

Classification of Highly-Rated Comments in Rural Short Videos Based on Multimodal Emotion Features

Yimeng Gu

School of Media and Design, Xi'an Peihua University, Xi'an 710100, Shaanxi, China

Corresponding author: Yimeng Gu (e-mail: gym9022@sina.cn)

ABSTRACT The rapid growth of rural short video platforms has generated large volumes of multimodal interaction data, where audiovisual signals and textual comments jointly reflect users' emotional preferences and engagement patterns. However, semantic inconsistency between heterogeneous modalities and the lack of interpretable emotion modeling limit the accurate classification of highly-rated comments. This study proposes a cross-modal emotion alignment framework that integrates temporal audiovisual features with textual semantic representations for fine-grained comment classification. A dual-stream encoding network combines temporal attention and Wav2Vec2.0 to capture scene evolution, dialect rhythms, and ambient acoustic characteristics, while a domain-adapted text encoder extracts implicit emotional semantics from user comments. A bidirectional cross-attention gating module dynamically aligns multimodal features and suppresses semantic noise caused by irony and metaphor, followed by a multi-head classifier for structured emotion prediction. Experiments on a self-constructed rural short video corpus demonstrate that the proposed model achieves a Macro-F1 score of 89.4% while reducing the confusion rate to 11.9%, outperforming representative baseline methods. Beyond improving multimodal sentiment analysis, the proposed architecture establishes an effective signal-level fusion and feature interaction strategy for heterogeneous audiovisual data, providing useful insights for intelligent information perception, multimodal signal transmission, and adaptive communication systems in electromagnetic sensing and next-generation interactive media applications.

INDEX TERMS multimodal sentiment analysis, cross-modal alignment, audiovisual signal fusion, intelligent information perception, electromagnetic sensing

I. INTRODUCTION

IN recent years, with the sinking of mobile Internet infrastructure and the vertical operation of short video platforms, rural theme content has gradually evolved into an important cultural carrier and emotional link for digital rural construction [1,2]. As a concentrated expression of user value identification and emotional projection, high praise comments not only directly reflect the audience's acceptance threshold of local narratives, but also deeply influence the platform's content distribution logic and community ecological evolution [3]. However, the existing comment analysis system mostly follows the unimodal text mining paradigm of general social media, which is difficult to adapt to the multimodal interwoven context unique to rural short videos [4]. There is a complex nonlinear mapping relationship between audio-visual elements in rural scenes and fragmented user comments. The visual atmosphere (e.g., color temperature and scenery) and acoustic prosody serve as the emotional 'anchor' that triggers user resonance, while the comments function as the textual 'echo'. To prove the

marginal contribution of the video modality, we introduce cross-modal gating to verify that visual priors significantly reduce semantic ambiguity in social interactions, and there is an urgent need to construct a fine-grained emotion analysis framework that fits the local digital context.

The core bottleneck facing the classification of high like comments on rural short videos currently lies in the structural mismatch between audiovisual symbols and textual semantics, as well as the unexplainability of emotional driving mechanisms [5,6]. On the one hand, the rural audiovisual modality contains a large number of unstructured, low signal-to-noise ratio implicit clues, such as the gradual evolution of cool and warm tones often implying a shift in emotional tone, while the fluctuations of dialect tones and the arousal of environmental sounds constitute implicit scales of emotional intensity. These clues and the frequent use of local metaphors, ironic expressions, or ellipsis in comments can easily lead to semantic drift, resulting in difficulty in aligning feature spaces. On the other hand, mainstream multimodal models rely heavily on end-to-end

black box optimization and lack explicit reasoning paths for abstract categories such as “emotional resonance” and “cultural identity”, making it difficult to provide traceable attribution basis at the algorithmic decision-making level. This type of technical defect not only hinders further breakthroughs in classification accuracy, but also weakens the practical usability of the model in accurate recommendation of agricultural and rural content and community governance.

In response to the above challenges, this article proposes a cross modal emotion alignment classification architecture for rural contexts, aiming to achieve fine-grained fusion and interpretable decision-making of audio-visual priors and text semantics. The study first constructs a dual stream feature encoding network, which jointly captures scene color evolution, dialect prosody fluctuations, and environmental sound arousal through TAM and acoustic forward pre-training models. At the same time, it relies on domain adapted text encoders to extract implicit emotional representations of comments. On this basis, a cross modal emotion gating interaction module is designed to dynamically calculate the alignment weights between modalities using a bidirectional cross attention mechanism, effectively suppressing feature noise caused by irony and metaphor, and outputting structured emotion labels through a multi-head fine-grained classifier. The system validation on the self built local short video high like comment corpus shows that the architecture achieves significant gains in classification accuracy and confusion suppression. It not only provides a new computational path for multimodal understanding of rural digital content, but also lays a methodological foundation for the quantitative governance and interpretable recommendation of high like comment ecology.

II. RELATED WORK

Multimodal sentiment computing has become a research hotspot in short video content understanding, with the core goal of breaking through the limitations of sentiment polarity discrimination in a single text modality and achieving collaborative modeling of audio-visual cues and language symbols [7,8]. Early research often used feature concatenation or decision level fusion strategies, which achieved certain results in general scenarios, but were difficult to deal with the complex emotional structure hidden in high like comments. In recent years, cross modal representation learning based on deep learning has gradually shifted towards fine-grained alignment, narrowing the semantic distance between modalities through shared latent space or contrastive learning [9,10]. However, existing work mainly focuses on standardized content such as film and television commentary and entertainment live streaming, and there are significant domain differences in the distribution of pre-training data compared to rural short videos. Rurality is characterized by the use of non-standard dialects, agricultural symbols, and local cultural representations. Due to issues with feature extraction, extensive emotional resonance models often suffer from representation failures, resulting in a blurred

boundary between emotional resonance and cultural identity. It is urgent to introduce domain priors to reshape semantic mapping relationships between modalities [11,12].

The key to cross modal alignment technology lies in establishing a reliable mechanism for inter modal attention allocation [13,14]. The current mainstream methods mostly rely on one-way or static cross attention, achieving feature interaction through query key value matching, but lack dynamic adjustment ability when dealing with semantic conflicts or low signal-to-noise ratio inputs [15,16]. The frequent use of ironic sentence structures and local metaphors in rural high praise comments can introduce a large number of misleading features, and if fully integrated directly, it exacerbates semantic ambiguity. Some scholars have attempted to introduce gating mechanisms or noise filtering modules to suppress abnormal activation through confidence thresholds or rule constraints, but such methods often rely on manual design and have limited generalization ability. On the other hand, the prosodic features and environmental sound arousal in audio modalities have not been fully explored, and their temporal coupling relationship with visual color evolution lacks structured modeling, resulting in a lack of multi-source evidence support for emotion intensity assessment and difficulty in supporting the discrimination requirements of fine-grained classification tasks [17,18].

Interpretability is a necessary prerequisite for the practical application of multimodal sentiment analysis. The existing classification models rely heavily on end-to-end optimization, and the decision-making process lacks explicit attribution of key modal contributions, making it difficult to meet the requirements for algorithm transparency in agricultural and rural content governance. Although some studies attempt to provide post hoc explanations through attention visualization or gradient inversion, such methods only reflect the distribution of feature weights and do not touch upon the causal logic of emotion driven mechanisms. At the same time, research on domain adaptation in rural contexts is still in its early stages, and most of the work only stays at the level of vocabulary expansion or lightweight fine-tuning, without addressing the semantic mismatch between audiovisual symbols and fragmented texts at the architectural level. In this context, this article integrates emotion alignment and gating interaction into a unified framework, dynamically calibrates modal priors through bidirectional cross attention, and constructs a traceable emotion inference path by combining domain fine tuned text encoders. This design not only addresses the shortcomings of existing methods in adapting to rural digital contexts, but also provides a computational paradigm that combines accuracy and interpretability for classifying high like comments.

III. METHOD

A. DATASET CONSTRUCTION AND PREPROCESSING

The data collection of this study focuses on the top content in the vertical field of agriculture, rural areas and farmers on mainstream short video platforms in China. Through

targeted crawler interfaces, the top 100 rural themed short videos from October 2024 to October 2025 were selected based on a weighted engagement score (likes, share ratio, and comment-to-view ratio) to mitigate the sampling bias introduced by platform algorithms and posting times. This ensures the ‘highly-rated’ nature of comments stems from deep emotional quality rather than transient visibility. To ensure the timeliness and representativeness of the samples, the collection process strictly follows the platform’s open protocol, eliminates advertising marketing, machine brushing volume, and duplicate forwarding content, and ultimately retains 12847 valid multimodal samples. In the data preprocessing stage, the original video is first subjected to keyframe extraction and audio separation, and frame rate resampling and loudness normalization are used to eliminate distribution shifts caused by device differences. This article focuses on the conversion of complex and simplified comments, cleaning of special symbols, and restoration of colloquial abbreviations. It also relies on open-source word segmentation tools and a self built local dictionary for fine-grained segmentation. To improve annotation consistency, this study invited six experts with a cross disciplinary background in rural sociology and computational linguistics to conduct independent annotation, covering six categories categorized by a dual-dimension logic of sentiment polarity and interactive value: (1) Emotional Resonance (high sentiment, high value), (2) Cultural Identity (collective sentiment), (3) Practical Advice (high interactive utility), (4) Questioning Feedback (critical interaction), (5) Entertainment and Leisure (low-intensity sentiment), and (6) Neutral Description (objective information). During the annotation process, a double-blind cross validation mechanism was used, and the Cohen’s Kappa coefficient reached 0.87, indicating high reliability in category definition. This article conducts word frequency statistics and semantic clustering on comment texts to generate high-frequency word cloud maps for short video high like comments, as shown in Figure 1:

From the perspective of semantic themes, high-frequency words can be classified into three categories: (1) emotional projection, such as “touched”, “homesick”, “warmth”, and “nostalgia”, reflecting users’ emotional resonance and value identification with rural narratives, confirming the core function of rural short videos as emotional connectors; (2) Cultural identity categories, such as “hometown”, “land”, “agriculture”, and “inheritance”, reflect users’ collective memory and identity belonging to regional cultural symbols. Their frequent appearance validates the representativeness of corpora in modeling local contexts; (3) Examples of practical interactive types of categories: “technology”, “experience”, “sharing”, and “learning”: all represent interactive comment types and have the capability to transmit knowledge as well as mutually assist other members of the community.

To meet exact alignment between the audiovisual and textual stream, a time-award synchronization technique has been developed in which the time of generation (of the

comment) is mapped to the video progress interval being played back. Such types include the use of a dynamic slide window to identify multimodal segments of correspondence. The degree of accuracy of the alignment can be controlled quantitatively by using modal integrity scores. An example of how to calculate an integrity score is shown in equation (1):

$$M_{\text{align}} = \frac{1}{N} \sum_{i=1}^N I(\Delta t(i) \leq \tau \wedge L_{\text{txt}}^{(i)} \geq \lambda) \quad (1)$$

In the equation, $\Delta t(i)$ represents the deviation between the comment posting time and the center of the video clip time, τ is set as the 5-second tolerance threshold, $L_{\text{txt}}^{(i)}$ is the comment character length, λ takes the minimum effective length of 3, and $I(\cdot)$ is the indicator function. Unlike other systems, which may allow user-generated comments that aren’t related to or are out of context with the video being posted, this also provides appropriate filters for comments based on how far away they are from the video’s original context. For example, because there can be significant differences between dialects, pronunciation, and background noise (especially in environments like rural areas), a complete set of features can also be identified using standard audio processing techniques (such as MFCC, Mel frequency cepstral coefficients) including segregating voice activity from other parts of the recording. The statistical information of the final constructed corpus is detailed in Table 1, covering core dimensions such as category distribution, multimodal completeness, and average sequence length. The training and testing sets are divided in an 8:2 ratio to provide a standardized data base for subsequent model training and evaluation.

B. DESIGN OF DUAL STREAM FEATURE ENCODING NETWORK

The dual stream feature encoding network aims to decouple the representation learning path of rural short video audiovisual clues and comment texts, and avoid modal interference caused by early fusion [19,20]. The visual and auditory branches use TAM and Wav2Vec 2.0 (Wave to Vector 2.0) acoustic forward pre-training models to work together. On the visual side, ResNet-50 is used as the backbone network to extract highlevel semantic features for each frame, and then input into the TAM module to calculate frame level weight distribution, focusing on capturing the evolution of cool and warm tones and spatial migration of work scenes [21,22]. The color tone evolution characteristics are quantified through HSV (Hue, Saturation, Value) color space mapping, and their cumulative representation is shown in equation (2):

$$C_t = \sum_{k=1}^T \alpha_k \cdot HSV(F_k), \quad \alpha_k = \frac{\exp(W_v^T h_k)}{\sum_{j=1}^T \exp(W_v^T h_j)} \quad (2)$$



FIGURE 1. High likes, comments, and high-frequency vocabulary in short videos

TABLE 1. Corpus statistical information

Category	Sub-item	Value / Description	Notes
Sample Statistics	Valid multimodal samples	12,847 samples	After deduplication, cleaning, and quality filtering
	Video-comment pairs	1,284 pairs	Each video corresponds to ~ 10 high-engagement comments
Class Distribution	Emotion	3,215 samples	Emotional resonance, sentiment expression, nostalgia
	Culture	2,826 samples	Rural cultural identity and agricultural values
	Advice	2,312 samples	Agricultural techniques and life experience suggestions
	Question	1,542 samples	Inquiries, critiques, and constructive feedback
	Entertainment	1,799 samples	Humorous and entertainment-oriented interactions
	Neutral	1,153 samples	Objective descriptions and factual statements
Textual Features	Average comment length	28.6 ± 12.3 characters	After removing emojis; punctuation included
	Dialect usage ratio	0.347	Verified via dialect lexicon matching and manual inspection
	Metaphor / irony samples	1,927 samples	Identified through semantic annotation
Audio-Visual Features	Average video duration	42.8 ± 18.5 seconds	30-second window around comment generation timestamp
	Keyframe sampling rate	2 frames / second	Used for temporal visual feature modeling
	Audio sampling rate	16 kHz	Valid speech segments after VAD processing
Modality Completeness	Fully aligned multimodal	11,562 samples	Video, audio, and text all available
	Bimodal samples	1,028 samples	Missing audio or low-quality visual frames
	Unimodal samples	257 samples	Text-only samples for robustness evaluation

In the equation, F_k is the k th frame image, h_k is the corresponding hidden layer feature, W_v is the learnable projection matrix, and α_k represents the temporal importance weight. On the audio side, the original waveform is input into Wav2Vec 2.0 to extract continuous acoustic representations. The dialect rhythm fluctuations and environmental sound arousal are modeled using a Long Short Term Memory (LSTM) network, and the acoustic emotion vector P_{dialect} is output. The calculation is shown in equation (3):

$$P_{\text{dialect}} = \text{LSTM}(\text{Wav2Vec2.0}(\text{audio}_{\text{dialect}})) \quad (3)$$

The text branch uses RoBERTa-wwm-ext (Robustly Optimized BERT Approach with Whole Word Masking Extended) as the basic encoder. To address common sense semantic shifts in rural contexts, we specifically fine-tuned the model on a localized dialect lexicon and rural metaphor corpus. This enhances the model's sensitivity to regional expressions and implicit emotions that general sentiment models typically overlook. To mitigate the domain distribution bias, feature regularization constraints are introduced, with the optimization objective shown in equation (4):

$$L_{\text{domain}} = - \sum_{c=1}^C y_c \log p_c + \lambda_1 \|E_{\text{rural}} - E_{\text{base}}\|_2^2 + \lambda_2 KL(q_{\text{rural}} \| q_{\text{base}}) \quad (4)$$

In the equation, y_c and p_c represent the true label and predicted probability, E_{rural} and E_{base} represent the embedding vectors labeled with [CLS] before and after domain fine-tuning, λ_1 and λ_2 are regularization coefficients, and Kullback-Leibler (KL) divergence term is used to constrain the smooth transition of the output distribution of the hidden layer. The dual stream outputs are projected onto a unified dimension d through fully connected layers, obtaining the audio-visual prior $H_{av} \in R^d$ and the text implicit emotion vector $H_t \in R^d$. This design provides a high discriminative feature basis for subsequent cross modal interactions while retaining modality specificity.

C. CROSS-MODAL EMOTION GATING INTERACTION MODULE

To resolve semantic mismatch between rural audiovisual symbols and text, this module employs cross-modal gating. This mechanism achieves fine-grained alignment and

dynamically suppresses feature noise [23,24]. Traditional one-way attention is easily influenced by modal dominant effects, leading to excessive amplification of misleading features introduced by ironic expressions or local metaphors. To this end, a Bidirectional Cross-Attention (BCA) structure is constructed, which uses the audio-visual prior H_{av} and text semantics H_t as query vectors and key value pairs for symmetric weight calculation [25,26]. The alignment score matrix for audio-visual to text is shown in equation (5):

$$A_{av \rightarrow t} = \text{softmax} \left(\frac{Q_{av} K_t^T}{\sqrt{d_k}} \right) V_t, \quad (5)$$

$$Q_{av} = H_{av} W_q, \quad K_t = H_t W_k$$

Similarly, reverse mapping yields $A_{t \rightarrow av}$. After residual connection and layer normalization, the bidirectional output enters the gated fusion layer. To suppress noise interference caused by semantic conflicts, an adaptive gating function $G \in R^d$ is designed, whose activation value is determined by both modal confidence and contextual consistency. The calculation is shown in equation (6):

$$G = \sigma(W_g[A_{av \rightarrow t} \odot A_{t \rightarrow av} + b_g]) \quad (6)$$

In the equation, $\sigma(\cdot)$ is the Sigmoid activation function, W_g and b_g are gating parameters, and $\odot$ represents element wise multiplication. This mechanism automatically reduces the weight contribution of low confidence modalities when the model encounters semantic tension scenarios such as “joyful images but implicit loss in comments”. The enhanced representation H_{enhanced} is generated by fusing the gate mask with the original interactive features, and its expression is shown in equation (7):

$$H_{\text{enhanced}} = G \odot (H_{av} \oplus H_t) + (1 - G) \odot \text{LayerNorm}(H_{av} + H_t) \quad (7)$$

In the equation, $\oplus$ represents the feature concatenation operation. To further enhance the discriminative power of alignment, a mutual information regularization term is introduced to constrain the information fidelity of cross modal representations in the shared subspace, as shown in equation (8):

$$L_{MI} = -E_{p(H_{av}, H_t)}[\log D(H_{av}, H_t)] + E_{p(H_{av})p(H_t)}[\log(1 - D(H_{av}, H_t))] \quad (8)$$

$D(\cdot)$ is a discriminator network used to maximize positive sample mutual information and minimize negative sample associations. This interaction module effectively decouples surface semantics from deep emotional cues [27,28].

D. MULTI-HEAD FINE-GRAINED CLASSIFIER AND OPTIMIZATION OBJECTIVES

The enhanced representation H_{enhanced} carries multimodal emotion information that has been filtered through noise and aligned at a fine-grained level, and is subsequently

connected to a multi-head fine-grained classifier to complete category discrimination [29,30]. The classifier models the complexity of multiple dimensions associated with emotions in comments made by praising rural properties, utilising a structure that places independent heads upon them, wherein each head is focused upon a separate semantic subspace, thus reducing the likelihood of confusion between categories that is caused by the parameter coupling of those heads. The probability distribution of each head output is calculated through linear transformation and normalization function, and its form is shown in equation (9):

$$p_k^{(h)} = \frac{\exp((W_h^{(k)})^T H_{\text{enhanced}} + b_h^{(k)})}{\sum_{j=1}^{C_h} \exp((W_h^{(j)})^T H_{\text{enhanced}} + b_h^{(j)})} \quad (9)$$

In the equation, h represents the classification header index, k is the category label, and W_h is the number of categories in the h -th header. To improve the recognition robustness of minority categories, the category aware weight ω_c is introduced and dynamically adjusted based on the inverse frequency of the training set distribution, as shown in equation (10):

$$\omega_c = \frac{\log(1 + N_{\text{total}}/N_c)}{\sum_{j=1}^C \log(1 + N_{\text{total}}/N_j)} \quad (10)$$

The optimization objective adopts a multi-task joint learning strategy, integrating classification cross entropy and cross modal contrastive alignment loss [31,32]. The alignment loss is constructed based on the Information Noise Contrastive Estimation (InfoNCE) framework, aiming to narrow the distance between audiovisual textual representations of similar samples and distance heterogeneous samples. Its calculation is shown in equation (11):

$$L_{\text{align}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\text{sim}(H_{av}^{(i)}, H_t^{(i)})/\epsilon)}{\sum_{j=1}^B \exp(\text{sim}(H_{av}^{(i)}, H_t^{(j)})/\epsilon)} \quad (11)$$

In the equation, B is the batch size, ϵ is the temperature coefficient, and $\text{sim}(\cdot)$ is the cosine similarity function. The total loss function is a weighted combination of classification term, alignment term, and mutual information regularization term, as shown in equation (12):

$$L_{\text{total}} = \sum_{h=1}^H \omega_h L_{\text{cls}}^{(h)} + \gamma_1 L_{\text{align}} + \gamma_2 L_{MI} \quad (12)$$

γ_1 and γ_2 are hyperparameters, and the optimal ratio is determined through grid search. This loss design explicitly constrains semantic consistency across modal representations while ensuring classification accuracy.

E. EXPERIMENTAL DESIGN

Evaluation indicators' selection is highly relevant to the multi-dimensional discrimination requirements of the high-praise classification task in rural areas. Due to the inherent long tail characteristics of abstract categories such as "emotional resonance" and "cultural identity" in the distribution of self-constructed sample corpuses, a single accuracy figure can easily be dominated by the majority class and can therefore be misleading regarding the model's ability to discriminate between minority classes. Therefore, introducing Macro-F1 as the core indicator, its calculation is shown in equation (13):

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^C \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \quad (13)$$

This indicator measures how well the model performed across all emotion classes by considering both precision and recall together (e.g., averaging them) to give an overall score across all classes. In addition, this paper used both precision and recall to measure the trade-off between capturing implicit emotion cues and creating false alarms. It also introduced AUC (Area Under Curve) for measuring the rank consistency of classifier outputs at different thresholds for making decisions where many emotions have been expressed with high amounts of "likes" on comments that have a wide variety of emotional expressions. The last section of this study addresses a key area of confusion for many rural residents regarding how "useful guidance" can differ from "how someone feels emotionally," with confusion rates being established as the percentage of a target class that gets classified incorrectly as an emotional expression that is not a target class. Its calculation is shown in equation (14):

$$\text{ConfusionRate}_{i,j} = \frac{N_{i \rightarrow j}}{\sum_{k=1}^C N_{i \rightarrow k}} \times 100\% \quad (14)$$

In the equation, $N_{i \rightarrow j}$ represents the number of samples whose true category i is predicted to be category j . This indicator directly quantifies the actual gains of cross modal alignment mechanism in resolving semantic ambiguity and suppressing ironic noise, compensating for the shortcomings of traditional indicators in characterizing fine-grained category boundary ambiguity problems.

The selection of baseline algorithms follows the principle of "simplicity to complexity, modal coverage, and architecture comparison", aiming to comprehensively position the relative advantages of the proposed method in the multimodal emotion computing spectrum. Text Convolutional Neural Network (TextCNN) and Bidirectional Encoder Representation from Transformers are selected as the single modal baselines for text: TextCNN represents the traditional local feature extraction paradigm, which is computationally lightweight but lacks contextual modeling capabilities, and is used to verify the necessity of introducing audiovisual modalities; BERT, as a universal pre-

trained language model, has strong semantic representation ability, but it is difficult to capture implicit emotions of local metaphors and dialect spoken language without domain adaptation. It can be used as a reference for the upper limit of text branches. Cross modal fusion baseline selection ViLBERT (Vision and Language BERT) and Multi-modal Transformer: ViLBERT adopts a dual tower structure combined with a joint self attention mechanism, which is a classic architecture for audio-visual language alignment. However, it faces problems of pre-training domain offset and computational redundancy in non-standard rural scenes. Multi-modal Transformer represents an end-to-end global fusion strategy, which directly concatenates multi-source features for deep interaction and is commonly used for pan entertainment content classification. This article uses the above model as a comparison to reveal the unique value of the architecture in the design path of "domain prior injection fine-grained gating alignment interpretable multihead output". All baselines are trained on the same hardware environment, and the optimizer uses AdamW (Adaptive Moment Estimation with Weight Decay). The initial learning rate is set to 2×10^{-5} , the batch size is fixed at 32, and the training epochs are 50.

IV. RESULTS AND DISCUSSION

A. PERFORMANCE COMPARISON OF MULTIPLE MODELS

To evaluate the comprehensive performance of the proposed Cross modal Emotion Alignment Classification Architecture (CEA-CA), this paper selects text unimodal baseline and cross modal fusion baseline as controls, and quantitatively compares them from three dimensions: classification accuracy, robustness, and fine-grained discriminative ability. The aim is to verify the actual gain of the dual stream encoding and gating interaction mechanism in resolving semantic ambiguity in rural context. The results are shown in Table 2:

Table 2 shows the comparative results of each model on the core indicators. The CEA-CA architecture in this article achieved 89.4% on Macro-F1, which is 7.8 percentage points higher than the optimal single modal baseline BERT. The precision and recall rates reached 91.2% and 87.6%, respectively, indicating that the model achieved a good balance in suppressing false positives and capturing implicit emotional cues. In addition, the Area Under Curve (AUC) reached 0.943, verifying the ranking stability of the model under different decision thresholds. The above results confirm that the collaborative design of domain prior injection and fine-grained gating alignment is the key path to improving the accuracy of rural high like comment classification.

B. IN DEPTH ANALYSIS AND VISUALIZATION OF CONFUSION CATEGORIES

To address the core pain point of category confusion, a heatmap was used to visualize the misjudgment flow

TABLE 2. Performance comparison results of multiple models

Model Architecture	Macro-F1(%)	Accuracy (%)	Precision (%)	Recall (%)	AUC	Confusion Rate (%)
TextCNN	72.3 ± 1.8	74.1 ± 1.5	73.6 ± 2.1	71.0 ± 2.4	0.812	28.7 ± 3.2
BERT	81.6 ± 1.2	83.4 ± 0.9	82.9 ± 1.4	80.3 ± 1.6	0.891	19.4 ± 2.1
ViLBERT	84.2 ± 1.0	85.7 ± 0.8	85.1 ± 1.2	83.3 ± 1.3	0.918	16.8 ± 1.8
Multi-modal Transformer	85.1 ± 0.9	86.3 ± 0.7	86.0 ± 1.0	84.2 ± 1.1	0.925	15.3 ± 1.5
CEA-CA	89.4 ± 0.6	90.8 ± 0.5	91.2 ± 0.7	87.6 ± 0.8	0.943	11.9 ± 1.2

between each label, intuitively revealing the fuzzy areas of category boundaries and the sources of uncertainty in model decision-making. The results are shown in Figure 2:

Figure 2 shows the classification confusion of six types of semantic labels. From the overall results, the values in the diagonal region are generally maintained at a high level, above 85%, indicating that the model's recognition ability in various categories is relatively stable and has strong overall discriminative performance. Among them, the correct classification rates of "neutral" and "entertainment" categories are the highest, reaching 95.1% and 92.8% respectively, indicating that the model has better adaptability on samples with clear semantic boundaries. In contrast, there is a certain degree of confusion between "advice" and "emotion", and their non diagonal elements are relatively high, reflecting that the model still has a certain degree of discrimination difficulty in situations where semantic expressions are similar. From the perspective of misclassification distribution, most of them are below 5%, with errors mainly concentrated between a few adjacent semantic categories, and the overall error propagation range is relatively limited.

To further verify the model's ability to focus on key features, Figure 3 visualizes the distribution of temporal attention weights on video high like comments, with the time axis as the horizontal axis and feature channels as the vertical axis. By encoding the weight intensity through contour lines, it reveals the collaborative attention pattern of the model on clues such as cool and warm tones, dialect rhythm, etc., at the peak of high like comment emotions, providing temporal evidence for interpretability analysis.

From Figure 3, it can be seen that there are significant stage differences in the response of different feature channels on the time axis. Within the 15-25 second interval, each channel generally shows a significant enhancement trend, indicating that this time period is a concentrated outbreak area of emotional expression and semantic information. Dialects, rhythms and tone show high levels of attentiveness in relation to motion changed here (i.e., high attention weights). The peaks of dialectically produced rhythm change are where the models most heavily utilized moves towards semantic based data and then utilized emotional speech and visual style change motions to derive the majority of its semantic data. Scene change, sound of the environment awakening at a moderate (less than peak) intensity level indicates that the background/scene is used as a secondary data point and did not contribute significantly to make the changes that occurred (i.e., secondary background data). Non-peak time frames show a relatively flat attention dis-

tribution, meaning that the model did not rely so much on the more semiotic properties of background scenes/video data and filtered in most ways with regards to semantic based data during this time frame. The comment generation time was indicated (by the location of the red dashed line) near the area of most significant value (centralized focus) and correspondingly close to the area of highest value for individual comment generation. The results indicate that the multimodal attention mechanism adaptively focuses on emotional peaks. To mitigate the risk of causal inversion where high-engagement comments might retroactively direct visual attention, our ablation studies (Section 4.3) specifically isolate the influence of multimodal features via the gating module, confirming that the initial audiovisual stimuli act as primary causal drivers for the generation of these specific emotional categories rather than merely following user attention shifts.

C. ABLATION EXPERIMENT

To quantify the independent contributions and synergistic effects of each core component on overall performance, this study designed four sets of ablation experiments, which removed the cross modal gating module, replaced bidirectional cross attention with a unidirectional mechanism, and turned off domain fine-tuning (Fine Tuning, FT). These experiments intuitively revealed the performance degradation mode caused by module loss, thereby verifying the necessity and rationality of the architecture design. The ablation results are shown in Figure 4:

Figure 4 compares the performance of the complete model with different ablation settings on metrics. Overall, the complete model achieved the best results in all indicators, indicating that the synergistic effect of each module can effectively improve the performance of the model. After removing the gating module, multiple indicators showed a significant decrease, particularly in Precision and AUC, indicating that the gating mechanism plays an important role in screening key features and stabilizing the discrimination process. After removing the bidirectional cross attention module, the model performance also showed a certain degree of decline, indicating that this structure helps to capture more comprehensive contextual information and enhance feature expression ability. In contrast, the performance degradation caused by closing domain fine-tuning is more significant, with significant reductions in various indicators, reflecting the importance of domain adaptation in improving model generalization ability. Overall, all functional modules contribute positively to the overall performance, with do-

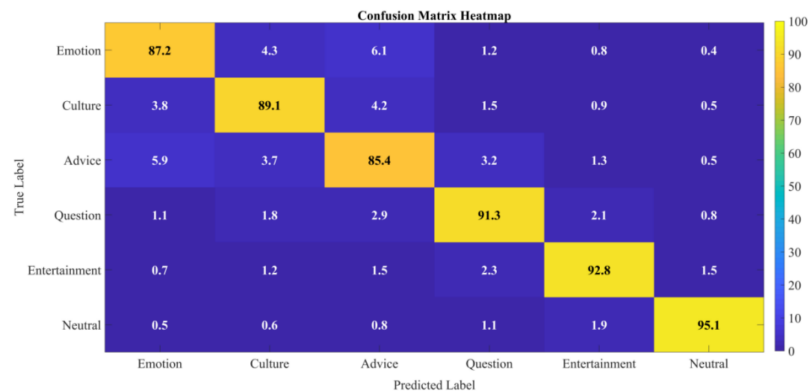


FIGURE 2. Classification confusion of semantic labels

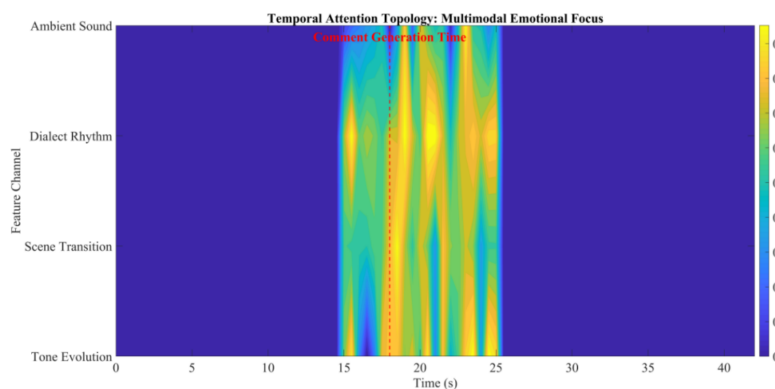


FIGURE 3. Temporal attention changes

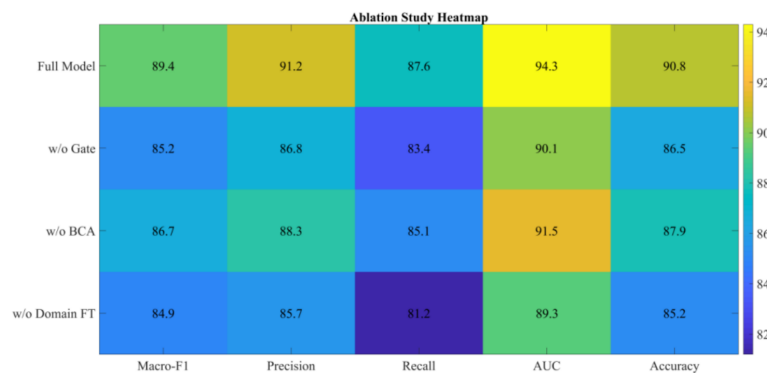


FIGURE 4. Results of ablation experiment

main fine-tuning and gating mechanisms being particularly critical, validating the rationality and effectiveness of the proposed model structure design.

V. CONCLUSION

The current research contributes a new architecture aimed at classifying emotions between two different media type (cross modal) in a rural setting, to address the problems associated with an incongruity of meanings between an audio-visual representation of an emotion and the fragmented representation of the text that often accompanies

that representation, as well as the “black box” nature of the emotion generating mechanisms in the analysis of the most widely liked comments on a given set of rural short videos. This cross modal emotion alignment classification framework employs a dual stream feature decoupling and domain adaptive fine-tuning methodology to capture the evolution of scene tone and rhythm of different dialects, as well as implicit representations of emotion. It further employs a bidirectional cross attention methodology combined with an adaptive gating mechanism to create a finer

resolution of modal alignment between the different modalities, thereby alleviating the effects of ironic expressions or local metaphorical expressions that tend to create drift in the features themselves. Experimental validation of the model demonstrates that the proposed approach possesses a significant structural advantage in establishing complex semantic mappings and discriminating between subtle emotions. The decision-making process associated with the method also offers a clear attribution of contributions from the various modalities used in the classification process. The contributions made through this research advance the understanding of the challenges associated with adapting the traditional fusion paradigm in rural digital contexts by providing a theoretical foundation for accurate agricultural content recommendations, and for implementing quantitative methods for governing ecological comments, and providing an interpretable emotional computation across multiple modalities.

AUTHOR CONTRIBUTIONS

Yimeng Gu designed, collected and analyzed the data, and drafted the manuscript. Yimeng Gu conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Yimeng Gu participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research was supported by,

2021 General Special Scientific Research Project of Shaanxi Provincial Department of Education ,A Study on Short Video Dissemination of Rural Culture in Shaanxi in the Context of Beautiful Countryside,21JK0276.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] J. Yu, "A Study on the narrative space of rural short video: A case study of rural short video in TikTok platform," *Frontiers in Art Research*, vol. 5, no. 2, pp. 6-10, 2023, doi: 10.25236/FAR.2023.050202.
- [2] S. Zou, "Curating a scopic contact zone: Short video, rural performativity, and the mediatization of socio-spatial order in China," *Television & New Media*, vol. 24, no. 4, pp. 452-470, 2023, doi: 10.1177/15274764221128925.
- [3] Z. Shuang, "Rural Short Videos' New Farmer Image Construction and Communication Strategy Presentation—the Example of "Winning the Heart," *Academic Journal of Humanities & Social Sciences*, vol. 6, no. 14, pp. 63-68, 2023, doi: 10.25236/AJHSS.2023.061411.
- [4] D. Huo, P. Zou, and Y. Lu, "Live vs," static comments: Empirical analysis of their differential effects on user evaluation of online videos. *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 20, no. 2, pp. 102-125, 2025, doi: 10.3390/jtaer2002102.
- [5] C. Kuchler, A. Stoll, M. Ziegele, and T. K. Naab, "Gender-related differences in online comment sections: Findings from a large-scale content analysis of commenting behavior," *Social Science Computer Review*, vol. 41, no. 3, pp. 728-747, 2023, doi: 10.1177/08944393211052042.
- [6] X. Dong, H. Liu, N. Xi, J. Liao, and Z. Yang, "Short video marketing: what, when and how short-branded videos facilitate consumer engagement," *Internet Research*, vol. 34, no. 3, pp. 1104-1128, 2024, doi: 10.1108/INTR-02-2022-0121.
- [7] P. Kumar, G. Bhatt, O. Ingle, D. Goyal, and B. Raman, "Affective Feedback Synthesis Towards Multimodal Text and Image Data," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 6, pp. 1-23, 2023, doi: 10.1145/3589186.
- [8] S. Wang, Y. Zhai, G. Xu, and N. Wang, "Multi-modal affective computing: An application in teaching evaluation based on combined processing of texts and images," *Traitement du Signal*, vol. 40, no. 2, pp. 533-542, 2023, doi: 10.18280/ts.400212.
- [9] M. Firdaus, G. V. Singh, A. Ekbal, and P. Bhattacharyya, "Affect-GCN: a multimodal graph convolutional network for multiemotion with intensity recognition and sentiment analysis in dialogues," *Multimedia Tools and Applications*, vol. 82, no. 28, pp. 43251-43272, 2023, doi: 10.1007/s11042-023-14885-1.
- [10] B. Schuller, A. Mallol-Ragolta, A. P. Almansa, I. Tsangko, M. M. Amin, A. Semertizidou, and S. Amiriparian, "Affective computing has changed: The foundation model disruption," *npj Artificial Intelligence*, vol. 2, no. 1, pp. 16-26, 2026, doi: 10.1038/s44387-025-00061-3.
- [11] E. Zhang, H. Zong, X. Li, M. Feng, and J. Ren, "ICSF: Integrating inter-modal and cross-modal learning framework for self-supervised heterogeneous change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, no. 1, pp. 1-16, 2024, doi: 10.1109/TGRS.2024.3519195.
- [12] T. J. Wen, C. H. Chuan, G. Anghelcev, S. Sar, J. T. Yun, and Y. Xu, "Infusing affective computing models into advertising research on emotions," *Journal of advertising*, vol. 53, no. 5, pp. 710-731, 2024, doi: 10.1080/00913367.2024.2409254.
- [13] Z. Li, L. Zhang, K. Zhang, Y. Zhang, and Z. Mao, "Improving image-text matching with bidirectional consistency of crossmodal alignment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6590-6607, 2024, doi: 10.1109/TCSVT.2024.3369656.
- [14] A. Zavras, D. Michail, B. Demir, and I. Papoutsis, "Mind the modality gap: Towards a remote sensing vision-language model via cross-modal alignment," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 228, no. 1, pp. 270-287, 2025, doi: 10.1016/j.isprsjprs.2025.06.019.
- [15] M. Lan, F. Rong, H. Jiao, Z. Gao, and L. Zhang, "Language query-based transformer with multiscale cross-modal alignment for visual grounding on remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, no. 1, pp. 1-13, 2024, doi: 10.1109/TGRS.2024.3407598.
- [16] Y. Ye, J. Zhang, Z. Chen, and Y. Xia, "CADs: A self-supervised learner via cross-modal alignment and deep self-distillation for CT volume segmentation," *IEEE Transactions on Medical Imaging*, vol. 44, no. 1, pp. 118-129, 2024, doi: 10.1109/TMI.2024.3431916.
- [17] L. Chen, Z. Deng, L. Liu, and S. Yin, "Multilevel semantic interaction alignment for video-text cross-modal retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6559-6575, 2024, doi: 10.1109/TCSVT.2024.3360530.
- [18] L. Ding, L. Liu, Y. Huang, C. Li, C. Zhang, W. Wang, and L. Wang, "Text-to-image vehicle re-identification: Multi-scale multi-view cross-modal alignment network and a unified benchmark," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 7, pp. 7673-7686, 2024, doi: 10.1109/TITS.2023.3348599.
- [19] F. Bi, T. He, Y. Xie, and X. Luo, "Two-stream graph convolutional network-incorporated latent feature analysis," *IEEE Transactions on Services Computing*, vol. 16, no. 4, pp. 3027-3042, 2023, doi: 10.1109/TSC.2023.3241659.
- [20] C. Cao, Y. Lu, and Y. Zhang, "Context recovery and knowledge retrieval: A novel two-stream framework for video anomaly detection," *IEEE Transactions on Image Processing*, vol. 33, no. 1, pp. 1810-1825, 2024, doi: 10.1109/TIP.2024.3372466.
- [21] K. Vo, S. Truong, K. Yamazaki, B. Raj, M. T. Tran, and N. Le, "Aoe-net: Entities interactions modeling with adaptive attention mechanism for temporal action proposals generation," *International Journal of Computer Vision*, vol. 131, no. 1, pp. 302-323, 2023, doi: 10.1007/s11263-022-01702-9.
- [22] S. Park, M. Mark, B. Park, and H. Hong, "Using speaker-specific emotion representations in wav2vec 2.0-based modules for speech emotion recog-

- nitition," *Computers, Materials and Continua*, vol. 77, no. 1, pp. 1009-1030, 2023, doi: 10.32604/cmc.2023.041332.
- [23] N. Wang and Q. Wang, "Dynamic weighted gating for enhanced cross-modal interaction in multimodal sentiment analysis," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 1, pp. 1-19, 2024, doi: 10.1145/3702996.
- [24] H. Kumar and M. Aruldoss, "Gated Cross-Modal Fusion Mechanism for Audio-Video-based Emotion Recognition," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 20835-20841, 2025, doi: 10.48084/etasr.9430.
- [25] A. G. Mengara Mengara and Y. K. Moon, "CAG-MoE: Multimodal emotion recognition with cross-attention gated mixture of experts," *Mathematics*, vol. 13, no. 12, pp. 1907-1944, 2025, doi: 10.3390/math13121907.
- [26] S. Zhang, J. Liu, Y. Jiao, Y. Zhang, L. Chen, and K. Li, "A multi-modal semantic fusion network with cross-modal alignment for multimodal sentiment analysis," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 10, pp. 1-22, 2025, doi: 10.1145/3744648.
- [27] R. G. Praveen and J. Alam, "Incongruity-aware cross-modal attention for audio-visual fusion in dimensional emotion recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 18, no. 3, pp. 444-458, 2024, doi: 10.1109/JSTSP.2024.3422823.
- [28] B. Yu, C. Li, and Z. Shi, "Multi-grained feature gating fusion network for multimodal sentiment analysis: B," *Yu et al. Knowledge and Information Systems*, vol. 67, no. 8, pp. 6879-6905, 2025, doi: 10.1007/s10115-025-02446-x.
- [29] R. Ji, J. Li, L. Zhang, J. Liu, and Y. Wu, "Dual transformer with multi-grained assembly for fine-grained visual classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 5009-5021, 2023, doi: 10.1109/TCSVT.2023.3248791.
- [30] B. Zhong, "Fine-grained sentiment analysis using multidimensional feature fusion and GCN," *Journal of Information and Telecommunication*, vol. 9, no. 1, pp. 91-112, 2025, doi: 10.1080/24751839.2024.2386785 .
- [31] Y. Cui, C. Han, and D. Liu, "Collaborative multi-task learning for multi-object tracking and segmentation," *Journal on Autonomous Transportation Systems*, vol. 1, no. 2, pp. 1-23, 2024, doi: 10.1145/3632181.
- [32] X. Hao, S. Yang, R. Liu, Z. Feng, T. Peng, and B. Huang, "SMTCC-L: Continuous learning via selective multi-task coordination for adaptive signal classification," *IEEE Transactions on Cognitive Communications and Networking*, vol. 11, no. 3, pp. 1664-1681, 2024, doi: 10.1109/TCCN.2024.3485083 .