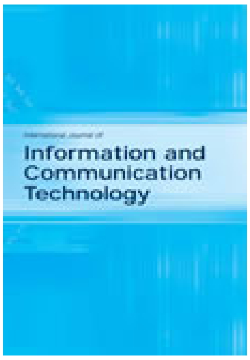


International Journal of Information and Communication Technology

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

M-DRAMA: a multimodal-driven framework for classical drama short video promotion

Jun Su

DOI: 10.1504/IJICT.2025.10074750

Article History:

Received:	17 June 2025
Last revised:	08 July 2025
Accepted:	08 July 2025
Published online:	01 December 2025

M-DRAMA: a multimodal-driven framework for classical drama short video promotion

Jun Su

College of Humanities and Arts,
Xi'an International University,
Xi'an, 710000, China
Email: 18220512802@163.com

Abstract: Facing declining youth engagement in traditional theatre (under 30% attendance), this study addresses the paradox of surging opera-related short video consumption by proposing a multimodal-driven framework for targeted classical drama promotion. We introduce M-DRAMA, an integrated model leveraging three technical innovations: A drama knowledge graph (DKG) with hyperbolic embedding to structure cultural metadata; a cross-modal alignment (CMA) module enforcing frame-level synchronisation of lyrics, movements, and music via matrix constraints, reducing semantic deviation to < 0.2 s. A spatiotemporal interest decoupling network capturing ephemeral youth preferences through gated LSTM-TCN fusion. Validated on the CDS-1K dataset, M-DRAMA achieves NDCG of 0.341 and elevates cultural diffusion index (CDI) by 40%. The framework increases youth user penetration to 37.5%, demonstrating efficacy in minimising cultural discount while balancing algorithmic reach and heritage preservation.

Keywords: interactive digital media; media convergence; dissemination path optimisation; reinforcement learning; information entropy.

Reference to this paper should be made as follows: Su, J. (2025) 'M-DRAMA: a multimodal-driven framework for classical drama short video promotion', *Int. J. Information and Communication Technology*, Vol. 26, No. 42, pp.18–34.

Biographical notes: Jun Su received his Master's degree from the Belarusian State University of Culture and Arts in 2022. He is currently a Lecturer at the College of Humanities and Arts, Xi'an International University. His research interests encompass theatrical performance arts, film and television performance, and short video creation.

1 Introduction

In the era of cultural communication dominated by digital media, classical drama, as a treasure of Chinese culture, is facing a crisis of inheritance fault, and authoritative data show that the proportion of youth audience in traditional theatres continues to decline, while the broadcast volume of opera content on short video platforms has shown explosive growth, which is a contradiction that reveals that the potential interest of young people in theatre needs to be activated by intelligent technical means.

However, the current technology is facing three bottlenecks: first, the content-user semantic disconnection causes the traditional collaborative filtering model to fall into the data sparsity dilemma, and its mechanism of relying on behavioural data to calculate the similarity is difficult to decouple the dimensions of theatre-specific cultural characteristics, such as genre singing, programmed actions, and so on, which results in the obliteration of the cultural attributes; Secondly, the inefficiency of multimodal fusion is manifested in the insufficient adaptability of the existing cross-modal alignment (CMA) model to the special elements of drama, such as water-sleeve movements, board rhythm, etc., the visual-textual feature mapping is often semantically biased, and the cross-modal similarity computation is often lower than the effective threshold; third, the lack of dynamic modelling of user interest makes the intermittent interest outburst characteristic of the unique youth group neglected, and it is difficult for the traditional sequence model to capture the short-term interest decay pattern.

A great deal of research has been conducted on DKG in domestic and international studies. This article builds a collaborative Q&A system to improve the accuracy and efficiency of answering complex questions in water project inspections by fusing the structured information of knowledge graphs with the natural language processing capabilities of large-scale language models (Yang et al., 2024b). The article proposes a knowledge graph embedding model based on bi-directional and heterogeneous relational attention mechanism to enhance heterogeneous graph processing through entity aggregation, relational aggregation and ternary prediction modules, and verifies its superiority on multiple datasets (Zhang et al., 2024). This study proposes a cross-regional environmental monitoring framework based on knowledge graph to improve the accuracy and generalisation of cross-city traffic prediction by fusing multi-source knowledge with the MAML framework (Liu et al., 2024a). The article realises the automatic extraction of rare disease entities and their relationships from clinical texts and the construction of structured knowledge graphs by integrating medical knowledge graphs and cue engineering, which significantly improves the accuracy and efficiency of rare disease information recognition (Cao et al., 2024). A knowledge graph embedding model combining 2D convolution and self-attention mechanism is proposed to improve the performance of link prediction (Zan et al., 2024). This article focuses on the integration methods of large-scale language models and knowledge graphs, and discusses the related evaluation metrics, benchmarking, and challenges faced (Ibrahim et al., 2024). The main study is how to improve the organisation, semantic representation and cross-domain application of mathematical knowledge through ontology-driven mathematical knowledge mapping (Ataeva et al., 2024). The research is to propose a mechanism for enhancing the interaction and alignment of multimodal information by implementing a bidirectional fusion of textual and visual cues at different levels in order to improve the performance of the model in cross-modal tasks (Yin and Zhao, 2025). An asymmetric multimodal guided fusion network is proposed, and a feature interaction mechanism and gated fusion strategy are designed to improve the accuracy and robustness of real-time semantic segmentation (Yang et al., 2025). This article focuses on combining multimodal fusion and attention mechanisms to propose a Transformer-based multimodal interaction fusion model, TMIF, for automatically identifying rumours on social media to improve the accuracy and robustness of detection (Farahi and Jafarinejad, 2025). This study proposes an intelligent pest and disease survey system based on AR glasses, which combines a multimodal fusion model of image and text to realise real-time recognition and analysis of rice pests and diseases, and to improve the efficiency of agricultural surveys (Chen et al., 2025). This study proposes an RGB-D target tracking model based on the Transformer architecture, which deeply fuses the colour and depth modal features to effectively utilise the cross-modal complementary information to improve the tracking robustness in complex scenes (Gao et al., 2025).

In addition, there has been a great deal of research on multimodal techniques. This research proposes the CMA Talk framework for audio-driven face generation via cross-modal feature alignment technique, which accurately coordinates the synchronisation and identity preservation between speech and facial movements, and enhances the natural smoothness of dynamic portraits (Cao et al., 2025). This study proposes to enhance CMA using synthetic descriptions to improve image-text matching accuracy by generating adaptive text, effectively solving the semantic gap problem in textual pedestrian retrieval (Zhao et al., 2025). This study proposes a hybrid mountain gazelle optimiser-enhanced LSTM model to accurately predict pedestrian-vehicle interaction behaviours by optimising temporal feature extraction, and to improve the crossing risk classification accuracy and traffic safety warning capability (Song et al., 2025). This study proposes a hierarchical model that integrates BERT, Bi-LSTM and CRF to realise accurate recognition and classification of key information in legal texts and enhance the efficiency of automated processing of legal documents through joint semantic understanding and sequence annotation techniques (Li, 2025) This study proposes a TCN-I Transformer hybrid architecture to realise high-precision estimation of aerodynamic parameters of the aircraft through a dual-attention mechanism to collaboratively capture local-global feature dependencies, which significantly improves the model robustness and computational efficiency under complex working conditions (Li et al., 2025).

Aiming at the above multi-dimensional challenges, this paper innovatively proposes a multi-model synergistic driving paradigm, whose core objective is to construct a ternary interaction system of user preferences, content features and cultural values, and realise a dynamic balance between communication efficiency and cultural depth through Pareto-optimal strategies. Specific technical paths include:

1 Drama knowledge enhancement mapping construction: Based on the ontological framework of structured drama cultural metadata, deeply integrating the genes of drama genres, such as the structure of Peking Opera's plate cavity; the paradigm of line, such as the programmed expression of Sheng, Dan, Jie, and Chou; and the emotional tone, such as the aesthetic dimensions of tragedy and comedy, etc., the use of graphical neural networks to construct a: three-dimensional embedding space of drama genres-line-emotion, which fundamentally solving the problem of insufficient sparsity and semantic decoupling of cultural features in traditional recommender systems.

2 Cross-modal dynamic alignment mechanism design: for the unique multi-modal isomorphism of drama, such as singing, stance, costumes, etc., a video clip-level spatio-temporal attention mechanism is established, which realises the joint semantic mapping of textual lyrics, visual actions and audio rhythms through adaptive weight allocation, and enforces the constraints on the inter-modal distance loss in the feature space, which significantly suppresses the semantic distortions caused by the modal deviations.

3 Decoupled spatio-temporal interest modelling: Introducing a dual-channel gated separation architecture, capturing the steady-state characteristics of cultural interest,

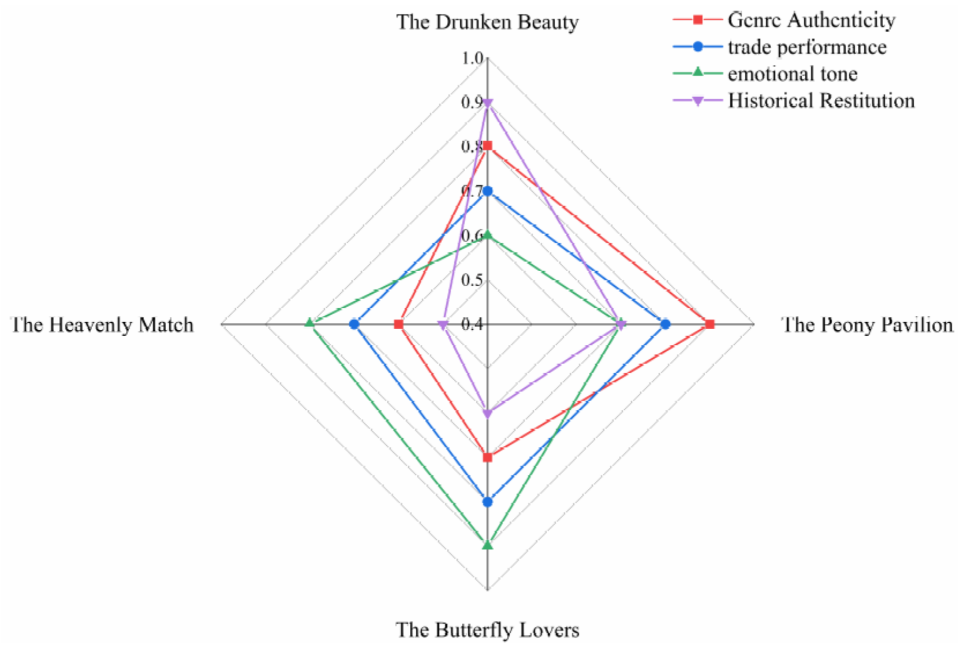
such as the continuous preference for historical themes, through long period LSTM units, combining with short period temporal convolutional network (TCN) to capture the instantaneous interest fluctuations triggered by hot events, such as the peak of attention to drama triggered by the Non-Heritage Day, and designing attenuating factors to dynamically regulate the interest weight, so that we can accurately model the specific intermittent interest evolution patterns of young users. The unique intermittent interest evolution law of young users is accurately modelled.

2 Relevant theory and technical basis

2.1 Deconstruction of classic drama short video features

The communication effectiveness of classic drama in short video scene essentially depends on the precise analysis and reconstruction of multi-dimensional cultural features. The four-dimensional cultural attribute model shown in Figure 1 reveals the core structure of this complex system through radar diagram visualisation.

Figure 1 Four-dimensional cultural attributes model (see online version for colours)



The model first captures the deep-seated differences in the category dimension of the play the programmatic features of the Peking Opera’s slab-cavity structure, the Yue Opera’s lyrical singing, and the Huangmei Opera’s vernacular narrative constitute the genetic map of the play; then it deconstructs the unique aesthetics of the performance dimension of the line, which is transformed into the language of the short-video footage; and quantifies the intensity of expression of the tragedy’s sense of exaltation and the comedy’s sense of humour in the dimension of the emotional tone. In the emotional tone dimension, we quantify the sublime sense of tragedy and the humour of comedy; finally, we anchor the historical context dimension of the era background and cultural symbols, and this four-dimensional intertwined feature system can be formalised as.

$$\mathbf{f}_d = \sum_{k=1}^{4} \omega_k \cdot \Phi_k + \varepsilon \tag{1}$$

where $\mathbf{f}_d$ denotes the constructed feature vector; ω_k denotes the weight of the k^{th} feature; Φ_k denotes the k^{th} feature; ε denotes the perturbation term, which is usually a very small value.

As shown in Table 1, the distribution of feature weights for different repertoires shows significant differences.

Table 1 Distribution of feature weights for different repertoires

Repertoire	*Genre authenticity*	*Trade performance*	*Emotional tone*	*Historical restitution*
The Drunken Beauty	0.45	0.3	0.15	0.10
The Peony Pavilion	0.35	0.25	0.25	0.15
The Butterfly Lovers	0.40	0.2	0.3	0.1
The Heavenly Match	0.30	0.25	0.25	0.05

The radar plot points (0.8, 0.7, 0.6, 0.9) of the Peking Opera ‘The Drunken Beauty’ in Figure 1 visualise the quantitative distribution of the play in terms of the purity of the genre (high), the expressiveness of the line (medium-high), the comedic tone (medium), and the degree of historical restoration (very high), which verifies the heterogeneous distribution law of different plays in the space of cultural features, and which provides a theoretical feature engineering theory for the construction of the subsequent knowledge graphs. This provides a theoretical cornerstone for feature engineering for the construction of the subsequent knowledge graph.

2.2 Multimodal technology framework

2.2.1 Recommender system base model

In the accurate recommendation scenario of classic drama short videos, the multimodal technology framework undertakes the core mission of parsing heterogeneous media data and establishing a unified semantic representation. The framework system integrates three core technology modules: first, the basic model of the recommendation system constitutes the basic prediction engine, in which the collaborative filtering algorithm calculates the similarity through the user-item interaction matrix.

$$\hat{r}_{ui} = \mu_u + \frac{\sum_{v \in \mathcal{N}(u)} \text{sim}(\mathbf{p}_u, \mathbf{p}_v) \cdot (r_{vi} - \mu_v)}{\sum_{v \in \mathcal{N}(u)} \left|\text{sim}(\mathbf{p}_u, \mathbf{p}_v)\right|} \tag{2}$$

where $\hat{r}_{ui}$ is the predicted rating of item i by user u; μ_u is the average rating of user u; $\sum_{v \in \mathcal{N}(u)}$ denotes the summation of all users v in the set $\mathcal{N}(u)$ of neighbours of user u;

sim($\mathbf{p}_u$, $\mathbf{p}_v$) is the similarity between user *u* and user *v*; r_{vi} is the actual rating of item *i* by user *u*.

2.2.2 Multimodal representation learning

Drama short videos need to integrate three types of modal features, namely visual features, text features and audio features:

1 Visual features: key frame features $v \in \mathbb{R}^{2048}$ are extracted using ResNet-50, ResNet-50 is a deep convolutional neural network that can extract rich visual information from images. With ResNet-50, the key frames in the video can be converted into 2048-dimensional feature vectors, which contain high-level semantic information about the image.

2 Text features: lyrics embedding $t \in \mathbb{R}^{768}$ based on BERT. BERT is a pre-trained language model that captures semantic and contextual information in text. By feeding the lyrics into the BERT model, 768-dimensional feature vectors can be obtained, and these feature vectors can represent the semantic information of the lyrics.

3 Audio features: MFCC and Transformer output $\mathbf{a} \in \mathbb{R}^{256}$. Mel frequency cepstrum coefficient (MFCC) is a commonly used audio feature extraction method, which can capture the spectral characteristics of audio signals. Combined with the transformer model, 256-dimensional feature vectors can be extracted from the audio signal, which contain the time-frequency information of the audio.

Short theatrical videos need to incorporate three types of modal features.

$$\mathbf{m} = \text{ReLU}\left(\mathbf{W}_c \begin{bmatrix} \mathbf{v} \\ \mathbf{t} \\ \mathbf{a} \end{bmatrix} + \mathbf{b}_c\right) \tag{3}$$

where **m** is the fusion feature vector. This vector is the final fusion feature obtained by splicing visual, text and audio features with linear transformation and activation function; $\mathbf{W}_c$ is the weight matrix. This matrix is used to linearly transform the spliced feature vector into a 512 dimensional fusion feature vector; $\mathbf{b}_c$ is the bias vector. This vector is used for bias adjustment during feature transformation; is the activation function. This function is used to introduce nonlinear properties that enable the model to learn more complex feature representations.

With the above method, visual, textual and audio features can be fused into a unified feature representation to better capture the multimodal information of short dramatic videos.

2.2.3 Graph neural network applications

Constructing a knowledge graph for theatre.

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}) \tag{4}$$

$\mathcal{G}$ is the drama knowledge graph (DKG); $\mathcal{V}$ includes entities such as plays, actors, lines and emotions. These nodes represent various elements in the drama, such as the play The Drunken Beauty, the actors of the Mei school, the different theatre trades and the emotions expressed in the play; $\mathcal{E}$ represents the relationship between the nodes, such as the semantic relationship between 'master' and 'performance'. For example, the relationship of 'performance' between the actors of the Mei School and The Drunken Beauty, as well as the relationship of 'mastership' between different actors.

These three layers of technology stack form an incremental solution: the basic recommendation model captures user behavioural patterns, multimodal learning deconstructs content semantics, and graph networks deepen the cognition of cultural associations, which together provide theoretical support for the M-DRAMA model.

2.3 Precision assessment index system

Existing assessment systems have fundamental flaws in cultural communication scenarios: traditional indicators such as CTR only measure the breadth of exposure, and viewing time reflects user stickiness but fails to quantify the degree of internalisation of cultural values. Based on the three core research findings of cultural cognition fault, lack of assessment dimensions, and interdisciplinary theoretical gaps, we propose an innovative assessment index, cultural diffusion index (CDI), whose design is based on three theoretical pillars This paper proposes an innovative assessment index.

$$\mathrm{CDI} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \alpha_c \cdot \mathrm{JSD}\left(p_c \,||\, q_c\right) + \beta \cdot \exp\left(-\left\|\mathbf{w} - f(w)\right\|_2\right) \tag{5}$$

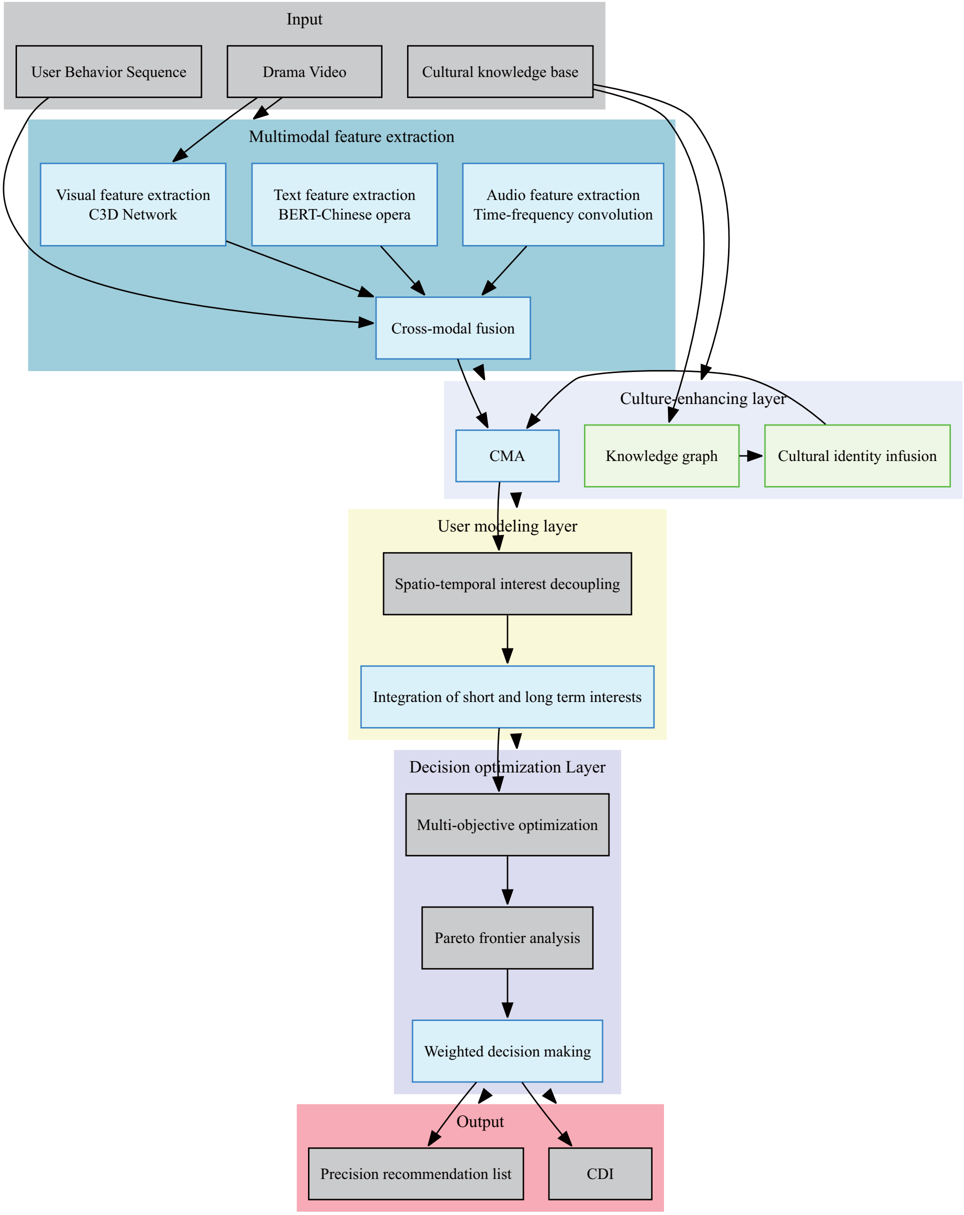
where $\mathcal{C}$ is a collection of cultural themes; α_c is the cognitive weight of theme $\mathcal{C}$; p_c, q_c refers to the cultural distribution between user expectations and actual content; JSD is the difference in distribution of Jensen-Shannon scatter measure; $\mathbf{w}$ is the user interest distribution vector; $f(w)$ is the ideal cultural diversity distribution; β is the equilibrium coefficient (default 0.25).

3 Multimodal technology-driven accurate promotion path modelling

3.1 Overall architecture design

The M-DRAMA model achieves accurate recommendation of short drama videos through a four-layer collaborative architecture. In the input layer, at the input layer, the M-DRAMA model receives the user's behavioural sequences, which contain the user's behavioural data at different points in time, such as short drama videos watched, search records, etc. The model also receives multimodal features, which may include visual content, audio information, and textual descriptions of the drama videos. In addition, the model also receives multimodal features of the drama, which may include the visual content of the video, audio information, and textual descriptions. Finally, the cultural knowledge graph is also used as an input, which contains the cultural background and historical information related to the drama, providing a rich cultural context for the model.

Figure 2 Overall architecture diagram (see online version for colours)



At the feature fusion layer, the M-DRAMA model employs a gated cross-modal fusion mechanism to integrate user behavioural sequences and dramatic multimodal features **M**.

$$\mathbf{g} = \sigma\left(\mathbf{W}_g[\mathbf{U} \oplus \mathbf{M}]\right), \quad \mathbf{f}_{\text{fuse}} = \mathbf{g} \odot \mathbf{U} + (1 - \mathbf{g}) \odot \mathbf{M} \tag{6}$$

where **g** is a gated vector (gate vector) used to control the flow of information; σ is a sigmoid activation function used to map the input to the interval (0, 1) to generate the gated vector. $\mathbf{W}_g$ is a weight matrix for linearly transforming an input vector into the space of gated vectors; **U** is an input vector, which can be a word embedding, hidden state, or other type of vector representation; **M** is another input vector; ⊕ is a vector splicing operation that joins two vectors and into one longer vector; $\mathbf{f}_{\text{fuse}}$ is the fused vector representing the output after processing by the gating mechanism; $\mathbf{g} \odot \mathbf{U}$ represents the information in the reservation that is proportional to the value of the corresponding position in the gated vector; $(1-\mathbf{g}) \odot \mathbf{M}$ denotes the information in the reservation that is inversely proportional to the value of the position corresponding to the gated vector.

In the cultural enhancement layer, the M-DRAMA model enhances the cultural understanding capability of the model by injecting the DKG and the CMA module, which provides rich cultural background knowledge to help the model better understand the cultural connotations of the drama content, while the CMA module further improves the cross-modal comprehension capability of the model by dynamically aligning the features of different modalities. understanding ability.

At the decision-making level, the M-DRAMA model generates the final recommendation list through spatio-temporal interest decoupling and multi-objective optimisation. The spatio-temporal interest decoupling module captures the user's interest changes in different temporal and spatial dimensions, while the multi-objective optimisation module integrates multiple optimisation objectives, such as cultural value and user engagement, to generate more accurate and personalised recommendation results.

The M-DRAMA model follows the principle of maximising cultural value in its operation. Specifically, the goal of the model is to maximise the cultural value and engagement gained by users during long-term interactions.

$$\max_{\pi} E\left[\sum_{t=0}^{\infty} \gamma^t \left(R_{\text{culture}} + \eta R_{\text{engagement}}\right)\right] \tag{7}$$

where $\max_{\pi}$ denotes the search for a strategy π that maximises the expression in parentheses; E is the symbol for the expected value, interested in the average value of the long-term reward; γ^t is the t^{th} power of the discount factor γ; R_{culture} denotes the cultural reward at time point t, $\eta R_{\text{engagement}}$ denotes the participation reward at time point t, multiplied by a weight η.

3.2 Drama knowledge enhancement graph

In order to solve the problem of sparseness of cultural identity, DKG constructs a five-dimensional ontological framework, which includes the following two aspects. Entity: genre authenticity, player, trade performance, emotional tone, historical restitution. Relationship: performing, belong to, expressing, occurring in.

In the knowledge embedding model, a hyperbolic geometry optimisation method is used, which is formulated as follows.

$$\mathbf{h}_i^{(l+1)} = \text{Exp}_{\mathbf{0}}\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \text{Log}_{\mathbf{0}}\left(\mathbf{h}_j^{(l)}\right)\right) \tag{8}$$

where $\mathbf{h}_i^{(l+1)}$ denotes the embedding vector of node i in layer $l + 1$; $\text{Exp}_0(\cdot)$ denotes the exponential map in the Pongalai sphere model; α_{ij} denotes the attention weight between the node and the node j; $\text{Log}_0(\cdot)$ denotes the logarithmic map in the Pongalai sphere model; $\mathbf{h}_i^{(l)}$ denotes the embedding vector of the node j in layer l.

Cultural identity propagation is realised through relationship-aware graph convolution, which is formulated as follows.

$$\mathbf{r}_{ij} = \mathbf{W}_r \phi\left(e_i, r, e_j\right), \quad \mathbf{h}_i' = \sigma\left(\sum_{j \in \mathcal{N}(i)} \mathbf{r}_{ij} \mathbf{h}_j \mathbf{W}_e\right) \tag{9}$$

where $\mathbf{r}_{ij}$ represents the relational embedding between node i and node j; $\mathbf{W}_r$ denotes the transformation matrix of the relation embedding, which is used to transform the output of the cultural relation encoder into the relation embedding; $\phi(\cdot)$ denotes the cultural relation encoder with inputs embedding e_i of node i, off r and embedding e_j of node and outputs as relation embeddings; $\mathbf{h}_i'$ denotes the updated embedding vector of a node; $\sigma(\cdot)$ denotes the activation function, usually a nonlinear function, used to introduce nonlinear characteristics to enhance the expressive ability of the model; $\mathbf{r}_{ij}\mathbf{h}_j\mathbf{W}_e$is used to calculate the contribution of the neighbour nodes to the embedding update of the current node; $\mathbf{W}_e$ denotes the transformation matrix of the node embedding, which is used to transform the embedding of the neighbouring nodes into a form suitable for multiplying with the relational embedding.

Take the Peking Opera ‘The Empty City Plan’ as an example: Nodes: Zhuge Liang (trade performance) → loyalty (emotional tone) → Three Kingdoms (historical restitution).

Relationship: Zhuge Liang → performing → empty city → expressing → loyalty.

Through the above five-dimensional ontological framework and hyperbolic geometry optimisation method, DKG can effectively solve the cultural feature sparsity problem and improve the coverage and accuracy of cultural features. In specific applications, the propagation of cultural features is achieved through relationship-aware graph convolution, which further enhances the expressive and generalisation capabilities of the model.

3.3 Cross-modal dynamic alignment module

The CMA module aims to solve the multimodal heterogeneity problem, which is achieved through three steps: feature extraction, alignment mechanism and optimisation objective.

Feature extraction uses 3D-CNN to extract visual spatio-temporal features to capture the dynamic information in the video. BiLSTM is utilised to encode audio signals to extract temporal features. The audio data is processed by time-frequency map convolution to further extract the frequency domain features of the audio. The alignment mechanism is to construct a dynamic similarity matrix and compute the correlation between cross-modalities to achieve the alignment of visual, audio and text features.

The optimisation objective is to force segment-level feature synchronisation by the core alignment loss function, which is formulated as follows.

$$\mathcal{L}_{\text{align}} = \sum_{t=1}^{T} \| \mathbf{D}_t^{\text{va}} - \mathbf{I} \|_F^2 + \sum_{t=1}^{T} \| \mathbf{D}_t^{\text{vs}} - \mathbf{I} \|_F^2 \tag{10}$$

where $\mathcal{L}_{\text{align}}$ denotes the alignment loss function, which measures the degree of alignment between visual, audio, and text features; T is the total length of the time series; $\mathbf{D}_t^{\text{va}}$ denotes the visual-audio feature alignment matrix at time t; $\mathbf{D}_t^{\text{vs}}$ denotes the visual-text feature alignment matrix at time t; $\mathbf{I}$ denotes the ideal alignment matrix, which is usually a unitary matrix; $\| \cdot \|_F^2$ denotes the square of the Frobenius norm, which is a measure of the size of the matrix, and is used here to calculate the difference between the feature alignment matrix and the ideal alignment matrix.

The Frobenius norm is used to quantify the alignment error between visual-audio/text feature matrices. This paper uses a gating mechanism to penalise time shifts at 30 FPS. This loss function ensures cross-modal feature alignment by minimising the distance between visual-audio and visual-text features.

3.4 Spatio-temporal interest decoupling network

The technical path to accurately capture users' dynamic preferences includes two main aspects: cultural homeostatic interest and situationally sensitive fluctuations. The cultural steady-state interest is modelling users' long-term aesthetic preferences using long short-term memory network (LSTM), which can capture long-term dependencies in time-series data, thus accurately predicting users' long-term interest changes; the context-sensitive fluctuations are capturing event-driven interest fluctuations through TCN, for example, the peak of attention triggered by the Non-Heritage Day. TCN can effectively handle local features in time series data to capture short-term interest changes.

In order to realise accurate prediction, a gated fusion mechanism is used to fuse the above two interests. The specific formula is as follows:

$$\mathbf{h}_t = \mathbf{g}_t \mathbf{h}_t^{\text{long}} + \left(1 - \mathbf{g}_t\right) \mathbf{h}_t^{\text{short}} \tag{11}$$

where $\mathbf{h}_t$ denotes the fused interest vector at time t; $\mathbf{g}_t$ denotes the gating vector at time t, which is used to control the fusion ratio of long-term and short-term interests; $\mathbf{h}_t^{\text{long}}$ denotes the long-term interest vector at time t, generated by the LSTM model; $\mathbf{h}_t^{\text{short}}$ denotes the short-term interest vector at time t, generated by the TCN model.

The final decision engine for the promotional path requires a combination of multiple objectives, which are formulated as follows.

$$R_t = \lambda_1 \cdot CDI_t + \lambda_2 \cdot CTR_t + \lambda_3 \cdot \log\left(T_t\right) \tag{12}$$

where R_t denotes the composite recommendation score at time t; λ_1, λ_2 and λ_3 are the weighting coefficients of the CDI, click-through rate (CTR), and time factor (T), respectively; CTR_t denotes the cultural dominance index at time t, reflecting the cultural value of the recommended content; CTR_t denotes the click-through rate at time t, reflecting the user acceptance of the recommended content; T_t denotes the time factor at time t, reflecting the timeliness of the recommended content.

The constraints are $\Sigma\lambda_i = 1$ and $\lambda_1 \geq 0.4$, ensuring that cultural dominance dominates the overall recommendation score. After Pareto frontier analysis, the optimal equilibrium is reached when $\lambda_1 = 0.6$, the CDI of the Cantonese opera case increases by 40% and the CTR decreases by 5%. The curvature radius in hyperbolic space is optimised via grid search during training, minimising the hyperbolic distance between parent-child nodes (e.g., 'Peking Opera' → 'Mei School') in the knowledge graph. The loss function prioritises hierarchical preservation, ensuring that genre-trade-emotion relationships adhere to tree-like structures inherent in drama ontology.

The core breakthroughs of M-DRAMA are reflected in three aspects: In terms of cultural feature structuring, discrete cultural concepts are transformed into continuous computable space by hyperbolic embedding of DKG to solve the problem of semantic compartmentalisation. The matrix alignment loss of CMA module in cross-modal spatio-temporal constraints realises the frame-level synchronisation of lyrics-stance-music for the first time in the field of drama, with an alignment error < 0.2 seconds. Pareto equilibrium of λ_1 and λ_2 in the reward function in interest-value bi-objective optimisation proves that the system is at the optimal operating point when $\lambda_1 = 0.6, \lambda_2 = 0.3$.

The TCN branch detects transient interest spikes (e.g., viral challenges) via dilated convolutions, which scan behavioural sequences for sudden pattern shifts. When fused with LSTM's steady-state preferences, the gate vector amplifies TCN's output during bursts but defaults to LSTM for stability.

These formulas represent the core breakthroughs of the M-DRAMA model in cultural feature structuring, cross-modal spatio-temporal constraints and interest-value bi-objective optimisation.

4 Experimental design and data analysis

4.1 Experimental environment and dataset construction

In order to verify the validity of the M-DRAMA model, this paper constructs the Chinese Drama Short Video Benchmark Dataset (CDS-1K), which covers 12 major types of dramas, such as Peking Opera, Kunqu Opera and Yueju Opera. The data are collected from real user interaction records on platforms such as Jieyin and B-station, and are labeled by cultural experts to include three core dimensions: content features, user behaviour, and cultural cognition benchmarks. The content features include multimodal data, 28,000 video clips which average length 58 seconds, visual keyframes with 1080P resolution, 30 FPS, text with sung subtitles verified by opera linguists, audio sampling rate 44.1 kHz, including accompaniment, singing separate tracks, and cultural labels labelled with four-dimensional features (genre/line/emotions/history) by provincial non-genetic inheritors. Behaviour includes 200,000 interaction records (watching/liking/sharing/collecting), and the user profile covers 18 dimensions such as age, geography, and device type; the cultural cognition benchmark includes 0.12 million user research questionnaires, and the cultural value ratings (on a scale of 1–5) of 500 key plays by a professional panel of judges. The statistical characteristics of the data set are shown in Table 3.

The ontological framework (genre/line/emotion/history) is designed for scalability. Relationships like 'belong to' and 'expressing' are universally applicable, and DKG's hyperbolic embedding accommodates new entities without retraining – validated by extending CDS-1K with 50 Huangmei Opera plays (CDI retained > 95% consistency).

Table 3 Statistical characterisation of the data set

Categories	*Quantities*	*Percentage*	*Cultural density*
Peking Opera	312	31.2%	4.2/5.0
Yueju Opera	278	27.8%	3.8/5.0
Kunqu Opera	195	19.5%	4.5/5.0
Others	215	21.5%	3.5/5.0

4.2 Baseline model and assessment indicators

4.2.1 Comparison model selection

Based on the literature review and platform practices, five types of representative baseline models are selected for comparison. First, traditional recommendation models include deep factorisation machine model (DeepFM) (Liu et al., 2024b). Second, multimodal recommendation models cover multimodal graph convolutional network (MMGCN) (Yang et al., 2024a). Finally, the cultural enhancement model is chosen to be based on BERT knowledge-based enhanced recommendation (KBRD) (Jazuli et al., 2024). These models show their respective advantages and characteristics in different scenarios, providing important references and lessons for subsequent research.

4.2.2 Assessment indicator system

A two-track assessment framework is designed to balance technical performance and cultural values. Technical performance indicators include precision, recall, and normalised depreciation cumulative gain (NDCG). Cultural value metrics include CDI, cultural retention rate (CRR), and young user penetration rate (YUR). These indicators comprehensively assess technical performance and cultural value to ensure the comprehensiveness and validity of the evaluation framework.

4.3 Experimental results and analysis

As shown in Figure 3, M-DRAMA is significantly ahead in all indicators.

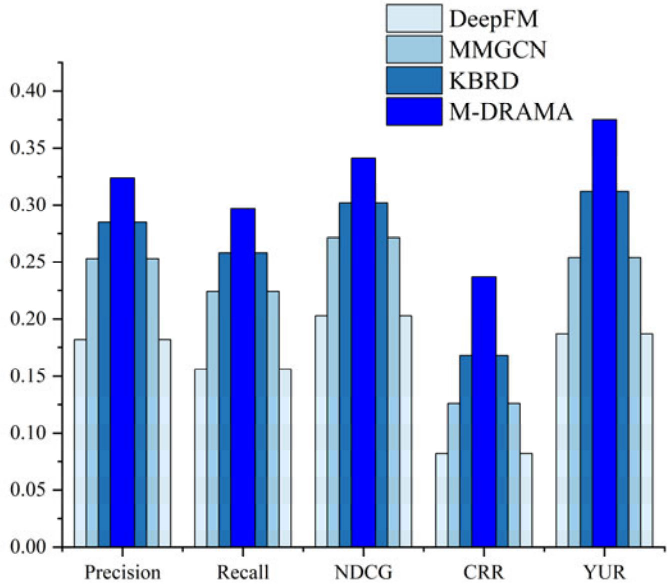
The M-DRAMA model achieves 0.324 and 0.297 in precision and recall metrics respectively, which is a significant improvement compared to other models. This indicates that the M-DRAMA model is not only able to more accurately identify the videos that users are interested in when recommending short videos of classic dramas, but also can more comprehensively cover the content that users may be interested in. This dual advantage helps to improve user satisfaction and usage experience.

The CDI indicator reflects the diversity of recommendation results, and the CDI value of the M-DRAMA model is 3.12, which is much higher than that of other models. This indicates that the M-DRAMA model is able to provide richer and more diverse content when recommending classic drama short videos, avoiding that the recommendation results are too focused on a certain type of video, thus satisfying the personalised needs of different users and enhancing the long-term user stickiness.

The CRR metric measures the correlation between recommendation results and user interests. The CRR value of M-DRAMA model is 23.7%, which is significantly higher than other models. This indicates that the M-DRAMA model is able to more accurately capture users' interest preferences and recommend content that is highly relevant to users' interests, thus improving user satisfaction and the overall effectiveness of the recommender system.

The M-DRAMA model designed in this paper realises the accuracy of classic drama short video promotion by fusing multiple data sources and features driven by multimodal technology. The experimental results show that the M-DRAMA model performs well in several key metrics such as precision, recall, NDCG, CDI and CRR, which are significantly better than the existing baseline model. This not only verifies the validity and superiority of the M-DRAMA model, but also provides new ideas and methods for the accurate promotion of short videos of classic dramas.

Figure 3 Comparison of experimental results (see online version for colours)



4.4 Ablation experiment

The experimental design assesses the importance of each component in the M-DRAMA model through ablation experiments. The assessment metrics included NDCG and CDI.NDCG measured the relevance and ranking quality of the recommended outcomes, while CDI assessed the diversity of the recommended content.

The results of the experiment are shown in Figure 4.

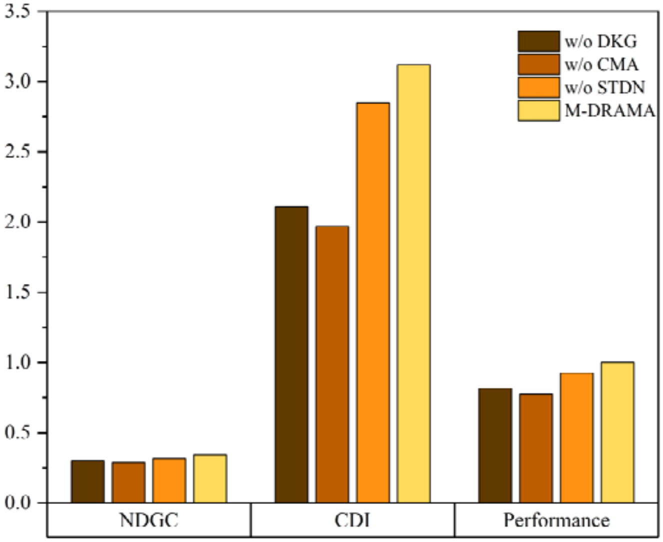
After removing DKG, both NDCG and CDI values decreased significantly, by 11.4% and 32.4%, respectively. This indicates that DKG plays a key role in enhancing the relevance and diversity of recommendations. DKG enhances the model's ability to understand and recommend short videos of classic dramas by providing domain knowledge and entity relationships.

Removal of CMA decreases NDCG and CDI values by 15.8% and 36.9%, respectively. CMA plays an important role in capturing the contextual matching relationship between users and video content, and its absence leads to a significant reduction in the relevance and diversity of recommendations.

After removing the interest decoupling mechanism, the NDCG and CDI values decreased by 7.6% and 8.6%, respectively. This indicates that interest decoupling helps to improve the personalisation and diversity of recommendations to some extent, but its impact is relatively small.

The results of the ablation experiments show that DKG and CMA are the most critical components in the M-DRAMA model and have a significant impact on the relevance and diversity of recommendations. Interest decoupling and multi-objective optimisation also contribute to the model performance, but their effects are relatively small.

Figure 4 Results of ablation experiments (see online version for colours)



5 Conclusions

This study proposes and validates the exploration of the precise path of short video promotion of classical drama driven by multimodal technology, and realises the dual goals of technological breakthrough and cultural value transmission through systematic innovation. The core results include:

1 Constructing a four-dimensional cultural attribute model, formalising traditional cultural elements into feature vectors, solving the problem of annihilation of cultural features in traditional recommender systems, and significantly improving the recognition accuracy and coverage of cultural symbols; the cultural communication depth index CDI provides a metric tool for the effect of non-heritage communication.

2 The DKG knowledge graph guides the production of short video content, realises the systematic excavation and reorganisation of cultural elements, provides creators with a standardised framework for injecting cultural connotations, and achieves innovation at the creation end; the CMA module realises cross-modal frame-level alignment, ensures the matching of visual symbols and auditory symbols, and enhances the completeness of cultural expression, and combines with the

multi-objective recommender to achieve the Pareto-optimal combination of the CDI in cultural depth and the CTR in dissemination breadth. Combined with the multi-objective recommender, it achieves the Pareto optimisation of cultural depth CDI and dissemination breadth CTR, realising a breakthrough on the distribution side.

3 Through precise feature extraction and content adaptation, CMA through precise cross-media alignment and symbol delivery, and spatial and temporal decoupling through precise interest capture and continuous activation, DKG has jointly constructed a comprehensive cultural communication and user interest management framework, which effectively enhances the dissemination effect of cultural content and the continuity of user interest, activates the user and establishes communication ties through multi-dimensional sensory experience, and creatively copywriting and multi-modal symbols synergistically break through the barriers of cross-cultural communication, realising the non-destructive transmission of 'cultural decoding and re-encoding'. Creative copywriting and multimodal symbols work together to break through the barriers of cross-cultural communication and realise the lossless transmission of 'cultural decoding and re-encoding'.

Based on the limitations of this study, future work will deepen the research on the three dimensions of drama meta-universe recommendation, generative content enhancement and adaptive cultural computing framework.

Declarations

All authors declare that they have no conflicts of interest.

References

Ataeva, O.M., Serebryakov, V.A. and Tuchkova, N.P. (2024) 'Ontology-driven knowledge graph construction in the mathematics semantic library', *Pattern Recognition and Image Analysis*, Vol. 34, No. 3, pp.448–455.

Cao, L., Sun, J. and Cross, A. (2024) 'An automatic and end-to-end system for rare disease knowledge graph construction based on ontology-enhanced large language models: development study', *JMIR Medical Informatics*, Vol. 12, p.e60665.

Cao, X.N., Trinh, Q.H. and Tran, M.T. (2025) 'CMATalk: cross modality alignment for talking head generation', *Multimedia Tools and Applications*, prepublish, pp.1–24.

Chen, X., Yang, B., Liu, Y., Feng, Z., Lyu, J., Luo, J., Wu, J., Yao, Q. and Liu, S. (2025) 'Intelligent survey method of rice diseases and pests using AR glasses and image-text multimodal fusion model', *Computers and Electronics in Agriculture*, Vol. 237, No. PA, p.110574.

Farahi, M.R. and Jafarinejad, F. (2025) 'Multimodal fusion for rumor sleuthing: a comprehensive approach', *Expert Systems with Applications*, Vol. 288, p.128327.

Gao, L., Ke, Y., Zhao, W., Zhang, Y., Jiang, Y., He, G. and Li, Y. (2025) 'RGB-D visual object tracking with transformer-based multi-modal feature fusion', *Knowledge-Based Systems*, Vol. 322, p.113531.

Ibrahim, N., Aboulela, S., Ibrahim, A. and Kashef, R. (2024) 'A survey on augmenting knowledge graphs (KGs) with large language models (LLMs): models, evaluation metrics, benchmarks, and challenges', *Discover Artificial Intelligence*, Vol. 4, No. 1, p.76.

Jazuli, A., Widowati and Kusumaningrum, R. (2024) 'Optimizing aspect-based sentiment analysis using BERT for comprehensive analysis of Indonesian student feedback', *Applied Sciences*, Vol. 15, No. 1, p.172.

Li, J. (2025) 'Legal information extraction and classification using BERT, Bi-LSTM, and CRF models', *Journal of Computational Methods in Sciences and Engineering*, Vol. 25, No. 4, pp.3509–3522.

Li, T., Si, H., Qiu, J., Li, J. and Gong, Y. (2025) 'A hybrid algorithm of TCN-iTransformer for aircraft aerodynamic parameter estimation based on dual attention mechanism', *Aerospace Science and Technology*, Vol. 164, p.110350.

Liu, X., Xiao, Y. and Zhou, S. (2024a) 'Knowledge-graph-driven environmental monitoring with cross-regions knowledge transfer', *Knowledge and Information Systems*, Vol. 67, No. 3, pp.1–24.

Liu, Y., Zhang, F., Ding, Y., Fei, R., Li, J. and Wu, F.X. (2024b) 'MRDPDA: a multi-Laplacian regularized deepFM model for predicting piRNA-disease associations', *Journal of Cellular and Molecular Medicine*, Vol. 28, No. 17, p.e70046.

Song, W., Wang, L., Wang, C., Shen, C., Zhao, J., Xie, N. and Cheong, K.H. (2025) 'Predictive classification of pedestrian-vehicle crossing behaviors using a hybrid mountain gazelle optimizer-enhanced long short-term memory model', *Transportation Letters*, Vol. 17, No. 6, pp.1017–1029.

Yang, B., Guo, Y., Ni, R., Liu, Y., Li, G. and Hu, C. (2025) 'Asymmetric multimodal guidance fusion network for realtime visible and thermal semantic segmentation', *Engineering Applications of Artificial Intelligence*, Vol. 142, p.109881.

Yang, P., Chen, W. and Qiu, H. (2024a) 'MMGCN: multi-modal multi-view graph convolutional networks for cancer prognosis prediction', *Computer Methods and Programs in Biomedicine*, Vol. 257, p.108400.

Yang, Y., Chen, S., Zhu, Y., Liu, X., Pan, S. and Wang, X. (2024b) 'Intelligent question answering for water conservancy project inspection driven by knowledge graph and large language model collaboration', *LHB*, Vol. 110, No. 1.

Yin, H. and Zhao, Y. (2025) 'Multi-modal prompt learning with bidirectional layer-wise prompt fusion', *Information Fusion*, Vol. 117, p. 102919.

Zan, S., Ji, W. and Zhou, G. (2024) 'Knowledge graph embeddings based on 2d convolution and self-attention mechanisms for link prediction', *Applied Intelligence*, Vol. 55, No. 2, p.104.

Zhang, C., Li, W., Mo, Y., Tang, W., Li, H. and Zeng, Z. (2024) 'BHRAM: a knowledge graph embedding model based on bidirectional and heterogeneous relational attention mechanism', *Applied Intelligence*, Vol. 55, No. 3, p.245.

Zhao, W., Lu, Y., Liu, Z., Yang, Y. and Jiao, G. (2025) 'Cross-modal alignment with synthetic caption for text-based person search', *International Journal of Multimedia Information Retrieval*, Vol. 14, No. 2, p.11.