Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 20 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i20.58431

PAPER

A Multimodal Emotion Recognition Framework for Dynamic Content Adaptation and Market Response Prediction in Mobile Advertising

Lingfei Wang(✉)

Communication University of Zhejiang, Hangzhou, China

20121255@cuz.edu.cn

ABSTRACT

With the proliferation of mobile internet and the widespread use of smart devices, mobile advertising has emerged as a central channel for enterprise promotion. However, challenges such as content homogeneity and insufficient alignment with users' real-time emotional states have constrained advertising efficiency and diminished user experience. Accurately identifying user emotions and dynamically adapting advertising content to enhance market responsiveness remains an urgent research imperative. Existing studies on emotion recognition and content optimization in mobile advertising exhibit several limitations: many rely solely on unimodal emotion recognition, neglecting the complementary nature of multimodal signals and thereby limiting recognition accuracy; content adaptation approaches are predominantly static, lacking mechanisms for real-time dynamic adjustment; and market response prediction models often fail to integrate multimodal emotional features, resulting in suboptimal prediction accuracy and generalizability. To address these gaps, a multimodal emotion recognition model for mobile advertising was developed, accompanied by a dynamic content adaptation system driven by fused multimodal emotional features. Furthermore, a market response prediction mechanism grounded in multimodal emotional feature fusion was established. These contributions fill critical theoretical gaps in the areas of static content adjustment and unimodal prediction and significantly enrich the theoretical framework for dynamic optimization in mobile advertising.

KEYWORDS

multimodal emotion recognition, mobile advertising, dynamic content adaptation, market response prediction, research progress

1 INTRODUCTION

With the deep penetration of mobile Internet technologies [1–3] and the widespread deployment of intelligent terminals [4–6], mobile advertising has been established as a central channel for product promotion and brand communication by enterprises [7, 8].

Wang, L. (2025). A Multimodal Emotion Recognition Framework for Dynamic Content Adaptation and Market Response Prediction in Mobile Advertising. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(20), pp. 69–83. https://doi.org/10.3991/ijim.v19i20.58431

Article submitted 2025-05-27. Revision uploaded 2025-08-24. Final acceptance 2025-08-31.

During fragmented moments of daily life, users are exposed to an overwhelming volume of mobile advertisements, and their emotional responses to these advertisements directly influence click-through rates, conversion behaviors, and brand perception [9, 10]. However, significant challenges persist within the current mobile advertising, including severe content homogeneity and insufficient alignment with users' real-time emotional needs [11, 12]. These limitations have hindered advertising delivery efficiency and diminished user experience. Thus, the precise recognition of user emotions and the dynamic adaptation of advertising content in response to such emotions have emerged as critical challenges in need of urgent resolution across the industry.

Existing research related to emotion recognition and content optimization in mobile advertising remains constrained in several key respects. A substantial portion of prior work has focused exclusively on unimodal emotion recognition, failing to integrate multimodal information such as speech and visual cues. This limitation has adversely affected recognition accuracy. For example, some studies [13, 14] relied solely on textual data for emotion analysis in advertising, neglecting emotional cues embedded in images and speech signals. With regard to content adaptation, current strategies are predominantly static and lack mechanisms capable of real-time dynamic adjustment in response to users' changing emotional states. The models proposed in some studies [15, 16] offered content optimization strategies that were not capable of real-time updates, thereby falling short of meeting rapidly evolving user emotional demands. Furthermore, most market response prediction models have not fully incorporated multimodal emotional features, leading to suboptimal accuracy and limited generalizability. For instance, the models in some studies [17, 18] considered only partial textual emotional features, overlooking the influence of other modalities on prediction outcomes.

This study includes three main parts. First, it builds a multimodal emotion recognition model for mobile advertising by combining data from text, images, speech, and video. This improves the accuracy and depth of emotion detection. Second, a dynamic content adaptation system is developed, using real-time emotional data to automatically adjust ad content, format, and timing, creating personalized and responsive ad experiences. Third, a market response prediction model is created based on these emotional features to forecast key metrics such as click-through and conversion rates. Overall, this research offers a comprehensive framework that links emotion recognition, dynamic ad adaptation, and response prediction, helping enhance the effectiveness and competitiveness of mobile advertising.

2 MULTIMODAL EMOTION RECOGNITION MODEL FOR MOBILE ADVERTISING

The proposed multimodal emotion recognition model for mobile advertising consists of three major components. The input and encoding layers serve as the foundational stage, primarily responsible for transforming the complex multimodal information present in mobile advertisements into computable feature vectors. Considering that mobile advertisements are frequently presented in the form of short videos comprising visual and auditory elements, bidirectional long short-term memory (Bi-LSTM) networks were employed to perform fine-grained modeling of visual and auditory modalities within advertisement clips. This enables the effective capture of key emotional cues, including variations in color tones, the display of product details, fluctuations in speech intonation, and musical rhythm—features closely associated with user emotional response. Encoding vectors were generated by the backbone model to unify these heterogeneous modality-specific features into

a consistent representation, thereby establishing the foundation for subsequent fusion processes. This ensures that emotional cues across different modalities within mobile advertisements can be effectively extracted and preserved.

The multimodal fusion layer and emotion recognition layer play critical roles in feature integration and emotion quantification, respectively, directly supporting the objective of accurately identifying users' emotional tendencies toward mobile advertisements. In the fusion layer, an attention-gating mechanism was adopted to dynamically integrate textual word embeddings, visual embeddings, and acoustic embeddings. This mechanism emphasizes emotionally salient information that significantly influences users' affective responses—such as emotional connotations in promotional slogans, facial expressions conveying affect in the visual stream, and the tonal characteristics of voice-overs—thereby addressing the limitations associated with unimodal information and enabling a more comprehensive representation of emotional features. Finally, the emotion recognition layer reformulates the emotion classification task as a regression problem, in which emotional polarity is determined through the calculation of an emotion score. This approach allows for the precise quantification of user emotional intensity toward mobile advertisements. The design of this layer is particularly well-suited to the emotional volatility and sensitivity of users in fragmented information environments, and it provides a reliable emotional basis for the subsequent dynamic adaptation of advertising content. In this way, content adaptation strategies can be aligned with users' real-time emotional states with high precision. The architecture of the proposed multimodal emotion recognition model is illustrated in Figure 1.

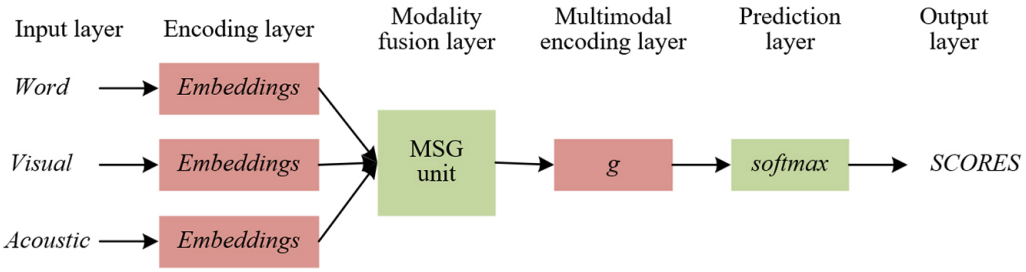

Fig. 1. Architecture of the multimodal emotion recognition model for mobile advertising

The underlying principles of the input and encoding layers are centered on the precise parsing of multimodal information embedded in mobile advertisements. Given the sequential nature of mobile video advertisements, three primary modalities—text, image, and audio—were preprocessed prior to encoding. The text modality focuses on extracting semantic information from advertising copy and product descriptions; the image modality captures visual features such as product details and scene colors from video frames; the audio modality extracts acoustic signals, including voice-over intonation and background music rhythm. Subsequently, each modality was individually encoded using LSTM encoders. Leveraging the LSTM's strengths in handling temporal data, dynamic changes in image frame sequences, temporal characteristics of audio signals, and the contextual logic of textual sequences were effectively modeled, resulting in initial encoding vectors. An attention mechanism was introduced to enhance the extraction of key information related to emotional polarity—such as close-up facial expressions in visual sequences and emotionally charged voice segments—while suppressing irrelevant background noise. Temporal alignment across modalities was then performed to synchronize the feature vectors from all three modalities, thereby establishing a unified spatiotemporal basis for

subsequent fusion. This alignment ensures the complete preservation of distributed emotional cues within the mobile advertisement. Formally, let the vector sequences of the input visual modality and audio modality be denoted as $N^{(u)} = \left[n_1^{(u)}, n_2^{(u)}, \ldots, n_s^{(u)}\right]$ and $X^{(u)} = \left[x_1^{(u)}, x_2^{(u)}, \ldots, x_s^{(u)}\right]$, where $N^{(u)}$ and $X^{(u)}$ represent the input vector representations for the image and audio modalities at time step s, respectively. The corresponding LSTM output vectors at time step s are denoted as $g_n^{(u)}$ for image and $g_x^{(u)}$ for audio. The u-th vectors in the encoded sequences are expressed as N_u and X_u. The attention mechanism is represented by ATT, and the concatenation operation is denoted by $\oplus$. The corresponding computational steps are defined as follows:

$$g_n^{(u)} = LSTM_{\rightarrow}(N^{(u)}) \oplus LSTM_{\leftarrow}(N^{(u)}) \tag{1}$$

$$g_x^{(u)} = LSTM_{\rightarrow}(X^{(u)}) \oplus LSTM_{\leftarrow}(X^{(u)}) \tag{2}$$

$$N_u = g_n^{(u)} \oplus ATT\left(\left[g_n^1, g_n^2, \ldots, g_n^S\right]\right) \tag{3}$$

$$X_u = g_x^{(u)} \oplus ATT\left(\left[g_x^1, g_x^2, \ldots, g_x^S\right]\right) \tag{4}$$

The fundamental principle of the multimodal fusion layer lies in achieving deep cross-modal emotional synergy within mobile advertising. A multimodal shift-gated (MSG) unit was employed, designed to dynamically adjust fusion weights by learning the interdependencies among modalities. The structure of the MSG unit is illustrated in Figure 2. Specifically, word embeddings from the text modality were first concatenated with the corresponding image and audio feature vectors. These concatenated vectors were then processed through weight matrices q_n and q_β, generating gate vectors $q_n^{(u)}$ and $q_\beta^{(u)}$, which quantify the emotional association strengths between the visual and auditory modalities and the textual semantics. Subsequently, shift vectors G_u^l for image and audio were computed based on the generated gate vectors. These shift vectors were modulated by a proportionality factor β and then integrated with the word vector R_u, yielding the multimodal word vector R_u^l. The computational process is defined as follows:

$$q_n^{(u)} = E\left(Q_{gn}\left[g_n^{(u)}; r^{(u)}\right] + y_n\right) \tag{5}$$

$$q_x^{(u)} = E\left(Q_{gx}\left[g_x^{(u)}; r^{(u)}\right] + y_x\right) \tag{6}$$

$$G_u^l = q_n^{(u)} \cdot \left(Q_n g_n^{(u)} + q_x^{(u)} \cdot \left(Q_x g_x^{(u)}\right)\right) + y_g^{(u)} \tag{7}$$

$$R_u^l = R_u + \beta G_u^l \tag{8}$$

$$\beta = MIN\left(\frac{\left\|R_u\right\|_2}{\left\|G_u^l\right\|_2}\alpha, 1\right) \tag{9}$$

This process simultaneously preserves modality-specific emotional information and enhances cross-modal emotional resonance through dynamical weights. As a result, inconsistencies among the emotional expressions of visual, auditory, and textual modalities in mobile advertisements are effectively addressed.

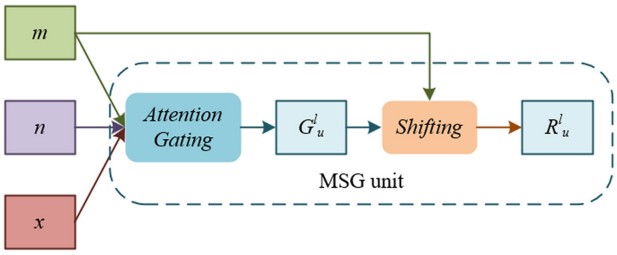

Fig. 2. Structure of the MSG unit

The fundamental principle of the emotion recognition layer is centered on the precise quantification of user emotions in response to mobile advertisements. Once the fused multimodal representation was obtained, a second-stage encoding was performed using an LSTM network to capture the temporal dynamics of emotional transitions embedded in the integrated features. The resulting encoded representation was then passed through a fully connected layer, which reframes the emotion recognition task as a regression problem. By mapping the user's emotional response to a continuous scale—such as a spectrum ranging from "aversion" to "affection"—a specific emotion score was produced as output. This design is well aligned with the emotional volatility observed during users' fragmented browsing of mobile advertisements, and it effectively avoids the oversimplification inherent in conventional classification models, which often fail to capture nuanced affective states. Based on the final emotion score, the polarity of the user's emotional response was determined by score intervals, providing a fine-grained, numerically grounded basis for subsequent dynamic content adaptation.

3 DYNAMIC CONTENT ADAPTATION SYSTEM BASED ON THE FUSION OF MULTIMODAL EMOTIONAL FEATURES IN MOBILE ADVERTISING

The construction of the dynamic content adaptation system based on the fusion of multimodal emotional features in mobile advertising is centered on achieving dynamic content matching. Building upon mature adaptation frameworks, the system emphasizes the deep integration of emotional features to enhance affective coherence. The system architecture is presented in Figure 3.

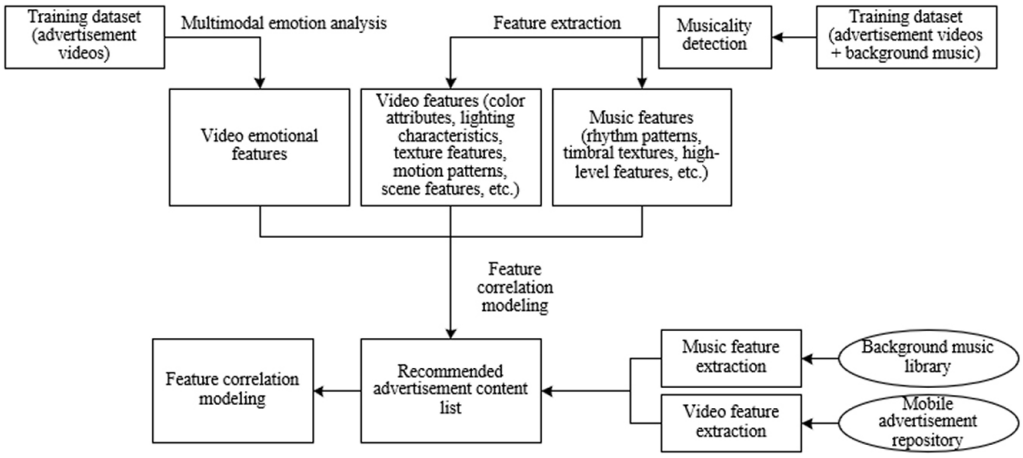

Fig. 3. Dynamic content adaptation system based on the fusion of multimodal emotional features in mobile advertising

The process begins with a training dataset of mobile advertisement videos, which is filtered using a musicality detection module to exclude samples with inadequate musical structure, thereby ensuring the quality of subsequent content recommendations. In the feature extraction stage, a dual-path strategy was adopted. The music feature extraction module focuses on the physical properties of audio signals, such as rhythm and spectral characteristics, to represent the style and cadence of background music. The video feature extraction module integrates visual features—such as color, lighting, and texture—with the emotional feature vectors produced by the multimodal emotion recognition model described in Section 2. This approach preserves objective visual attributes while quantifying the emotional tone conveyed by the video content. Such a feature design enables the system not only to perceive the visual content but also to interpret its emotional core, providing a multidimensional foundation for dynamic adaptation. The core logic of system construction is embodied in the recommendation module, which is responsible for modeling the associations between multimodal features and enabling real-time adaptive responses. Specifically, the background music recommendation module for mobile advertisements analyzes the semantic correlation between video and music features to uncover latent emotional congruencies. For instance, emotional features labeled as "joyful" in the video domain are associated with music characterized by "fast tempo" and "high frequency," while "calm" emotional features are matched with "slow tempo" and "low frequency" musical patterns. When a new mobile advertisement video is input, the system extracts its fused emotional feature representation in real time. By comparing this representation with those of audio files stored in the content database, similarity weights can be calculated, and a ranked recommendation list can be generated accordingly.

The core of multimodal feature correlation modeling in this system is based on a latent semantic analysis model across multiple feature types, with the objective of transforming multimodal characteristics of mobile advertisements into a computable entity association network. This network serves as a structured foundation for dynamic content adaptation. Each feature dimension—whether an emotional feature, a visual property such as color or lighting, or an auditory property such as rhythm or spectral content—is treated as an independent entity $\{A_1, A_2, \ldots, A_v\}$. A heterogeneous graph $H(N, R)$ is constructed, where each vertex corresponds to a specific feature, and an edge r_{uk} indicates a co-occurrence relationship between different features observed in mobile advertisement samples. This graph-based representation enables the explicit mapping of cross-modal associations. For example, the frequent co-occurrence of video emotional features and musical rhythm components is made computationally visible, laying the foundation for subsequent quantitative analysis and ensuring that the system can detect latent associations among features to support dynamic adaptation decisions.

The construction of a unified matrix I plays a pivotal role in quantifying the strength of associations among multimodal features. The design of matrix I directly supports the system's demand for precision in dynamic content adaptation. The matrix I is composed of $V \times V$ correlation matrices L_{uk}. For features of the same type, L_{uk} is defined as an identity matrix to preserve their structural independence. For features of different types, L_{uk} is constructed by statistically analyzing the co-occurrence frequency and strength of the two feature types in the training data of mobile advertisements. For example, if in emotionally "soothing" videos, music with "low-frequency" and "slow tempo" characteristics tends to appear with high probability, then the corresponding element in the correlation matrix is assigned

a relatively large value. This formulation preserves the intrinsic properties of unimodal features while enabling the cross-modal matrices to capture the affective compatibility rules—such as between emotion and music. This provides a quantitative standard for feature association, ensuring that the dynamic content adaptation process can be guided by the learned strength of historical co-occurrence patterns. The following equation defines the correlation matrix between emotional features and rhythmic features:

$$L_{E\_B} = L_{E\_D} \times L^{S}_{B\_D} \tag{10}$$

The process of extracting latent semantics based on the principle of mutual reinforcement provides the core decision-making foundation for dynamic content adaptation, thereby enabling the transition from feature correlation to intelligent matching. The system was conducted to compute a saliency vector e_u, which reveals dominant concepts underlying the associations between multimodal features. These concepts represent optimal alignment patterns between video and audio features in mobile advertisements, such as "joyful emotion + fast tempo + high-frequency spectrum" or "calm emotion + slow tempo + low-frequency spectrum." A latent space of dimensionality j, constructed from the first j feature vectors, encapsulates the most representative patterns of emotional and stylistic alignment. When the system performs dynamic adaptation for a given mobile advertisement, its video features—encompassing emotional attributes—are projected into this latent space. This enables the rapid identification of the corresponding salient concept, which is then used to retrieve audio features from the music library that conform to the identified pattern. This ensures that the recommended background music remains emotionally consistent with the video's affective tone. Through this mechanism, the system transcends the limitations of surface-level features by operating on a deeper semantic level. As a result, precise and dynamic content adaptation can be achieved, significantly enhancing emotional coherence and user receptivity to mobile advertisements. The saliency vector e_u is defined as:

$$e_u = \sum\nolimits_{\forall k:k\neq u} L_{uk} e_u \tag{11}$$

It can be reformulated using the unified matrix I as:

$$e = I \times e \tag{12}$$

The core principle underlying the system's dynamic content adaptation lies in the accurate matching of query video content and candidate audio tracks through vector transformation and space mapping. When a mobile advertisement video query is input, its multimodal features are extracted and integrated to form a unified query vector w, which includes both the objective visual attributes and the embedded emotional signals. This query vector is then projected into the latent semantic space defined by the top j salient concepts using the operation: $w_z = w \times I_j$. Through this transformation, the video features are mapped onto a shared semantic dimension that is also interpretable by music features. This enables cross-modal emotional and stylistic associations to be quantitatively captured. For instance, a video conveying a "joyful" emotional tone would be represented in proximity to audio features with "fast tempo" in the same space, thereby establishing a coordinate basis for subsequent matching.

To complete the content adaptation loop, similarity comparison is performed using cosine similarity, enabling the generation of a dynamic recommendation list.

Within the latent semantic space, the projected query vector w_z is compared against each row of the music feature matrix using cosine similarity. A cosine value closer to 1 indicates a higher degree of emotional and stylistic alignment between the video and audio content. For instance, a video characterized by "dim lighting + calm emotion" demonstrates high similarity with music that exhibits "slow tempo + low-frequency spectrum." Based on the similarity scores, a ranked recommendation list is dynamically generated to adapt to the real-time features of the query video. In cases where the emotional features of the video shift due to scene transitions, the system re-extracts relevant features and repeats the above process to promptly update the recommendation results. This mechanism ensures that the background music remains consistently aligned with the video's emotional expression and stylistic tone, thereby enabling dynamic optimization of mobile advertisement content through deep multimodal feature association and real-time computation. Specifically, assuming the projected music feature x is denoted as x_z, the similarity between a given music feature x and the query vector w is computed as follows:

$$SIM(x,w) = \frac{X_z \cdot X_z}{\left|X_z\right|\left|w_z\right|} \tag{13}$$

Finally, the top-v ranked advertisement contents are selected based on the computed scores. Let X represent the set of music features, and let s denote a candidate item. The score is computed using the following equation:

$$SCORE_{CO}(s) = \sum \beta \in X l_{x\_s} \cdot SIM(x,w) \tag{14}$$

4 MARKET RESPONSE PREDICTION MECHANISM BASED ON THE FUSION OF MULTIMODAL EMOTIONAL FEATURES IN MOBILE ADVERTISING

The core of the proposed market response prediction mechanism lies in establishing a quantitative relationship between the fused multimodal emotional features of mobile advertisements and market response indicators. Initially, emotional features are extracted from text, image, and audio modalities. These are then integrated using the multimodal fusion model to produce a unified emotional feature vector. In parallel, non-emotional features—such as advertisement delivery time and user profile attributes—are incorporated to construct a comprehensive feature matrix. Based on historical advertising campaign data, a predictive model was constructed using either gradient boosting decision trees (GBDT) or neural network architectures. The model was trained with the fused emotional feature vector as a central input, learning the mapping between multimodal emotional signals and market response indicators such as click-through rate and conversion rate. For example, a pattern such as "joyful emotional features + young user profile" may correspond to a higher click-through rate, enabling the model to capture deep correlations between affective content and user feedback.

To ensure predictive accuracy, the mechanism is designed to support dynamic iteration and real-time correction, thereby realizing dynamic tracking of market response. After a new mobile advertisement is launched, user interaction data are collected in real time by the system and used to update the feature matrix alongside freshly extracted multimodal emotional features. An online learning algorithm

is then employed to incrementally fine-tune the model parameters, thereby adjusting the weighting of feature–response associations. For instance, if the "soothing emotional feature" is observed to elicit a sudden decline in user engagement among a specific user segment, the influence coefficient of that feature can be promptly recalibrated. The prediction outcomes are simultaneously fed back into the advertising strategy layer, supporting the dynamic optimization of content and audience targeting. In doing so, a closed-loop process is formed—feature fusion → prediction → strategy adjustment—which enhances the controllability and precision of market response outcomes for mobile advertising.

5 EXPERIMENTAL RESULTS AND ANALYSIS

Table 1. F1-scores (%) of emotion recognition using different algorithms

Method	Multimodal			
	Positive	Negative	Neutral	*w-avg*
LMF	52.36	54.23	72.62	61.23
MULT	52.41	52.36	73.54	62.54
CMU-MOSI	55.69	58.64	71.26	63.25
CLIP	52.31	55.21	72.65	63.65
VisualBERT	54.69	56.32	73.54	64.52
Proposed method	57.89	58.64	75.23	66.23

As shown in Table 1, the proposed multimodal emotion recognition model demonstrates clear advantages across all emotional categories and the weighted average F1 metric when compared with existing algorithms. In recognizing positive emotions, the proposed model achieved an F1-score of 57.89%, outperforming low-rank multimodal fusion (LMF) (52.36%), multimodal transformer (MULT) (52.41%), Carnegie Mellon University-Multimodal Opinion Sentiment Intensity (CMU-MOSI) dataset (55.69%), contrastive language-image pretraining (CLIP) (52.31%), and Visual Bidirectional Encoder Representations from Transformers (VisualBERT) (54.69%). This result highlights the model's enhanced capability for capturing positively valenced emotional expressions. For negative emotion recognition, the proposed method achieved a score of 58.64%, equal to CMU-MOSI, but its superior performance across other categories reflects greater model stability. In the neutral emotion category, the proposed model reached an F1-score of 75.23%, significantly surpassing CMU-MOSI (71.26%) and all other baselines, indicating improved precision in recognizing emotionally neutral content. In terms of weighted average performance, the proposed model attained an F1-score of 66.23%, which notably exceeds the scores of all comparative methods, including CMU-MOSI (63.25%) and VisualBERT (64.52%). These results collectively demonstrate the overall superiority of the proposed model in the multimodal emotion recognition task. The effectiveness of the model is attributed to the integrated processing of text, image, speech, and video modalities, which significantly improves the accuracy and comprehensiveness of emotional feature extraction and fusion. These advancements provide a strong technical foundation for enhancing emotional perception capabilities in the dynamic content adaptation system and for improving the precision of the market response prediction mechanism.

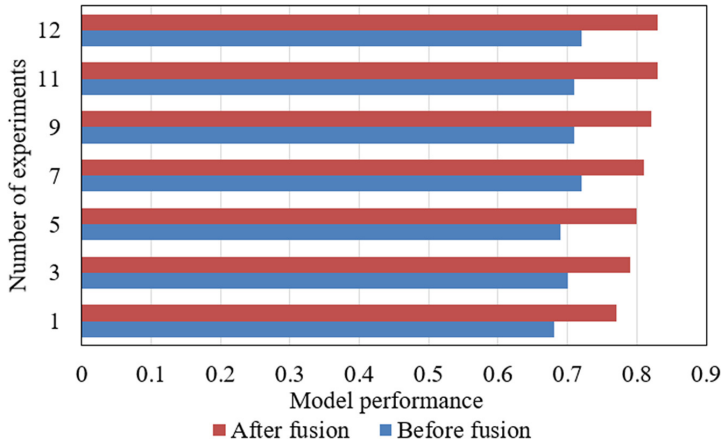

Fig. 4. Comparison of model performance before and after multimodal feature fusion

As illustrated in Figure 4, a consistent improvement in model performance was observed across all twelve experimental trials. In each case, the red bars extend farther to the right than the corresponding bars before fusion, indicating that the incorporation of multimodal fusion consistently enhanced model performance. The data provides direct empirical validation for the effectiveness of the feature fusion strategy within the proposed multimodal emotion recognition model. The model was designed to extract and integrate emotional features from multiple data modalities—including text, image, audio, and video—allowing for complementary reinforcement across single-modality limitations. The model's performance prior to fusion reflects the predictive capacity in the absence of integrated features. In contrast, the post-fusion performance gains clearly demonstrate the synergistic advantages of multimodal fusion, where the joint representation enhances the expressiveness and discriminative power of emotional features.

Table 2. Comparative results of content adaptation across different models

Algorithm Model	*Recall*@5 (%)	*NDCG*@5 (%)	*Recall*@10 (%)	*NDCG*@10 (%)
HSTU	0.5231	0.4125	0.6125	0.4215
MTGR	0.5896	0.4786	0.6789	0.5236
GRAB	0.5874	0.4752	0.6852	0.5124
AUGRU	0.6325	0.5231	0.7241	0.5569
Deep & Cross Network	0.6452	0.5369	0.7456	0.5896
Proposed model	0.6639	0.5426	0.7689	0.6123

As shown in Table 2, the proposed dynamic content adaptation system based on multimodal emotional feature fusion in mobile advertising demonstrates clear superiority across all key evaluation metrics. In both Recall@5 and Recall@10, the proposed model achieved scores of 0.6639 and 0.7689, hierarchical sequential transduction units (HSTU) (0.5231 and 0.6125), hybrid generative recommendation model (MTGR) (0.5896 and 0.6789), and other baseline models. These results indicate a notably enhanced capacity to retrieve advertisement content that aligns with users' emotional states and preferences, thereby improving the breadth of content coverage. In addition, the model exhibited superior performance in NDCG@5 and NDCG@10, with values of 0.5426 and 0.6123, respectively. These scores reflect improved relevance ranking in content recommendation, enabling the delivery of

advertisements that are both emotionally congruent and of higher perceived quality, thereby increasing the recommendation accuracy. The comparative results substantiate the effectiveness of the system's design, particularly the use of multimodal emotional feature fusion. This system not only improves the personalization of content recommendations but also enhances the ranking quality of recommended content. As such, the proposed system provides a reliable technical foundation for dynamic adaptation in mobile advertising.

Table 3. Results of the module ablation study

	Recall@5 (%)	*NDCG*@5 (%)	*Recall*@10 (%)	*NDCG*@10 (%)
Without the attention-gating mechanism	0.6123	0.5124	0.7152	0.5123
Without Bi-LSTM	0.6354	0.5123	0.7124	0.5326
Using the traditional latent semantic analysis model	0.6258	0.5148	0.7236	0.5348
Complete model	0.6659	0.5468	0.7653	0.6123

As observed in Table 3, the complete model outperforms all ablated variants across key performance metrics, including Recall@5, NDCG@5, Recall@10, and NDCG@10. These results provide strong empirical evidence supporting the effectiveness of the system's modular co-design architecture. For example, the removal of the attention-gating mechanism resulted in a significant performance drop—Recall@5 decreased from 0.6659 to 0.6123, and NDCG@5 dropped from 0.5468 to 0.5124. This demonstrates the critical role of attention-gating in enhancing multimodal feature interaction and improving the precision of content matching. Similarly, the adoption of a traditional latent semantic analysis model led to reduced performance across all metrics, highlighting the superior capacity of the proposed modern fusion architecture to capture complex emotional associations across modalities. The ablation study further confirms the necessity of each component: the attention-gating mechanism was shown to refine attention allocation over emotional features, improving the system's focus during matching; the Bi-LSTM network enabled the capturing of temporal emotional dynamics, which are essential for adapting to the variability of advertisement contexts; and the modern fusion architecture significantly enhanced the expressiveness of emotional representations. The superior performance of the complete model indicates that the integration of multimodal emotional feature fusion with a well-coordinated modular design has led to substantial improvements in both advertisement recall and recommendation quality, thereby fulfilling its core objective of delivering personalized and dynamically adaptive content within mobile advertising environments.

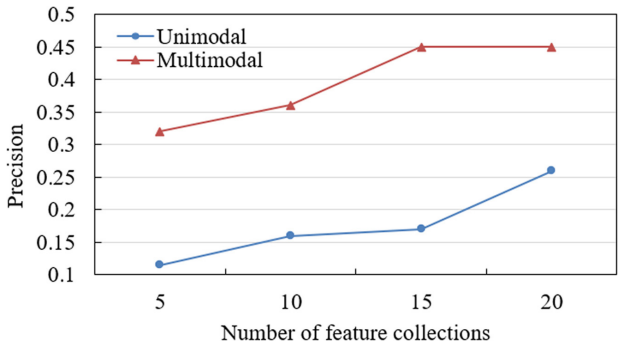

Fig. 5. Precision of the market response mechanism

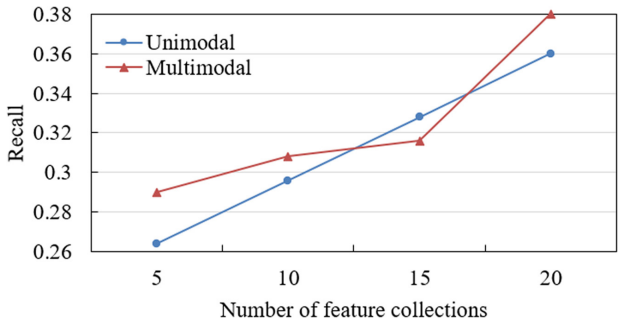

Fig. 6. Recall of the market response mechanism

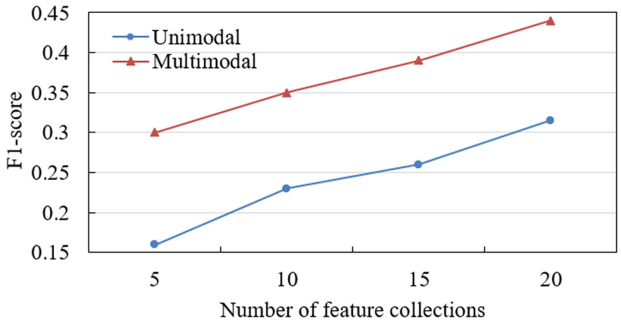

Fig. 7. F1-score of the market response mechanism

Analysis of the experimental results presented in Figures 5, 6, and 7 demonstrates that the market response prediction mechanism based on multimodal emotional feature fusion significantly outperforms its unimodal counterpart in terms of precision, recall, and F1-score. Furthermore, the performance advantage was observed to widen progressively as the number of feature collection iterations increased. In Figure 5, the precision of the multimodal model consistently exceeded that of the unimodal model, with a steeper growth trajectory. This indicates that the fusion of textual, visual, acoustic, and video-based emotional features enabled a more accurate determination of market responses, thereby improving the model's ability to identify high-response advertisement instances. Figure 6 shows that the recall of the multimodal model surpassed that of the unimodal model once the number of feature collection iterations reached or exceeded 15, with continued acceleration thereafter. This suggests that a more comprehensive capture of emotion-response associations was achieved, reducing false negatives and improving detection of high-response advertising content. As illustrated in Figure 7, the F1-score curve for the multimodal model not only started at a higher baseline but also exhibited a greater rate of increase, confirming its superior performance in balancing precision and coverage. The proposed mechanism relied on a multimodal emotional recognition model that deeply integrated emotional signals from text sentiment, visual emotion, vocal tone, and video ambiance, allowing for the construction of a predictive model. The complementarity of multimodal data enabled the extraction of holistic, scene-level emotional associations between advertisements and user interactions—far beyond the fragmented information provided by any single modality. Consequently, the mapping from emotional expression to market response was modeled with greater fidelity. The experimental results validate that the fusion of multimodal emotional features substantially enhanced all three critical performance metrics—precision, recall, and F1-score—thereby providing a reliable quantitative foundation for the dynamic optimization of advertising delivery strategies.

6 CONCLUSION

A systematic investigation into the application of multimodal emotional recognition in the mobile advertising domain was conducted. A multimodal emotional recognition model was constructed, a dynamic content adaptation system was designed, and a market response prediction mechanism was established—together forming a comprehensive technical framework encompassing emotional perception, content adaptation, and effectiveness prediction. The experimental results confirmed that the multimodal emotional recognition model, through the fusion of textual, visual, vocal, and video-based data, substantially improved the accuracy and completeness of emotional feature extraction. This advancement served as a critical foundation for downstream applications. Building upon this model, the dynamic content adaptation system was shown to effectively capture users' emotional shifts in real time, enabling the personalized adjustment of advertisement content, form, and delivery timing, thereby enhancing both user experience and advertising acceptance. In parallel, the market response prediction mechanism leveraged the fused emotional features to accurately estimate performance metrics such as click-through rate and conversion rate, providing quantitative guidance for the optimization of advertising delivery strategies. Through the synergy of these three components, a dual contribution was achieved: theoretically, by enriching the intersectional study of multimodal affective computing and dynamic mobile advertising optimization, and practically, by offering an implementable solution for precision marketing. The limitations of unimodal systems were addressed, and the shift from broad-spectrum outreach to targeted engagement in mobile advertising was enabled. These findings underscore both the academic significance and the industrial applicability of the proposed approach.

Nevertheless, certain limitations persist. First, the training data for the models were primarily drawn from a specific category of mobile advertisements, potentially limiting generalizability to cross-industry contexts. Second, the efficiency of real-time emotional recognition and content adaptation efficiency remains constrained by device-side computational capabilities, particularly on low-specification hardware. Moreover, the subjectivity in emotional feature interpretation may introduce variance in prediction accuracy. Future research may proceed along three directions. First, broader multimodal data sources—including physiological signals—should be integrated to enhance the granularity and generality of emotional cues. Second, algorithmic lightweighting strategies should be pursued, in conjunction with edge computing technologies, to improve real-time processing capabilities and ensure compatibility across heterogeneous terminal devices. Third, application scenarios should be further expanded, with adaptation strategies explored across platforms such as short video applications and social media. Reinforcement learning techniques may also be introduced to dynamically optimize the correlation model between emotional features and market response, thereby unlocking the full commercial potential of multimodal emotional recognition in marketing contexts.

7 REFERENCES

[1] F. Sakka, A. Gura, V. Latysheva, E. Mamlenkova, and O. Kolosova, "Solving technological, pedagogical, and psychological problems in mobile learning," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 16, no. 2, pp. 144–158, 2022. https://doi.org/10.3991/ijim.v16i02.26205

[2] F. Yang, "Leveraging mobile interaction technologies for real-time decision making in enterprise management systems," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 19, no. 2, pp. 65–78, 2025. https://doi.org/10.3991/ijim.v19i02.53743

[3] A. I. Suroso, I. Fahmi, and H. Tandra, "Adoption of mobile internet and the implication on palm oil productivity: Case study in Siak regency," *International Journal of Sustainable Development and Planning*, vol. 18, no. 1, pp. 335–342, 2023. https://doi.org/10.18280/ijsdp.180135

[4] N. D. Azeez and N. Y. Mohammed, "Factors influencing adoption of mobile health monitoring system: Extending UTAUT2 with trust," *Ingénierie des Systèmes d'Information*, vol. 27, no. 2, pp. 223–232, 2022. https://doi.org/10.18280/isi.270206

[5] F. Wang, D. Jiang, H. Wen, and S. Qi, "Security level protection for intelligent terminals based on differential privacy," *Telecommunication Systems*, vol. 74, pp. 425–435, 2020. https://doi.org/10.1007/s11235-020-00665-x

[6] F. Kojima, H. Harada, and M. Fujise, "Inter-vehicle communication network with an autonomous relay access scheme," *IEICE Transactions on Communications*, vol. 84, no. 3, pp. 566–575, 2001.

[7] C. Feijóo-González, J. L. Gómez-Barroso, and I. J. Martínez-Martínez, "Nuevas vías para la comunicación empresarial: Publicidad en el móvil," *Profesional de la Información*, vol. 19, no. 2, pp. 140–148, 2010. https://doi.org/10.3145/epi.2010.mar.04

[8] C. Jebarajakirthy, H. I. Maseeh, Z. Morshed, A. Shankar, D. Arli, and R. Pentecost, "Mobile advertising: A systematic literature review and future research agenda," *International Journal of Consumer Studies*, vol. 45, no. 6, pp. 1258–1291, 2021. https://doi.org/10.1111/ijcs.12728

[9] F. Saadeghvaziri and H. K. Hosseini, "Mobile advertising: An investigation of factors creating positive attitude in Iranian customers," *African Journal of Business Management*, vol. 5, no. 2, pp. 394–404, 2011.

[10] S. Sanz-Blas, C. Ruiz-Mafé, and J. Martí-Parreño, "Message-driven factors influencing opening and forwarding of mobile advertising messages," *International Journal of Mobile Communications*, vol. 13, no. 4, pp. 339–357, 2015. https://doi.org/10.1504/IJMC.2015.070058

[11] P. L. P. Rau, Q. Liao, and C. Chen, "Factors influencing mobile advertising avoidance," *International Journal of Mobile Communications*, vol. 11, no. 2, pp. 123–139, 2013. https://doi.org/10.1504/IJMC.2013.052637

[12] B. Yang, Y. Kim, and C. Yoo, "The integrated mobile advertising model: The effects of technology-and emotion-based evaluations," *Journal of Business Research*, vol. 66, no. 9, pp. 1345–1352, 2013. https://doi.org/10.1016/j.jbusres.2012.02.035

[13] C. Livas, F. Theofanidis, and N. Karali, "Consumer sentiment toward international activist advertising," *Innovative Marketing*, vol. 19, no. 2, pp. 250–260, 2023. https://doi.org/10.21511/im.19(2).2023.20

[14] V. P. Rosas, R. Mihalcea, and L. P. Morency, "Multimodal sentiment analysis of Spanish online videos," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 38–45, 2013. https://doi.org/10.1109/MIS.2013.9

[15] P. Li, S. Mei, and W. Zhong, "Fee or subsidy? Pricing strategies for digital content platforms with different content and advertising," *Managerial and Decision Economics*, vol. 44, no. 8, pp. 4482–4506, 2023. https://doi.org/10.1002/mde.3960

[16] S. Thomaidou, K. Liakopoulos, and M. Vazirgiannis, "Toward an integrated framework for automated development and optimization of online advertising campaigns," *Intelligent Data Analysis*, vol. 18, no. 6, pp. 1199–1227, 2014. https://doi.org/10.3233/IDA-140691

[17] M. P. Vandenbergh, K. T. Raimi, and J. M. Gilligan, "Energy and climate change: A climate prediction market," *UCLA Law Review*, vol. 61, no. 6, pp. 1962–2017, 2014.

[18] C. Zhang, N. N. Sjarif, and R. B. Ibrahim, "Decision fusion for stock market prediction: A systematic review," *IEEE Access*, vol. 10, pp. 81364–81379, 2022. https://doi.org/10.1109/ACCESS.2022.3195942

8 AUTHOR

Lingfei Wang graduated from Wuhan University with a master's degree and is currently working at School of Cultural Creativity and Management, Communication University of Zhejiang. His main research interests include brand management, public interest communication, and new media operations (E-mail: 20121255@cuz.edu.cn).