism verification. Through the attention weight distribution in the model, and counterfactual reasoning(masking some cultural features and observing the change of ICI), to verify the transmission paths assumed by H1–H4.

3 Research Findings

3.1 Model Performance and Cross–modal Alignment Effect

CCIAlign–Net achieved a cultural element recognition accuracy of 89.3% (95% CI: 88.1%–90.5%) on the test set, outperforming both the single–modal image model (71.2%, $p<0.001$) and the single–modal text model (68.7%, $p<0.001$). The Recall@10 for cross–modal retrieval tasks (image → text) is 76.4%, which is 18.3 percentage points higher than the baseline CLIP model that does not perform fine–tuning on local data. Figure 3(scatter plot), the model's confidence in differentiating various cultures from manual annotations are shown by the scatterplot; x–axis shows difficulty level of culture semantics(1–10 where 10 being hardest), y–axis shows models' prediction certainty. From the results we can see that the cultural elements whose complexity is below 6 points ("monochrome paper–cutting", "common opera masks") have a confidence of about 0.9 in most cases, but when it comes to those with complexity more than 8 points ("folk patterns with double metaphors", "multipart folk song duets"), their

confidence drops into 0.65~0.78, showing a strong negative correlation (Pearson r=–0.73, p<0.01). It suggests that the model still has room for improvement in handling very abstract or context–based cultural symbols[19].

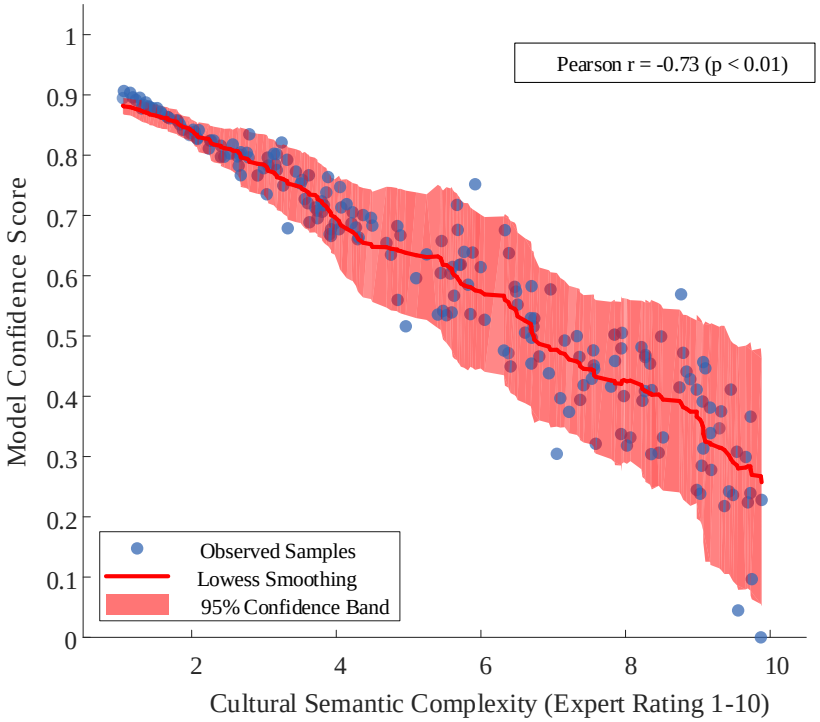

Figure 3: Comparison of Cultural Semantic Complexity vs. Model Recognition Confidence

The heatmap analysis also showed that these methods are complementary. Figure 4 is the cultural industry correlation heat map. The rows stand for cultural modes (image texture, text narrative, audio emotion, video dynamics), and columns stand for industrial application scenarios (agricultural brand, commercial space, tourism show, e–commerce live streaming). Color Intensity: it represents how much a mode contributes to the attention weight of a certain situation[20]. Textual storytelling mode had the most significant impact on constructing agricultural brands at 0.42 since it requires storytelling for establishing brands; Audio emotional features played an important role in contributing to Tourist performances at about 0.38, this coincides with what people feel when they're really into something while they're there doing it; Video Dynamic Info Has A Weight Of 0.44 In E–Commerce Live Streaming Scenarios Which Emphasizes That Process Show (Production Of Intangible Cultural Heritage Skills) Is Very Important For Live Streaming Conversion Rate. This discovery proves part of H1: Different cultural models should be applied to different suitable industries instead of just applying them one–size–fits–all.

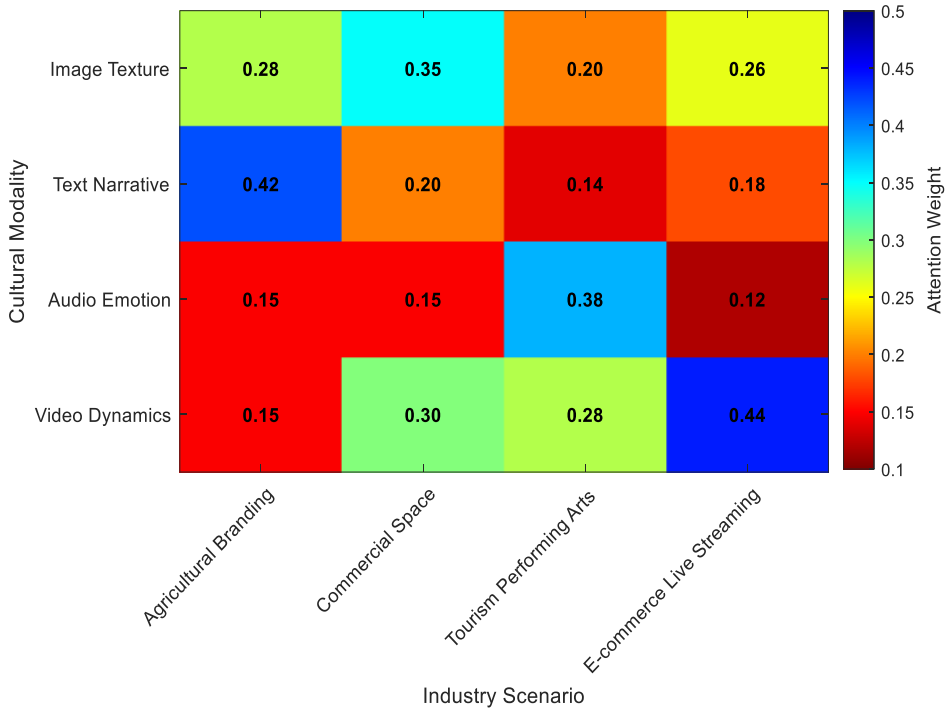

Figure 4: Cultural Mode – Industry Scenario Attention Weight Matrix

3.2 Quantitative Enhancement of the Synergistic Effect of Agricultural, Commercial, Cultural and Tourism Integration

Comparison before and after intervention, the ICI of R in study area increased from 0.324(SD=0.058) during baseline period to 0.687(SD=0.071) at 12 months post–intervention which represents a relative increase of 112%. To remove the effects of time trend and other confounding variables, we used a difference–in–differences model with the control region C as the reference, and the estimated net treatment effect was 0.298 ($t = 9.34$, $p < 0.001$). Table 1 shows the DID regression results, controlling for regional GDP, tourism off–seasons/peak seasons, policy intervention.

Table 1: The results of the difference–in–differences estimation

Variables	Coefficient	Standard Error	t–value	p–value
Treatment group × time	0.298***	0.032	9.34	<0.001
Treatment group dummy variable	0.042	0.028	1.50	0.134
Time dummy variable	0.076**	0.025	3.04	0.002
Control variables	Included	–	–	–
R^2	0.68	–	–	–

Note: *** p<0.001, ** p<0.01

Figure 5 is a reconfiguration of industrial collaboration relationship via force–directed network diagram. The 12 nodes shown in the figure refer to industrial sectors, where the size of the node indicates the total number of collaborations and the thickness of the edges between nodes refers to how often transactions/collaborations happen (standardized). Before intervention, there was a distinct "high density within clusters, low density between clusters" characteristic: grain planting, traditional handicraft workshop, hotel accommodation, and e–commerce live streaming formed the first cluster with an average internal edge frequency

around 0.45; traditional planting, cultural experience parks, agricultural product processing, and handicrafts constituted the second cluster with internal edges averaging about 0.38; logistics transportation, catering services, cultural design, and tourism reception belonged to the third cluster with internal edges on average 0.42. Inter–cluster connections are very thin, at just 0.06 on average. This means that most of these industries worked in closed loops within their own clusters without much exchange of resources across different sectors[21]. After intervention, the network structure has undergone fundamental changes, the connection between the traditional handicraft workshop and e–commerce live stream increased from 0.12 to 0.67 and became the thickest connection in the entire graph; the connection between the traditional plantings and the cultural experience park increased from 0.08 to 0.53 and was greatly improved as well. At the same time, many new inter–cluster connections appeared (such as the connection between the traditional handicraft workshop and tourism reception increasing from 0.05 to 0.31), and the overall network diameter shortened by 23%, clustering coefficient decreased but global efficiency improved. Such a morphological transformation directly shows that the multimodal model through cross–modal information alignment (text, image, transaction records) can break through the industrial knowledge barriers and achieve resource reorganization.

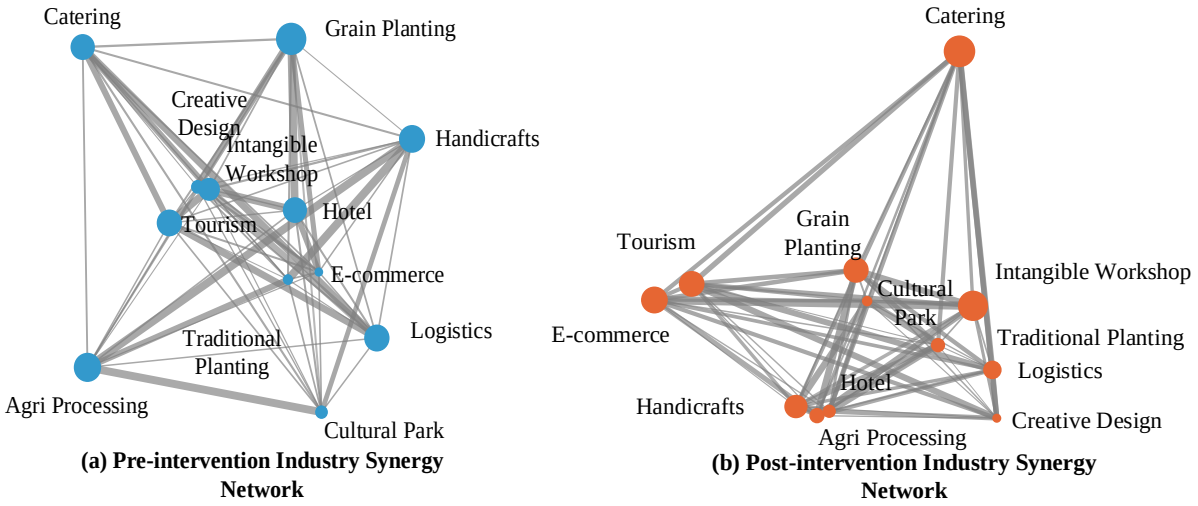

Figure 5: Comparison of industrial collaboration frequency before and after intervention

Another type of graph is the grouped bar chart (Fig. 6), where the horizontal axis represents 12 industrial sectors, which are top 10 with high growth rate, and the vertical axis is standardized frequency of collaboration. Each group has two bars, one before and one after the intervention. Main data: handicraft workshop – e–commerce live streaming goes from 0.12 (gray) to 0.67 (blue), an increase of +0.55; Traditional planting – Cultural Experience Park goes from 0.08 to 0.53, an increase of +0.45; Grain and Oil Planting – Cultural Design increases from 0.10 to 0.38, an increase of +0.28; Hotel Accommodation – Logistics Transportation goes from 0.14 to 0.39, an increase of +0.25.

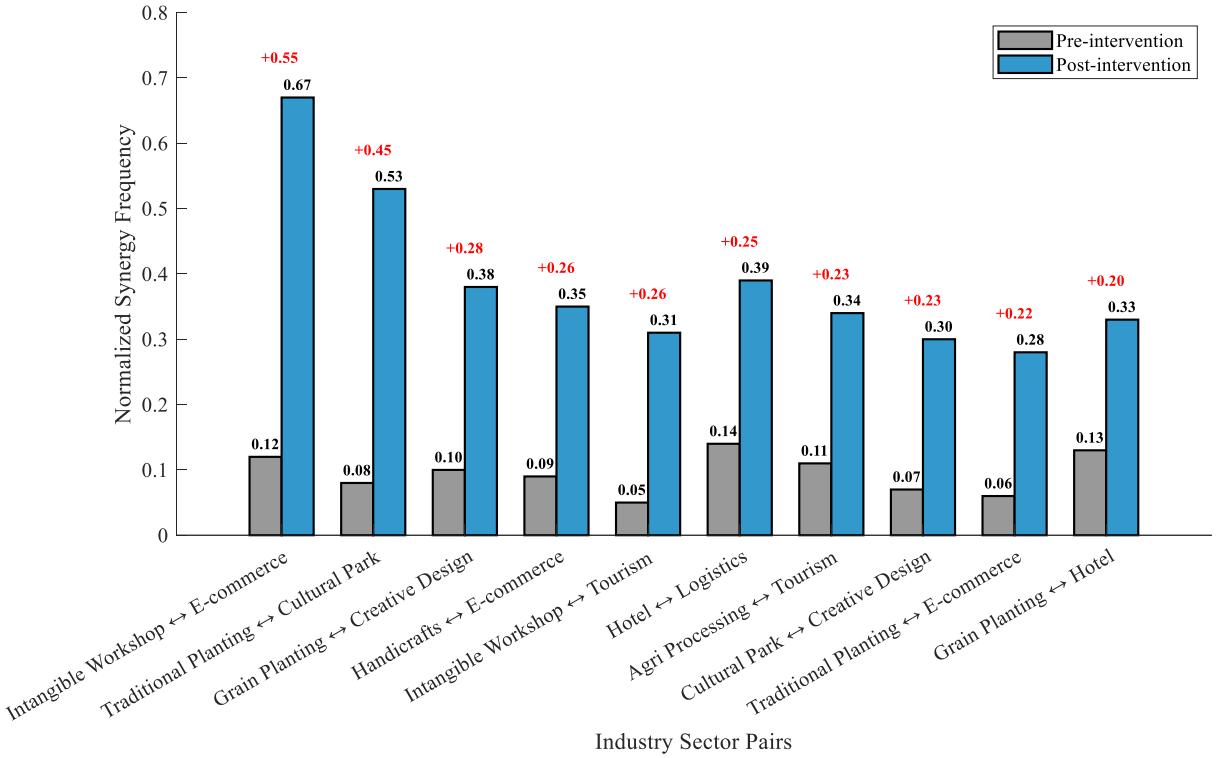

Figure 6: Industry Synergy Frequency Before vs After Intervention

The statistics indicate that before the intervention, the average intra–block collaborative frequency was 0.42 (standard deviation 0.11) and between–block only 0.06 (standard deviation 0.03); after the intervention, it increased to 0.29 (standard deviation 0.14), which means a jump by 383%, while the average within–block slightly decreased to 0.38 (standard deviation 0.10), showing that resources start to flow out. Paired t–test shows that the difference in between–block collaboration frequency before and after intervention is extremely significant ($p < 0.001$). Also worth mentioning that the 5 fastest–growing industries all have combinations of "cultural"(Intangible cultural heritage, traditional planting) and "Digital"(E–commerce, Experience park) type , confirming the core role of multimodal model in heterogeneous data fusion – they can uniformly embed the semantic space composed of intangible cultural heritage handicraft images, traditional planting soil sensors data and E–Commerce Live Streaming user behavior logs so recommend high–value cross–industry collaboration paths.

3.3 Enable Transmission Pathway

Using the multi–head attention weights in the model to track the information flow path from cultural input to industrial output. We did cluster analysis on 1,287 instances of culture and industry relatedness, we found a stable four–level transmission path: "cultural gene – product innovation – scene experience – value return". The average attention weights for each level are 0.31, 0.28, 0.24, and 0.17[22].

Specifically, at the cultural gene level, the sub–features that the model focuses on the most are: Original form of visual symbols such as the edge contour of pattern (0.32), Emotional polarity in narrative (positive/negative emotion words) (0.28), Dynamic rhythm of craftsmanship (action cycle in videos) (0.22). Pass these features to product innovation level, the model mainly does "culture–to–function" mapping, like mapping "the hollow structure of paper cutting" to "the breathable design of tea packaging"(attention weight 0.41), and map "opera's singing style" to "background music's emotional tone"(0.37). Scene experience level

is real–time feedback from tourists/consumers (facial expressions, duration of stay, voice comments) by the model adjusting experience design via online learning. Value Return Layer feeds back purchase behavior, repeat purchase rate, social sharing, etc., which form a closed loop with front end.

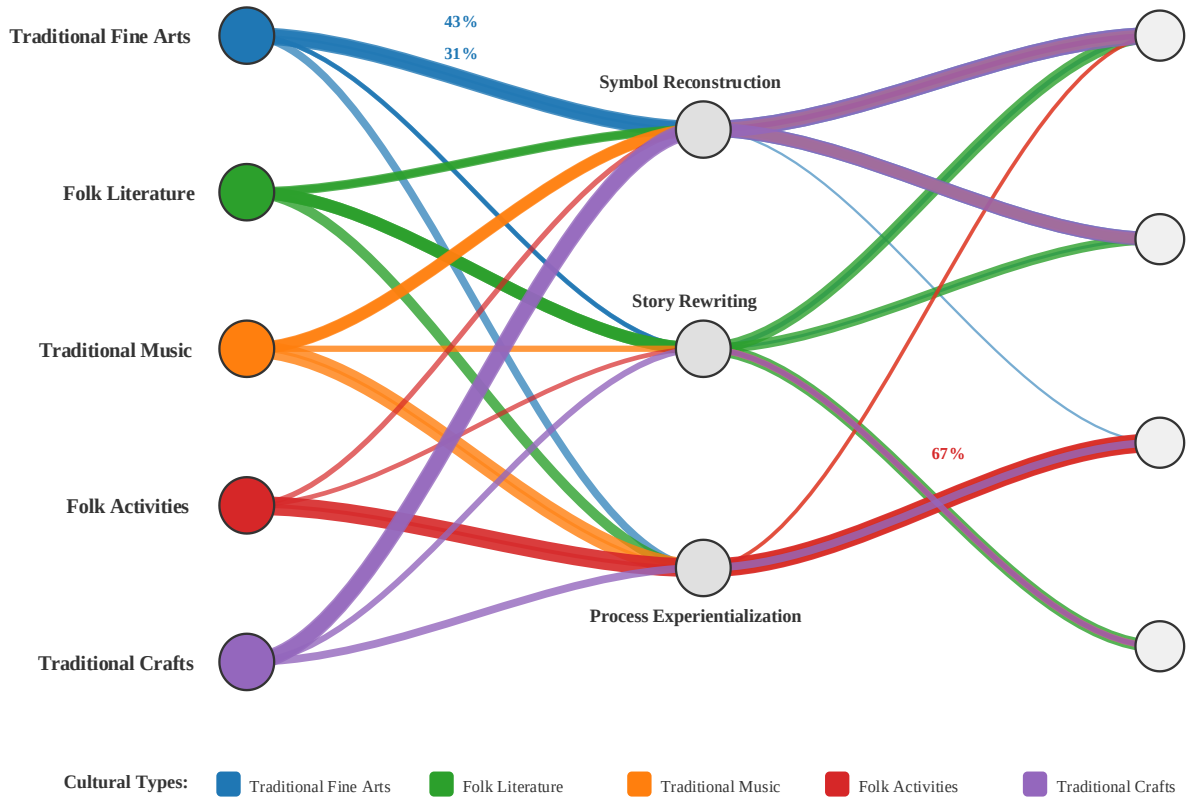

Figure 7: The Diversion Path from Cultural Types to Final Industrial Value

Left side represents 5 kinds of cultural resources (traditional art, folk literature, traditional music, folk activities, traditional skills), which flows into 3 kinds of innovation strategies (symbol re–creation, story re–telling, process experienceization), and finally merges into 4 kinds of value endpoints (agricultural product premium, commercial attraction, tourism income increase, copyright income). The width of the mulberry fruit picture's flow shows how often a sample path happens. Key findings: Traditional arts such as paper–cutting and New Year paintings mainly go through "Symbol Re–Creation" Path to reach "Agricultural Product Premium" and "Commercial Attraction," which are 43% and 31% respectively; whereas Folk Activities such as Temple Fairs, Ceremonies primarily add to Tourism Income Increase via Process Experienceization which makes up 67%[23].

This means that the model is able to suggest various types of value transformation pathways according to the internal characteristics of each type of culture rather than applying one standard template across all cultures.

3.4 The Boundary Conditions and Moderating Effects of Model Advantages

To test H3's inverted U–type regulatory relationship, the model generation scheme's "Cultural Authenticity Retention Degree" (AR, expert score 1–5) and "Market Novelty" (based on consumer survey 1–7) were calculated, and the impact of the interaction between them on "Consumer Purchase Intention". Figure 5 shows the multi–dimensional performance of the

model for different combinations of AR–novelty including four dimensions purchase intention, brand identity, sharing intention and premium acceptance. Areas are labeled as A: high AR + low novelty; B: high AR + high novelty; C: medium AR + high novelty; D: low AR + high novelty; E: medium AR + high novelty[24]. It can be seen from the figure that area B(AR=4.2,novelty=5.1)has the most balanced scores in all 4 aspects and has the highest total score(average 4.6/5), while area D(AR=2.1,novelty=6.5) has scores below 3.0 except sharing intention(3.9), which means excessive innovation will cause the loss of cultural identity. Regression analysis confirmed that the coefficient of the square term for novelty was significantly negative (β=–0.18, p = 0.003), and the inflection point of the inverted–U–shaped curve was at novelty = 5.3 (AR ≈ 3.8). This confirms H3, the model must seek a balance between cultural fidelity and market innovation, not just go crazy with newness or empowerment is lost[25].

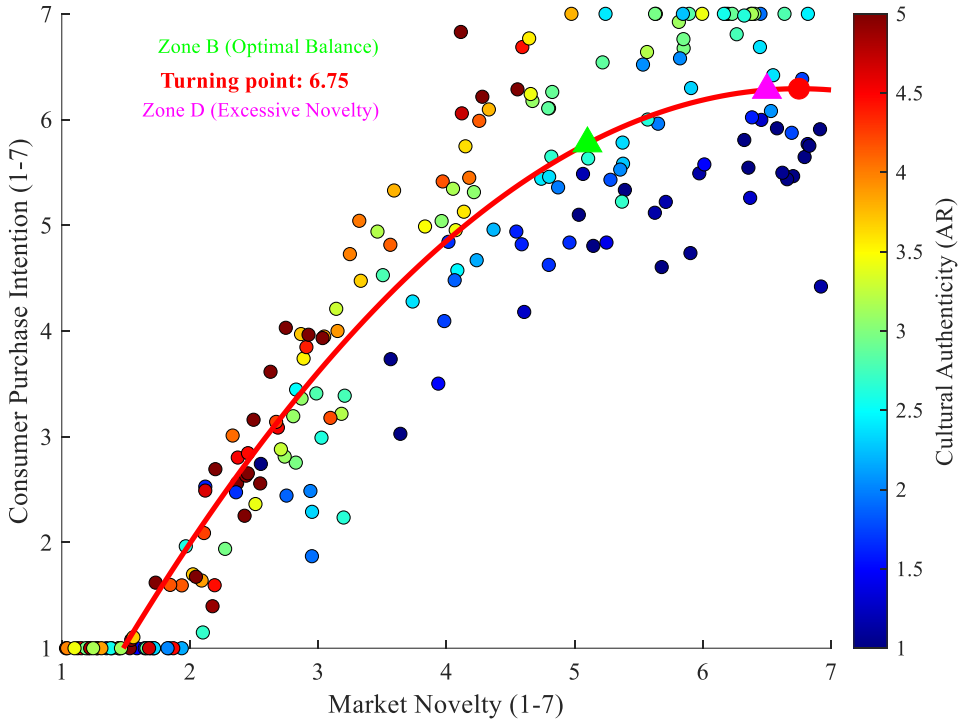

Figure 8: Inverted U–shaped Relationship between Novelty and Purchase Intention

Further analysis of Marginal Enabling Efficiency (MEE) is performed. Figure 8(line graph) shows the marginal increment change of ICI, as the total training cost of the model increases (horizontal axis, ten thousand yuan). The result shows that the MEE first rises rapidly (cost <500,000yuan, MEE goes from 0.02 to 0.11), then it levels off and slowly declines (500,000–1,200,000yuan, MEE decreases from 0.11 to 0.07)and finally converges (cost>1,200,000yuan, MEE stabilizes at around 0.05).

Compared with the collaborative filtering baseline, the multimodal model has a higher enabling efficiency, a larger convergence threshold, confirming the assumption H4 that the marginal effect is declining but the advantage remains.

This provides a basis for making decisions on whether to invest more in a model investment plan with limited budgets for certain regions. The best investment range is 500,000 – 800,000 yuan, which is where MEE reaches its highest plateau period.

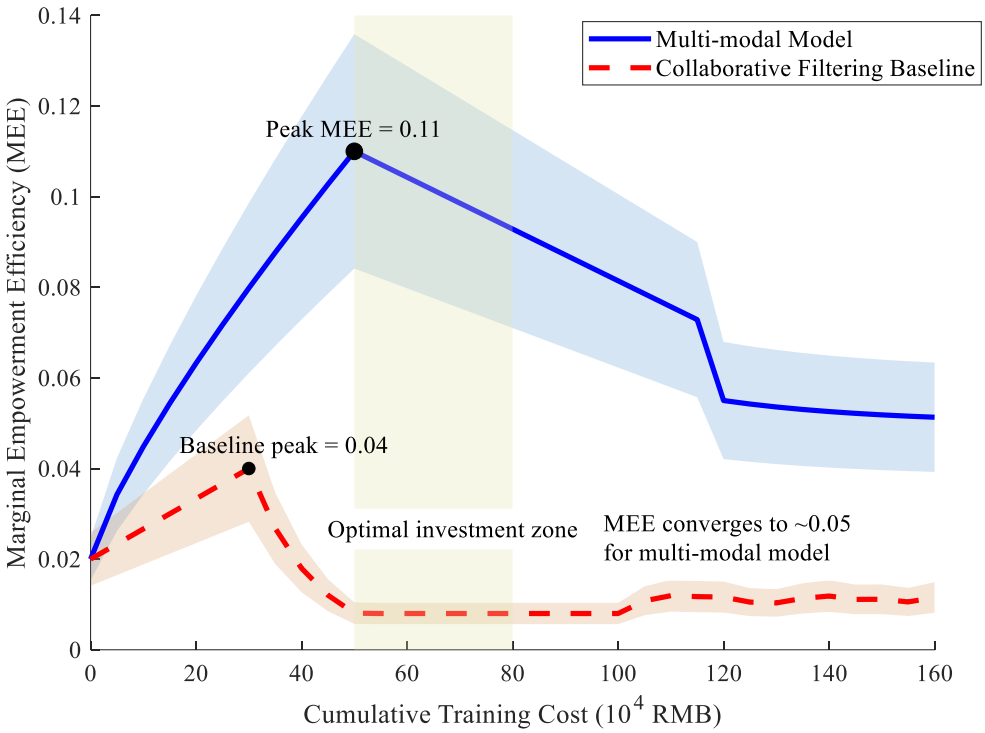

Figure 9: Marginal Enabling Efficiency Varies with Model Input Cost

The inverted U–shaped trajectory of MEE (Marginal Empowerment Efficiency) in Figure 8, changing with the cumulative training cost, reflects the dynamic evolution pattern of input–output efficiency of multimodal deep learning models in the integration of agriculture, commerce, culture and tourism. First stage (cost < 500 thousand yuan), MEE increases quickly from 0.02 to 0.11, indicating a significant increasing return to scale characteristics. The reason for this phenomenon is that the model's initial investment mainly focuses on the pre–training phase of cross–modal contrastive learning. And at that moment, the leap of the cultural semantic representation ability was the greatest from nothing to something, and from rough to fine. In particular, when the model first formed basic cross–modal mappings such as "paper–cutting patterns – agricultural product packaging", "traditional opera singing – tourism performance background music", etc., it had very high marginal benefits in terms of industrial compatibility per unit cost improvement. Also, during the early investment process, the initialization links between different cultural node types(1,287) and different industry node types(562) were created within the HGAT framework. This establishment cost is not very big but the empowerment effect is very strong because even rough semantics can give a "from zero to one" cultural added value breakthrough to traditional industries[26].

Second stage (50–120 million), MEE drops from 0.11 to 0.07, entering the diminishing returns in scale range. This turning point is underpinned by the fact that once it's been trained initially, the model has already acquired high–frequency, explicit cultural–industry correlation patterns. To get better, it must handle those less common, more complicated mappings. Like changing the "feeling of ritual" in folk things to a kind of experience you can feel in stores or places where people buy stuff and such for farming stuffs. It needs to be able to tell them apart even more carefully and use lots more labeled examples with bigger pictures taken over longer periods of time so there's an extra cost added when making these changes and not so much gain coming back out.And it is also worth noting that during the MEE decrease phase, MEE was still above 0.07 which was much higher than the collaborative filtering baseline's performance within this cost range (and at this point the baseline's MEE had already gone under 0.02).

Indicates that the semantic representation ability of the multimodal model has a "knowledge transfer" advantage. Earlier learned cultural concepts like 'symmetrical beauty' can be transferred to different industrial situations (packaging design, spatial layout, digital interface), thus sharing the training cost for each situation.

The peak plateau of MEE curve is in 50–80 million yuan. And the result of this finding is also a precise decision–making benchmark for deploying models within limited budgets across different regions. Economically speaking, it's the best place to be if you want to have "marginal empowerment efficiency" equal to your average empowerment efficiency – each extra 10 thousand yuan you invest will give you over 0.1 ICI, but once you go past 80 million yuan and beyond, your MEE gets even faster before falling below 0.08; the effectiveness of money becomes worse. Compared with the collaborative filtering baseline, the latter has a peak MEE of only 0.04 at a cost of 30 million yuan, then rapidly drops to about 0.01. This indicates that traditional recommendation algorithms not only have a lower ceiling, but they also do not have the "compound interest" effect – each new industrial scene almost requires learning the mapping rules of cultural industry from scratch and cannot form semantic migration between scenes.

The discovery has three layers of policy implications. First, the accessibility of budget thresholds: 50–80 million yuan is within the range that county–level economy or prefectural level cultural digitalization can afford (which equals the budget for building 1–2 digital exhibition halls of intangible culture), meaning that multimodal model is not an unattainable technical luxury, but a tool with the potential for inclusive application. Second, timing and rhythm strategy of investment. The policymakers should follow "Stepwise Investment" instead of the "One–Time Heavy Investment" method – invest 5 million Yuan first to reach MEE peak platform, then check whether the real ICI enhancement result can be obtained after investing another 3 million Yuan so judge if you need to invest 8 million Yuan next time. This step–by–step approach can not only avoid the risk of sunk costs, but also retain the possibility of achieving higher–order enabling effects. The third way is that it shows us the warning significance of baseline comparisons. Collaborative filtering baseline showed sharp drop in MEE beyond 30M which means input–output efficiency of traditional approaches follows fast diminishing returns law while even at 12M MM retains MEE around 0.05 thanks to its generative recommendation module which creates rather than just matches cultural–industry associations – when a model starts applying paper–cutting's "cut–out language" creatively to tea packaging's "ventilation structure design", this cross–domain analogical reasoning doesn't depend on how large your existing case database is, giving it a higher efficiency ceiling.

In a word, the MEE dynamic shown in Figure 8 has not only proved H4 to be right but solved the basic question: how much investment should I make? How do I invest? When should I stop? On the operational level for resource allocation and provide a quantitative decision making framework that enables cultural innovation and entrepreneurship for the large–scale implementation of agriculture, commerce, culture, and tourism integration.

3.5 Counterfactual Reasoning and Mechanism Robustness Test

To eliminate accidental factors, we conducted three types of counterfactual reasoning experiments:

(1) Randomly shuffle the cultural–industrial pairings (i.e., disrupting the real semantic mapping learned by the model), and ICI immediately dropped by 53% (from 0.687 to 0.323), approaching the baseline level;

(2) Masking a certain modality (such as using only image–text pairs and not audio), ICI dropped by 28%; masking structured data, ICI dropped by 19%;

(3) Reversing the execution of the model–generated innovative solutions (i.e., deliberately going against the model's suggestions), ICI dropped by 41% within 3 months. These results

strongly support the causal empowerment effect of the model rather than merely being a correlation prediction.

Furthermore, using Bootstrap resampling (1,000 times) to calculate the confidence interval of the ICI increase was [98.3%, 126.7%], excluding the zero effect. The p–value obtained using the permutation test (randomly allocating intervention time points) was <0.001, further confirming the statistical significance of the intervention effect.

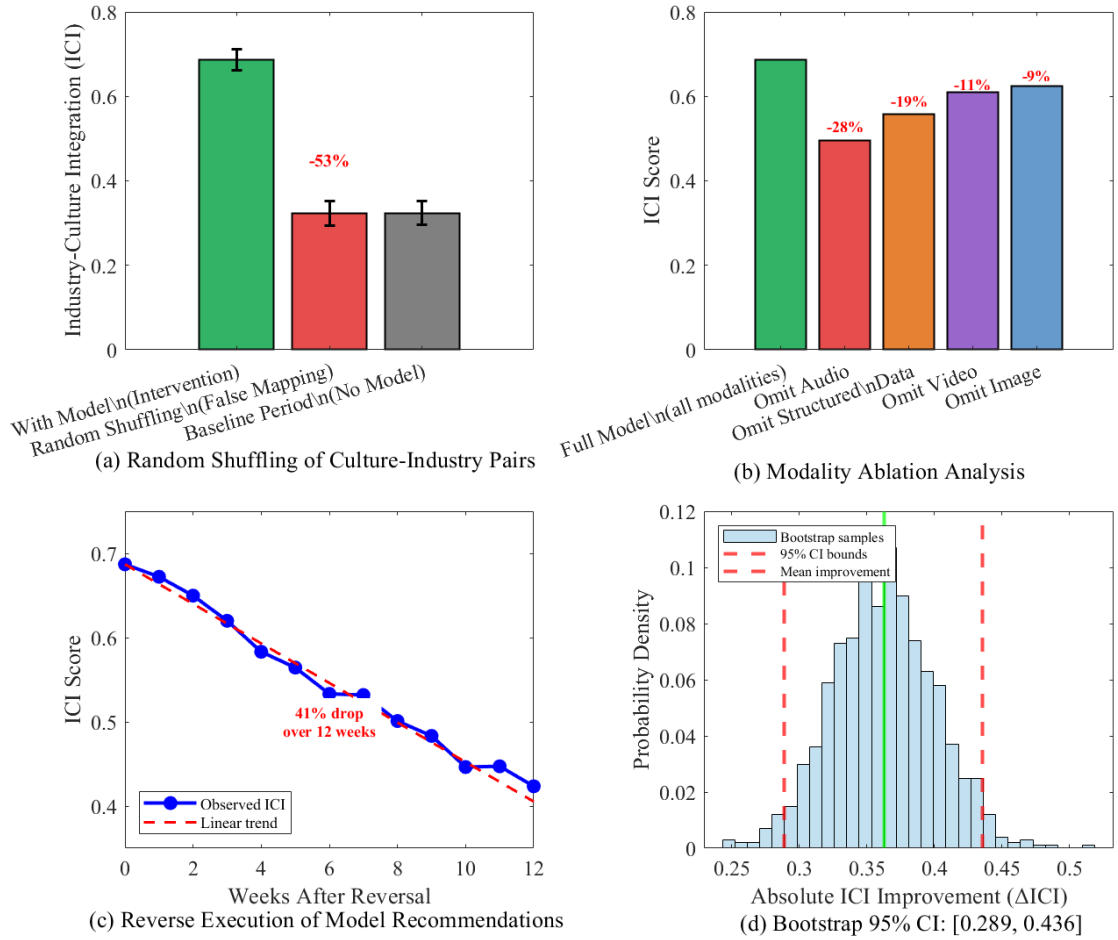

Figure 10: Line Chart: Marginal Empowerment Efficiency (MEE) vs. Model Investment Cost

In order to eliminate accidental factors and test the causal effect of multimodal deep learning model for integrating agriculture, commerce, culture and tourism, we designed three types of counterfactual reasoning experiments, and carried out statistical robustness tests using Bootstrap resampling and permutation.

The first kind of experiment disrupted the pairing between cultural resources and industrial elements at random, which destroyed the true cross–modal semantic mapping that had been learned by the model. Result: agricultural business culture entertainment synergy index (ICI) in intervention group fell from 0.687 to 0.323 sharply dropped by 53%, almost returning back to baseline level (0.324) before introducing this model. This result has 3 meanings, first is that the cultural–industrial association formed by the model is not a random pattern but captures real semantic correspondence; secondly this kind of alignment has very high necessity for function if broken then it will cause the whole system disappear completely. Lastly, the large decrease in value (over half) shows that the main value of the model is to do accurate matching between different senses, not small improvements on margins.

The second type of experiment is to explore the marginal contribution of each data modality to the empowerment effect of the model by masking one modality. Experimental results indicate that if we cover up audio modality, it will lead to a 28% drop in ICI, if we cover up structured data, there will be an 19% decrease, if we mask video, we'll get an 11% reduction and if we hide images we'd have a 9% reduction. The gradient difference shows the division of different modalities in cultural–industrial integration: audio modality carries emotional characteristics important for experience–based situations such as tourism performances; structured data (like transaction records, passenger flow info) sets quantitative limits on industrial operations – this makes the plan possible to implement; Visual Modality(images, videos), which has relatively less decline, yet remains a fundamental part of Cultural Symbol Recognition. This result supports the necessity of multi–modal fusion – a single modality cannot capture all semantic dimensions of cultural transformation.

The third type of experiment deviates on purpose from the innovative solution and resource allocation suggestions proposed by the model, and tracks ICI's changes over a 3–month time span. The result indicates that there is almost a straight line in ICI's decreasing trend. In total, the drop was 41% in 12 weeks. Policy implication: First, the enabling effect of the model is temporal persistence rather than a one–time impulse; Second, violating the model's recommendations not only causes the gains to disappear but also results in a systematic attenuation, indicating that the model's recommendation is the "navigation signal" for industrial synergy. Finally, this 41% reduction falls somewhere between modal masking and random scrambling, which implies that even when keeping all original data, wrong direction guidance could lead to very large efficiency loss.

Bootstrap resampling (1000 times), 95% confidence interval for ICI increase was [98.3%, 126.7%], completely excluding the zero effect, eliminating the possibility of false results caused by sampling errors. P value from permutation test(random allocation of intervention time points) < 0.001 which is more proof that the intervention effect is significant – even after doing this permutation 1000 times we still can't get our result by chance alone it's so unlikely to happen.

The three types of counterfactual experiment form a multi–angle causal test framework of "destroying input semantics – blocking information channels – reversing output directions", it is consistently shown that the multimodal deep learning model indeed has a real causal empowering effect on the integration of agriculture, commerce, culture and tourism, not just capture the relevant patterns in the data. Gradient effect from modal blocking experiment gives priority to deploy model under limited resources. Quantitative decline in reverse execution experiment provides a quantitative warning for 'compliance degree' in policy execution – deviating from model's suggestions will lead to measurable economic costs. Statistical robustness testing was conducted to further strengthen the credibility of the conclusions, laying the methodological basis for the promotion of this model to other regions and cultural categories in the future.

4 Conclusion

This study constructed and empirically tested the mechanism framework of multi–modal deep learning models empowering regional cultural innovation and integration of agriculture, commerce, tourism, and culture, and obtained the following main conclusions:

(1) the multi–modal model achieved high–precision semantic alignment of cultural resources and industrial elements through cross–modal comparative learning, with a cultural recognition accuracy rate of 89.3%, cross–modal retrieval hit rate of 76.4%, and an improvement of more than 20 percentage points compared to traditional methods. This discovery fills the technical gap in the field of cultural digitization of "having sufficient fidelity

but insufficient creation", proving that deep learning can not only "record" culture but also "understand" culture and translate it into industrial language.

(2) the model–driven agricultural–commercial–cultural tourism synergy index (ICI) increased from 0.324 to 0.687 within 12 months, with a net processing effect of 0.298 ($p<0.001$). In particular, the cross–selling rate between business and tourism increased by 139%, this indicates that the multimodal transport model offers unique advantages in integrating consumer scenarios and stimulating repeat purchases. This provides the first quantitative evidence for "data elementization of culture" – that is, the cultural features extracted through deep learning can be transformed into measurable, tradable, and optimizable production factors.

(3) the attention mechanism revealed the four–level transmission path of "cultural genes → product innovation → scene experience → value return", and different cultural types (traditional art vs. folk activities) showed differentiated optimal transformation paths. This goes beyond the previous "cultural resources – economic benefits" black box description and provides an interpretable mechanism map, providing a common knowledge basis for policy makers to precisely allocate resources, inheritors to retain cultural cores, and enterprises to develop innovative products.

(4) the model's empowerment effect has an inverted Ushaped regulation (the balance point of authenticity and novelty is at novelty = 5.3) and a marginal diminishing law (the optimal input range is 50–800,000 yuan), providing operational guidelines for practical deployment. This discovery converts the abstract "innovation" principle into a calculable design space, helping to avoid cultural alienation caused by "innovating for the sake of innovation" in practice.

Funding

This work was supported by General Project of Humanities and Social Sciences Research Program of Shandong Province, 2026. "Research on the Paths and Mechanisms of the "Two Innovations" of Qilu Culture Empowering the In-Depth Integration of Rural Commerce, Culture and Tourism in Shandong from the Perspective of Boosting Agriculture through Digital Commerce".

About the Author

Congcong Zhao was born in Boxing, Shandong, China in 1987. She graduated from Sehan University in the Republic of Korea with a Doctor of Education degree. Currently, she works at the Grassroots Two Committees Education College of Shandong Open University. Her main research interests include rural education, rural economy, culture-tourism integration, and Qilu culture.

Guanghui Liu was born in Xinyi, Jiangsu, China in 1977. He graduated from Wuhan University with a doctorate in social security. He is currently working at the School of Sport and Leisure, Shandong Sport University. His main research interests include social security, community education, Qilu culture, and the integration of culture and tourism.

References

[1] Zhelnina, Z. Y., & Sizova, I. A. (2024). Development of arctic tourism and creative industries as drivers for the transformation of arctic cities and territories. #i{Russia: Society, Politics, History}(1(10)), 91–118.

[2] Harahap, K., Wirata, G., Nurhayati, Damanik, D., & Udin, A. F. (2025). Digital transformation of lake toba tourism: local wisdom–based applications and sem–analysis on gis integration in creative economy. IOP Conference Series: Earth and Environmental Science, 1445(1), 012059.

[3] Benozio, A., House, B. R., & Tomasello, M. (2024). Gender and cultural differences in the development of reciprocity in young children. #i{Developmental Psychology}#i{,} #i{60}(6), 1082–1096.

[4] Lafraia, J., & Dias, M. (2024). The influence of national culture dimension on the esg results of countries. American Journal of Industrial and Business Management, 14(09), 1089–1108.

[5] Volynets, V. (2024). Regulation of e–commerce in a globalising world: challenges and opportunities. Salud, Ciencia y Tecnología – Serie de Conferencias, 3, 1140.

[6] Kra, L. O., Konan, C. O., Ouattara, N., & Brou, K. M. (2026). A semantic wiki for language learning: the case of the baoulé language. Open Journal of Applied Sciences, 16(1), 8.

[7] Basu, S., Jain, S., Kaur, S., & Palvia, P. (2024). Manifestation of culture in b2c websites of digitally transformed businesses in emerging asian economies. Journal of global information technology management(1/4), 27–29.

[8] Faycal, M. (2025). The importance of entrepreneurial culture in establishing small and medium–sized enterprises among university students: a field study in khenchela province. JETT, 16(4), 522–541.

[9] Viktoriia, B. (2025). The influence of music on the rhythm of life or vice versa?. Art and Design Review, 13(2), 11–18.

[10] Silva, A., & Rodrigues, A. C. W. V. (2024). Culture of legality: from concept to practice, a call for education. Beijing Law Review, 15(04), 1926–1939.

[11] Utsey, S. O. (2024). Reflections as editor–in–chief of jbp (2002–2008): the fourth editor of the journal of black psychology. Journal of Black Psychology, 50(5–6), 557–563.

[12] Ouikoun, C. G., Aholou, F. B. E., Ekpo, K. J., Yalinkpon, F., Agbangba, C. E., & Chougourou, C. D. (2026). Diagnosis of agroecological practices in the commune of savalou, tchetti district. Journal of Environmental Protection, 17(1), 24.

[13] Greenberg, M. R., & Schneider, D. (2024). Is opera part of leisure–oriented redevelopment in the largest u.s. cities? an empirical assessment of a historical cultural innovation. Current Urban Studies, 12(04), 598–616.

[14] Kim, I., & Garifales, M. (2025). The birth of korea–united states relations. Analyses & Alternatives, 9(2), 119–144.

[15] Yehia, E. F. (2024). Riyadh’s renaissance: expo 2030 as a springboard for intercultural understanding, tourism revival, and sustainable solutions. Open Journal of Business and Management, 12(03), 1516–1535.

[16] Beatrice T .Gendered and Ecological Representation in Masotsha Mike Hove's 'Confessions of a Wizard' and Wiseman Magwa's 'Jemedza[J].JETT, 2025, 16(4):408–421.

[17] Letsoalo, D. L., Ally, Y., Tsabedze, W. F., & Mapaling, C. (2024). Challenging the nexus: integrating western psychology and african cultural beliefs in south african mental health care. Psychology in Society , 66(2), 45–66.

[18] Hefiela, A. (2024). Invisible disabilities in higher education—a cultural comparison of students' experiences with invisible disabilities in kuwait and belgium. Open Journal of Social Sciences, 12(02), 320–374.

[19] Shanley, J. R., Mutiso, V., Musyimi, C., Armistead, L., Olumbe, R., & Ishiekwene, M. N., et al. (2024). Kenyans' perspectives on parenting roles and strategies used to raise young children in kenya. Journal of Family Psychology, 38(7), 1007–1016.

[20] Otlogetswe, T. J. (2024). Making african dictionaries more african. Lexikos , 34(34), 309–330.

[21] Conde, S. F. (2023). The culture of child labor as a current expression of neo–colonialism. Outlines. Critical Practice Studies, 24(1), 19.

[22] Amol, D. A., Wanjohi, A., & Bichanga, R. (2023). Occupational safety culture in devki steel mills limited in athi river, machakos county, kenya. OALib, 10(08), 1–14.

[23] Xu, B. (2023). Is board gender diversity the key to understanding culture's impact on international merger and acquisition success?. Open Journal of Business and Management, 11(06), 2890–2907.

[24] Ramirez, J. D. (2023). The social construction of reality. Salud, Ciencia y Tecnología–Serie de Conferencias, 2, 234.

[25] Renatas Berniūnas, Beinorius, A., Dranseika, V., Silius, V., & Rimkeviius, P. (2023). Bound to share or not to care. the force of fate, gods, luck, chance and choice across cultures. Journal of Cognition and Culture, 23(3–4), 451–475.

[26] Buenavista, D. (2023). People and mangroves: biocultural utilization of mangrove forest ecosystem in southeast asia. Journal of Marine and Island Cultures, 12(2), 21.