Exploring the Differences in Audience Emotional Arousal Patterns Between Virtual and Real-Life Anchors Using BERT Multimodal Sentiment Analysis

Di Zhang[1], Shiqi Li[2], Shuhan Xu[3]
[1]Department of New Media Communication, School of Humanities, Jiamusi University, Jiamusi 154007, Heilongjiang, China
[2]College of Information Science and Electronic Technology, Jiamusi University, Jiamusi 154007, Heilongjiang, China
[3]Department of Broadcasting and Hosting Art, School of Music, Jiamusi University, Jiamusi 154007, Heilongjiang, China

Corresponding author: Shuhan Xu (e-mail: 13846192497@163.com)

ABSTRACT This study investigates audience emotional arousal patterns in virtual versus real-life anchors using a multi-channel signal processing framework. A Text-Anchored Cross-modal Alignment for Arousal Modeling (TACAM) method is proposed, integrating text, visual, and audio signals from live streaming platforms. Textual features are extracted via fine-tuned BERT-large, while visual motion energy and audio prosodic features are temporally aligned and fused through a cross-modal attention mechanism. A Transformer-based temporal regression network generates high-resolution emotional arousal curves, capturing dynamic intensity, rise rate, and high-arousal duration. Experimental results show that the model achieves a high-arousal duration ratio of 0.38 for virtual anchors and 0.52 for real-life anchors, with adaptive weighting reflecting modality reliability (text weight 76% for virtual anchors). By interpreting audience reactions as multi-channel signal flows and TACAM as a signal fusion and temporal alignment system with closed-loop modulation, this work provides an engineering-oriented perspective for designing robust, interpretable, and quantitative frameworks for human-computer emotional interaction, enabling more accurate modeling of real-time affective dynamics in digital media environments.

INDEX TERMS emotional arousal, cross-modal signal processing, temporal regression, multimodal alignment

I. INTRODUCTION

THE audience's emotional interaction with digital media is evolving rapidly as virtual anchors interweave collaborative communication into the global digital ecosystem. This convergence of virtual presence and digital narrative creates a multi-layered sensory fabric that redefines how cultural information is perceived and emotionally processed by a global audience. Virtual anchors use technologically produced images to provide users with a unique interactive experience that mimics human-like qualities. By analyzing the similarities and differences between the emotional induction pathways associated with virtual anchors versus those associated with traditional (or "real") anchors, researchers can gain insights into the ways in which audiences perceive and connect emotionally with both types of media. Understanding these pathways enables researchers to optimize the user experience when interacting with virtual idols in the digital marketplace, while simultaneously assisting researchers who are working to develop a better theoretical framework for emotional authenticity, social presence, cognitive projection, and other phenomena associated with human-machine communication [1,2]. In light of the above, it is imperative that researchers assess, on a systematic basis, the ways in which virtual anchors and real anchors stimulate the same or similar patterns of emotional arousal in their respective audiences in order to develop an understanding of how emotional communication occurs within today's digital age. The use of new technical analysis tools, such as BERT multimodal sentiment analysis, to quantitatively identify these differences in emotional arousal patterns has emerged as a major area of interest among researchers today. It is worth noting that emotional arousal, as quantified in this work, reflects the intensity dimension of affective response rather than its valence. While elevated arousal may correspond to either positive (e.g., excitement) or negative (e.g., stress) states, our focus is on its dynamic patterns as an indicator of emotional engagement level, which is a key metric in evaluating media interaction effectiveness.

The current state of research has three main obstacles. Firstly, the audiovisual representation of virtual anchors is often very stylized (i.e., using fixed facial expressions and synthesized voices that lack natural rhythm, as well as pre-

defined driver movements for the body). Therefore, there is a great deal of bias when extracting features from the feature extraction level in traditional multimodal emotion analysis methods that model actual human conduct [3,4]. Secondly, since emotion might be described as the intensity of emotional activation in this dimension, it might be necessary to model emotional arousal dynamically and with a high temporal resolution. However, existing emotion recognition research mostly focuses on discrete emotion categories or static valence judgments, making it difficult to capture the instantaneous fluctuations and persistent characteristics of arousal intensity [5,6]. Third, in virtual anchor scenarios, audience emotional expression is highly dependent on text interactions such as bullet comments and reviews. However, visual and audio modal information is often scarce due to limitations in character modeling, resulting in an imbalance of modal weights during multimodal fusion and further weakening the reliability of cross-anchor type comparisons [7,8].

To address the aforementioned issues, recent research has attempted to approach the problem from different angles. Deep learning-based multimodal fusion frameworks have achieved high-precision sentiment regression on real human video datasets, but their assumption that each modality possesses rich semantic information leads to a significant performance degradation in virtual anchor scenarios due to the degradation of visual and speech modal information [9,10]. Some works have introduced pre-trained language models such as BERT to model text sentiment, achieving good results in live stream comment analysis, but these rely solely on a single modality and cannot characterize the multi-channel sentiment coordination mechanism [11,12]. Other studies have attempted to assess the emotional appeal of virtual anchors through social presence scales or questionnaires. While these methods have theoretical explanatory power, they lack objective, continuous, and quantifiable behavioral or physiological data support, making it difficult to support fine-grained dynamic arousal comparison [13,14]. Despite the progress made in multimodal sentiment analysis and virtual idol research, existing methods still haven't solved the problem of how to construct a multimodal sentiment modeling framework centered on the arousal dimension and aligned across anchor types under the condition of limited audiovisual signals for virtual anchors.

To address the multimodal imbalance and lack of dynamic arousal caused by stylized virtual expressions, this paper constructs a cross-modal aligned emotional model anchored by BERT semantics. This approach ensures that the "weave density" of semantic information and emotional fluctuations remains consistent, providing a high-precision digital textile for more authentic virtual anchor interactions. A fine-tuned BERT-large is used to perform context-aware encoding on time-synchronized audience text interaction sequences, which are then mapped to continuous arousal values via a regression head. Based on this, visual motion energy (based on optical flow amplitude integral) and speech prosodic features aligned with the text are extracted. A cross-modal alignment module is designed, using the text arousal sequence as a guiding signal, and dynamically adjusting the contribution weights of audiovisual features at each time step through a learnable attention mechanism. Finally, the aligned three-modal features are concatenated and input into a Transformer temporal regression network to generate a high-fidelity emotional arousal curve. This framework transforms BERT from a simple text analysis tool into a semantic benchmark for multimodal emotion modeling, effectively mitigating modal noise caused by unnatural audiovisual expressions in virtual content, and achieving comparable quantification between virtual and real anchors in terms of emotional arousal intensity, rhythm, and modal dependency structure.

II. EMOTIONAL AROUSAL MODELING BASED ON COMPARISON BETWEEN VIRTUAL AND REAL-LIFE ANCHORS

A. OVERALL ARCHITECTURE DESIGN FOR ANCHOR EMOTIONAL AROUSAL MODELING

This paper constructs a cross-modal emotional arousal modeling framework dominated by text semantics to address the multimodal emotional modeling bias caused by the limited audiovisual expression of virtual anchors. The overall implementation framework is shown in Figure 1.

Figure 1 shows that the framework takes time-synchronized viewer comments, live video, and audio as input, and uses a fine-tuned BERT-large to perform context-aware encoding of the comment sequence and map it to the continuous arousal dimension. Simultaneously, visual motion energy and speech prosodic features aligned with the text window are extracted; then, a cross-modal alignment module is introduced, using the text arousal sequence as a guiding signal, and dynamically calibrating the contribution weights of the audiovisual modal at each time step through a learnable attention mechanism. Finally, the three-modal alignment features are input into the Transformer temporal regression network, outputting a high-resolution emotional arousal curve, achieving comparable modeling of the emotional arousal modes of virtual and real anchors.

B. BERT-BASED TEXT SEMANTIC AROUSAL ENCODING

The text semantic arousal encoding module takes time-aligned viewer comment sequences as input and uses a pre-trained BERT-large model for context-aware deep semantic representation [15,16]. Given a live stream segment divided into T equal-length time windows (each window is 10 seconds), let the set of bullet comments (comments) in the t-th window be denoted as $U_t = \{u_{t1}, u_{t2}, \ldots, u_{tn_t}\}$. First, all bullet comments are concatenated into a single sentence, and special markers are added to form the input sequence $X_t = [\mathrm{CLS}] \oplus u_{t1} \oplus \mathrm{SEP} \oplus \cdots \oplus u_{tn_t} \oplus \mathrm{SEP}$. This sequence is then mapped to a context embedding by a BERT encoder.

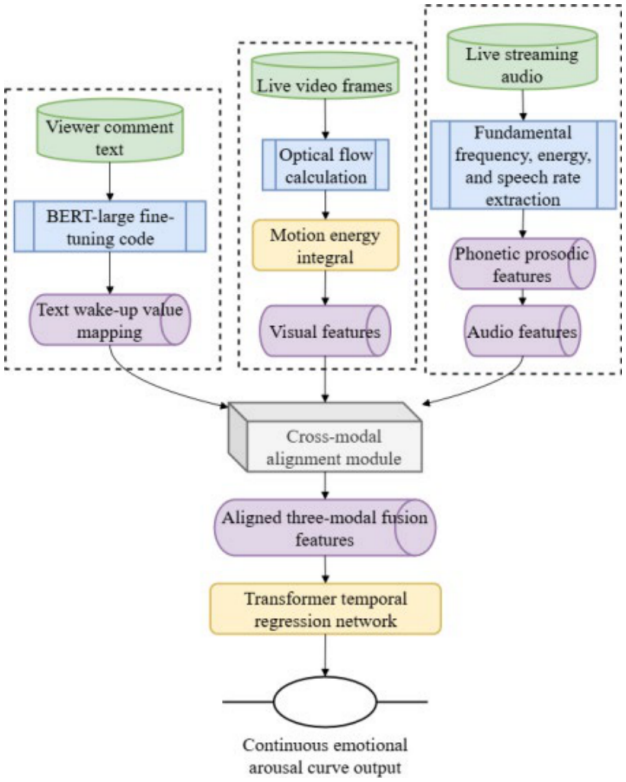

FIGURE 1. Overall architecture of the TACAM emotional arousal framework

$$H_t = \mathrm{BERT}(X_t) \in \mathbb{R}^{L\times d}. \tag{1}$$

In Equation 1, L is the sequence length, and $d = 1024$ is the hidden layer dimension. The output vector $h_t^{[\mathrm{CLS}]} \in \mathbb{R}^d$ at position [CLS] is taken as the global semantic representation of that time window. It is then mapped to continuous arousal values using a learnable linear regression head $w \in \mathbb{R}^d$ and a bias $b \in \mathbb{R}$:

$$a_t = \sigma\left(w^\top h_t^{[\mathrm{CLS}]} + b\right). \tag{2}$$

In Equation 2, $a_t \in [0, 1]$ represents the emotional arousal intensity of the t-th time window, and $\sigma(\cdot)$ is the Sigmoid activation function to constrain the output range [17,18]. The BERT-large parameters and the regression head are jointly fine-tuned on a labeled arousal dataset, aiming to minimize the mean squared error between the predicted arousal value and the reference arousal label, using the AdamW optimizer with a learning rate of 2e-5 and L2 weight decay of 0.01, thereby establishing an end-to-end mapping relationship between bullet screen semantics and emotional arousal intensity. The ground-truth arousal labels were derived from a multi-annotator subjective assessment process. For each 10-second time window, three trained annotators independently rated the overall intensity of audience emotional activation on a scale of 0–1, based on the aggregated content of bullet comments (reflecting collective sentiment). Annotators were instructed to focus on arousal intensity (level of activation) regardless of valence. The final label for each window is the average of the three ratings, with inter-annotator agreement exceeding 0.85. While subjective, this method provides a consistent and reproducible proxy for audience arousal in the absence of large-scale real-time physiological data, aligning with common practice in affective computing for user-generated content.

C. AUDIOVISUAL MODAL FEATURE EXTRACTION

The audiovisual modal feature extraction module extracts low-dimensional but emotionally discriminative physical features from video and audio streams that are strictly time-aligned with the text window. The video modality takes the continuous frame sequence $\{I_\tau\}_{\tau=t_1}^{t_2}$ within a 10-second window as input and uses the Farneback dense optical flow algorithm to calculate the displacement field $V_\tau \in \mathbb{R}^{H\times W\times 2}$ between adjacent frames, where H and W are the frame height and width, respectively [19,20]. The visual motion energy m_t is defined as the spatiotemporal integral of all optical flow amplitudes within the window:

$$m_t = \frac{1}{N}\sum_{\tau=t_1}^{t_2}\sum_{i=1}^{H}\sum_{j=1}^{W}\left\|V_\tau(i,j)\right\|_2. \tag{3}$$

N in Equation 3 is the normalization factor, and m_t reflects the overall dynamic intensity of the anchor's picture, serving as the emotional proxy feature of the visual modality. It is acknowledged that in virtual anchor scenarios, optical flow captures pixel displacement that may originate from pre-programmed animations rather than spontaneous emo-

tional expression. Nevertheless, m_t serves as a quantitative, modality-consistent proxy for overall screen dynamism, providing a baseline visual input whose relative informativeness is later calibrated by the cross-modal alignment module. The audio modality extracts ternary prosodic features from the synchronized single-channel speech signal: the fundamental frequency f_0 is calculated by the autocorrelation method and the mean $\bar{f}_0$ within the window is taken, and the short-time energy E is defined as:

$$E = \frac{1}{M}\sum_{n=1}^{M} s^2(n). \tag{4}$$

The speech rate r is estimated in syllables per second and calibrated using a forced alignment model. These three elements constitute the audio feature vector $a_t = [\bar{f}_0, E, r]^\top \in \mathbb{R}^3$. Both the video feature m_t and the audio feature a_t are Z-score normalized and resampled to be perfectly aligned with the text wake-up sequence $\{a_t\}_{t=1}^{T}$, forming a multimodal input with uniform temporal resolution [21,22].

D. CROSS-MODAL ALIGNMENT MODULE

The cross-modal alignment module uses the text wake-up sequence as a semantic anchor to dynamically calibrate the contribution of the audiovisual modality to emotional arousal modeling [23,24]. Let the text wake-up value at time step t be $a_t \in [0,1]$, and its corresponding learnable query vector be $q_t = W_q a_t \in \mathbb{R}^{d_k}$, where $W_q \in \mathbb{R}^{d_k \times 1}$ is the projection matrix; the visual feature m_t and the audio feature a_t are linearly mapped to obtain the key vectors $k_t^v = W_k^v m_t$, $k_t^a = W_k^a a_t$, and the value vectors $v_t^v = W_v^v m_t$ and $v_t^a = W_v^a a_t$, with dimensions d_k and d_v, respectively. The attention weights of the text for the visual and audio aspects are calculated:

$$\alpha_t^v = \frac{\exp\left(\frac{q_t^\top k_t^v}{\sqrt{d_k}}\right)}{\sum_{m\in\{v,a\}} \exp\left(\frac{q_t^\top k_t^m}{\sqrt{d_k}}\right)}. \tag{5}$$

$$\alpha_t^a = 1 - \alpha_t^v. \tag{6}$$

The final fused feature is $z_t = \alpha_t^v v_t^v + \alpha_t^a v_t^a$. This mechanism enables the model to automatically increase text dependence when the virtual anchor's audiovisual signal is weak, and to achieve balanced fusion when the real anchor's multimodal features are rich. The original dimensions, normalization methods, and projection configurations of each modality are shown in Table 1.

TABLE 1. Multimodal feature dimensions and alignment configuration

Modal	Original features	Original dimension	Normalization method	Projection target dimension	Projection method
Text	BERT [CLS] embedded	1024	LayerNorm	$d_k = 64$	Linear + Sigmoid (projecting the generated scalar a_t)
Visual	Action energy (optical flow amplitude integral)	1	Z-score	$d_k = 64$	Linear transformations W_k^v, W_v^v
Audio	[Fundamental frequency, energy, speech rate]	3	Z-score	$d_k = 64$	Linear transformations W_k^a, W_v^a

Table 1 details the feature representations and alignment configurations of the text, visual, and audio modalities within the TACAM framework. The text modality is played in BERT [CLS] format and aligned to an arousal scalar via LayerNorm and sigmoid before being projected onto the alignment space. The modalities of visual (1D) and audio (3D) include action energy and prosodic features, respectively. Both are normalized using Z-scores and mapped to a unified 64-dimensional key/value space through independent linear transformations to support attention alignment. The end result is that high-dimensional (verbo-semantically) semantic text represents the majority or parameter of the alignment process while at the same time holds on to the emotional discriminativeness of very low-dimensional audio-visual features, providing a modal comparability basis for cross-anchor type emotional arousal modeling.

E. EMOTIONAL AROUSAL TEMPORAL REGRESSION

The emotional arousal temporal regression module gets the fused feature sequence $\{z_t\}_{t=1}^{T}$ from the cross-modal alignment module, with each instance $z_t \in \mathbb{R}^{d_v}$ being an alignment representation at time step t. Each element of the sequence is first embedded with positional encoding before being processed through a 6-layer Transformer encoder (with 8 attention heads, hidden dimension is 64, and dropout rate of 0.1) for temporal regression. Each layer consists of multi-head self-attention and a feedforward network [25-26]. Let the output of the l-th layer be $Z^{(l)} = [z_1^{(l)}, \ldots, z_T^{(l)}]$, then:

$$Z^{(l)} = \mathrm{LayerNorm}\left(Z^{(l-1)} + \mathrm{MHSA}\left(Z^{(l-1)}\right)\right). \tag{7}$$

$$\mathrm{MHSA}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O. \tag{8}$$

$\mathrm{head}_i = \mathrm{softmax}\left(\frac{QW_i^Q (KW_i^K)^\top}{\sqrt{d_k}}\right) VW_i^V$ is used to capture the long-range dependence and local fluctuations of arousal intensity over time. Finally, the top-level output is input into a fully connected regression head after global average pooling to generate a continuous arousal value $\hat{a}_t =$

$w_r^\top z_t^{(6)} + b_r$, where w_r and b_r are learnable parameters [27-28]. This structure effectively models the temporal dynamics of emotional arousal and provides high-resolution output for cross-anchor type comparisons.

III. EXPERIMENTS AND EVALUATION

A. DATASET CONSTRUCTION AND ANCHOR GROUPING

The dataset is constructed based on 200 complete live stream replays collected from publicly available live streaming platforms. 100 of these are from Chinese virtual anchors, and 100 are from real-life anchors. The two types of anchors are strictly paired in terms of content type, covering three mainstream live streaming scenarios: game streaming, daily chat, and product sales. Each category comprises approximately one-third of the content to control the confusion effect of content themes on emotional arousal. While this pairing controls for broad topic categories, we acknowledge that finer-grained production elements (e.g., editing pace, use of background music, or clickbait tactics) may still vary and constitute a potential confounding factor in the observed arousal differences. All live stream videos have a resolution of at least 1080p, a frame rate of 30fps, and an audio sampling rate of 16kHz. The accompanying bullet screen data includes timestamps accurate to milliseconds.

Only one representative live stream from each anchor is selected, with a duration of at least 30 minutes, to ensure sufficient space for emotional dynamics to develop. Although only one live stream was selected for each streamer to control for confounding effects from individual style and content theme, the selected streams all met the following criteria: 1) they belonged to the streamer's regular content type; 2) their interaction metrics were located at the 40–60th percentile of their last 30 live streams; and 3) there were no obvious technical anomalies. While this strategy cannot capture intra-individual variation, it ensures fairness in inter-group comparisons and aligns with the mainstream paradigm in current research on virtual streamer sentiment computing. Virtual anchors use 2D or 3D real-time driven models with stable facial expressions and body movements; real-life anchors are natural on-screen anchors without virtual avatars. The raw data is manually reviewed to remove streams containing more than 20% of muted, black screen, non-interactive, or non-Chinese bullet screen content. The final dataset has about 120 total hours of valid live streaming video content split into 3 parts - training set, validation set, and test set - at an 8:1:1 proportion. To prevent "data leakage" between these sets coming from any given anchor, the split between these sets is determined according to the number of anchors. This allows models to be generalized for evaluation and provides a fair basis for comparing between different groups.

B. EVALUATION INDEX SYSTEM

The arousal pattern comparison indicator group evaluates an individual's arousal response along four different axes: the mean arousal score (MAS) reflects overall arousal; the peak frequency (PF) indicates the number of measurements above the 75th percentile of arousal for a given time; the rise rate (RR) captures the velocity of a response with the average slope of the rising portion of the arousal curve; the high-arousal duration ratio (HADR) reveals how frequently an individual has high arousal relative to the duration at a low level of arousal. Each of these indices is determined using all available tests for each individual in the group, and then the average for the whole group based on the means of their respective test results can be computed. MAS, PF, RR, and HADR characterize emotional arousal patterns from four orthogonal dimensions—intensity, explosiveness, response speed, and persistence—avoiding the one-sidedness of a single indicator. This system not only meets the needs of model validation but also provides an interpretable and reproducible quantitative basis for inter-group difference analysis between virtual and real anchors.

C. BASELINE METHODS AND ABLATION EXPERIMENT SETUP

Baseline methods select representative models in the current multimodal sentiment analysis field to verify the performance of the proposed method:

1) Single-modal baseline BERT-only, using only a finely tuned BERT-large to model the bullet screen text and regress the arousal value.

2) Early Fusion method, which simply concatenates text (BERT embedding), vision (motion energy), and audio (prosodic features) and inputs them into a multilayer perceptron [29].

3) Tensor Fusion Network (TFN), which explicitly models high-order interactions between modalities through outer product [30].

4) Multimodal Transformer (MulT), which uses a cross-modal attention mechanism to achieve information exchange between modalities [31].

All baselines are trained under the same data partitioning and input alignment conditions. The ablation experiments revolve around the core design of this paper:

1) TACAM-w/o-CAM removes the cross-modal alignment module and adopts fixed-weight fusion of three modalities;

2) TACAM-FixedW replaces the dynamic attention weights with preset mean values;

3) TACAM-w/o-Trans replaces the Transformer temporal regression network with a single-layer LSTM. These settings collectively verify the necessity of text anchoring mechanisms, dynamic alignment strategies, and temporal modeling structures for emotional arousal modeling in virtual anchor scenarios.

IV. SYSTEMATIC DIFFERENCES IN EMOTIONAL AROUSAL PATTERNS BETWEEN VIRTUAL AND REAL ANCHORS

A. OVERALL COMPARISON OF AROUSAL INTENSITY LEVELS

To quantify the overall differences in audience emotional activation levels between virtual and real anchors, this section calculates the average arousal level for each live stream and compares the distribution between groups, revealing the systematic differentiation between the two types of anchors at the baseline of emotional intensity. The analysis is based on 100 paired live streams, and the individual MAS index is calculated. The results are shown in Figure 2.

Figure 2 compares the average arousal level for both virtual anchors and real-life anchors as measured through the use of anchors. The left box in this figure is associated with the virtual anchor group (blue), while the right box corresponds with the real-life anchor group (red). Box plot indicates that the MAS (0.61, median 0.60) for the real-life anchor group is greater than the averages of the virtual anchor group (0.52, 0.51). In fact, the interquartile range reflects greater arousal ranges than do the virtual anchors. The differences arise from two aspects; real-life anchors can provide real-life physiological feedback mechanisms and dynamic contextual response capabilities; real-life anchors show a wide range of multi-channel emotional stimuli in their vocal rhythm as well as the use of subtle facial expressions and other forms of non-verbal feedback, and interaction, with others in real-time; while virtual anchors operate under a structured, pre-established driving model, and have no fluctuation in their voice and/or behaviour. All these reasons cause a virtual anchor to have an extremely low level of emotional arousal relative to other types of media. Current technological developments indicate that real-life anchors continue to have an overall higher level of emotional arousal than their virtual counterparts. This conclusion is based on the dataset curated with content-type pairing. Future studies employing stricter control of production variables or experimental designs are needed to further isolate the effect of the anchor's 'virtual vs. real' nature from that of ancillary production styles.

B. DIFFERENCES IN DYNAMIC RHYTHM OF EMOTIONAL AROUSAL

This section uses a method to summarize and classify the overall emotional arousal differences between virtual anchors and real-life anchors. It does this by displaying a standard deviation band overlaid on arousal curves created for each anchor based on 10-second intervals (timesteps). In addition, the analysis shows how frequently various types of anchors cause emotional outbursts, as well as the rate of ascent and how much variation there is within that time period. These results are based upon thirty minutes worth of live broadcast time-series data, and the results of the TACAM model are used to determine the average group response. The results can be found in Figure 3.

Figure 3 shows a 30-minute time frame on the horizontal axis. Emotional arousal for anchors in real-life reaches its highest point (peak) of 0.76 at 1.1 min; the peak intervals fluctuate with very short intervals and have an overall increase rate RR of approximately 0.31/min (per minute). On the other hand, the virtual anchor's initial peak lags to 1.5 minutes and is less than 0.66. The increase rate for virtual anchors is less than 0.17/min, and it appears to have had relatively smooth fluctuations over the trial period; in contrast, because of the capability of real anchors to respond dynamically to events occurring around them, the sudden change in the rhythm of speech has the ability to produce high-frequency emotional stimuli through subtle changes in facial expressions, as well as their interaction with their audience. Virtual anchors, on the other hand, rely on pre-set action scripts and synthesized speech, lacking natural physiological feedback mechanisms, resulting in delayed and limited arousal signals. The standard deviation band width also reflects stronger intragroup response heterogeneity in real-life anchors, indicating higher individual adaptability in real interpersonal interactions. The results show that current virtual anchors still lag significantly behind real-life anchors in the emotional rhythm dimension, making it difficult to achieve high-density, rapid-response emotional activation.

C. PERSISTENCE OF HIGH AROUSAL

To evaluate the differences in the ability of different sentiment analysis methods to characterize the persistence of high arousal, this section calculates the proportion of emotionally high-arousal time for each model on virtual and real-person anchor test sets, i.e., the proportion of time viewers are in a state of high emotional activation, thereby revealing the effectiveness of the methods in modeling the depth of emotional immersion for the two types of anchors. The analysis covers the TACAM proposed in this paper and four baseline methods, and the results are shown in Figure 4.

Figure 4 presents the percentage of high arousal duration for five methods in both virtual and real anchor scenarios. The horizontal axis represents the method category, and the vertical axis represents the mean HADR within each group. Figure 4(a) corresponds to virtual anchors, and Figure 4(b) corresponds to real anchors. Error bars indicate the 95% confidence interval. The results show that TACAM achieves an HADR of 0.38 in the virtual anchor group, significantly higher than methods such as TFN (0.30) and MulT (0.31). This is because the latter relies on visual-motor energy and speech prosodic features, while the stylized expression of virtual anchors results in a lack of such modal information. The real anchor group has a TACAM score of 0.52, which outperforms all of the baselines, indicating that TACAM has found a way to intelligently combine and utilize both visual (audio/visual) and tactile (touch/feeling). The overall HADR is lower for the virtual anchor groups compared with their counterparts in the real anchor group due to the differences in the levels of natural micro-expressions and tone variation as well as real-time interactive feedback. The virtual anchor models have been limited by the requirement of following a pre-defined driving model so that they exhibit limited

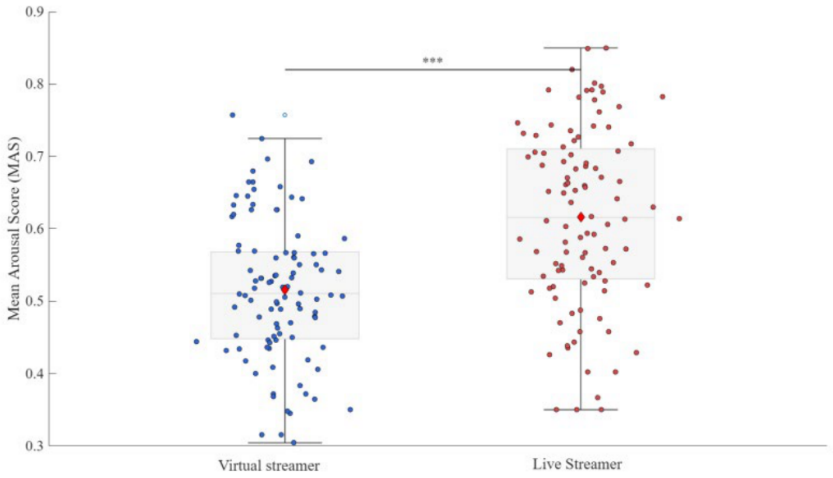

FIGURE 2. Comparison of average arousal distribution between virtual and real-life anchors. Note: "****" indicates $p < 0.001$, representing the statistical significance level of the result

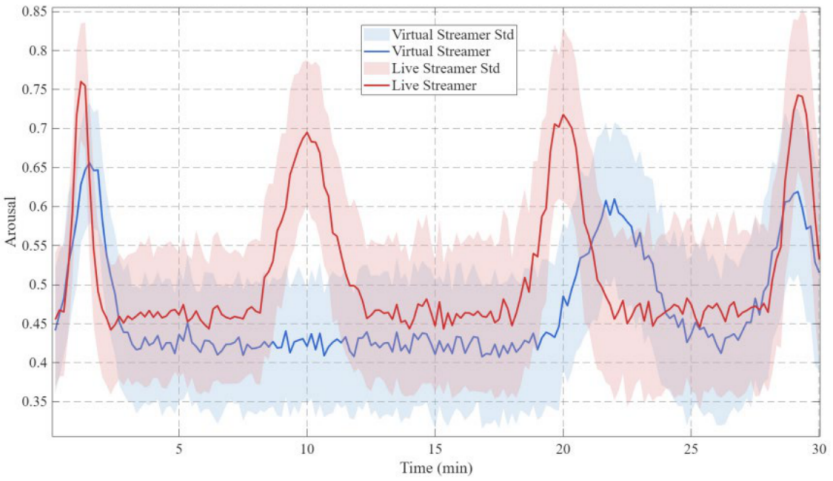

FIGURE 3. Comparison of emotional arousal dynamics curves between virtual and real-life anchors

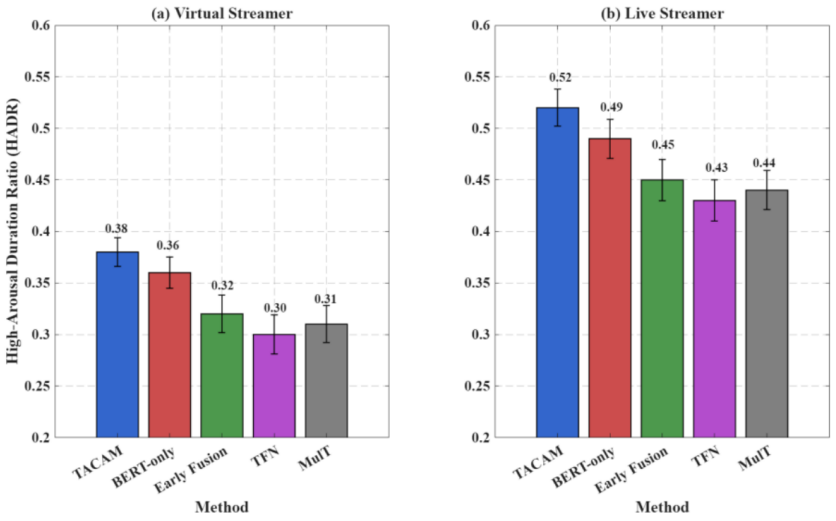

FIGURE 4. Comparison of high arousal duration percentage for different sentiment analysis methods: (a) Virtual anchor scenario (b) Real anchor scenario

emotional expression dimensions, making it more difficult to sustain high levels of audience goodwill and attention

relative to the real anchor groups. The findings illustrate that most multi-layer models used for sentiment analysis have some degree of modeling bias in virtual information, while the text anchoring strategy (an alignment strategy dominated by text semantics) proposed in this paper can effectively bridge this gap.

D. DIFFERENCES IN MULTIMODAL CONTRIBUTION STRUCTURE

To reveal how different models allocate multimodal information weights in virtual and real anchor scenarios, this section calculates the average contribution ratio and modal contribution entropy of each method across the text, visual, and audio modalities based on the dynamic attention weights output by the cross-modal alignment module. It quantifies the model's adaptive ability to modal dependency structures and analyzes the TACAM main model and its three ablation variants. The results are shown in Figure 5.

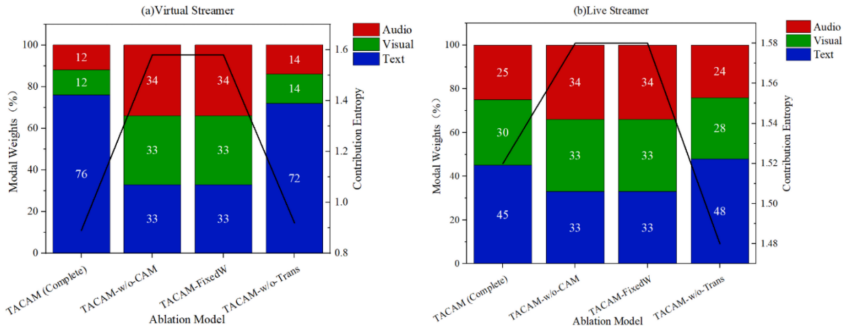

FIGURE 5. Comparison of multimodal contribution structure under different model configurations: (a) Virtual anchor scenario (b) Real anchor scenario

Figure 5(a) shows the modal contribution structure of four models in the virtual anchor scenario, and Figure 5(b) corresponds to the real anchor scenario. The horizontal axis represents the model configuration, including the complete TACAM master model and its three ablation variants; the left and right vertical axes represent the average contribution percentage weight and modal contribution entropy of each modality in emotional arousal modeling. The data results show that in the virtual anchor group, the complete TACAM has a text weight as high as 76%; visual and audio each account for 12%; the contribution entropy is only 0.89. This high text weight (76%) reflects the model's adaptation to the actual distribution of emotional information in the data: in virtual anchor scenarios, the stylized and less variable audiovisual expressions carry limited discriminative power for arousal, while audience reactions are predominantly channeled through text interactions. Thus, the model's reliance on text does not ignore anchor influence but rather captures the shifted pathway of emotional influence—where the audience's textual response becomes the primary carrier of arousal dynamics, as opposed to the direct, multimodal expressiveness of real anchors. The low weight assigned to visual features (12%) in virtual scenarios directly reflects the model's assessment of their low emotional discriminability—a consequence of the artificial, often repetitive nature of avatar motion which reduces the signal-to-noise ratio of raw optical flow as an arousal indicator. While in the real anchor group, the three modal weight distribution is 45% text, 30% visual, and 25% audio, and the entropy value rises to 1.52, reflecting the complementary value of multi-channel information in real interaction. In contrast, the ablation models TACAM-w/o-CAM and TACAM-FixedW, which remove the cross-modal alignment module and force equal weighting of the three modalities (33%, 33%, 34%), both have an entropy value of 1.58. They cannot adjust the weights according to content characteristics, leading to an over-reliance on low signal-to-noise ratio audiovisual features in virtual scenes. The reason for this variation is that virtual anchors do not have true physiological feedback systems. As a result, the speech and actions of virtual anchor do not contain emotions and therefore the model is forced to use text anchors in place of them. The data demonstrates that fixed fusion and multimodal splicing alone do not accommodate the different patterns of emotional expressions contained in the different types of anchors. The cross-modal alignment method in this study allows for dynamic adjustments to modal contribution for improving the robustness of emotional modeling.

V. CONCLUSIONS

This paper proposes TACAM, a cross-modal emotional arousal modeling method that utilizes text semantics as a stable anchor to integrate diverse emotional data into a cohesive multimodal representation. By ensuring precise cross-modal alignment, TACAM ensures that multimodal information remains balanced and tightly bound within the narrative weave. It extracts the arousal semantics of viewer comments by fine-tuning BERT, combines optical flow and speech prosodic features, and designs a cross-modal alignment module to dynamically calibrate the audiovisual

modal contribution using text as an anchor. Finally, it uses Transformer temporal regression to generate high-resolution emotional arousal curves. Experiments show that TACAM achieves a high arousal duration ratio of 0.38 in virtual anchor scenarios, significantly o utperforming t he T FN and MulT multimodal methods. In terms of overall arousal intensity, the average arousal level of real anchors reaches 0.61, higher than the 0.52 of virtual anchors. Modal analysis further shows that TACAM automatically increases the text weight to 76% in virtual anchors, effectively mitigating the modeling bias caused by the limited audiovisual expression. This method provides a robust and quantifiable computational framework for emotional interaction, offering clear application value in virtual idol optimization and the structural reinforcement of multimodal emotional computation.

AUTHOR CONTRIBUTIONS

Conceptualization –Di Zhang, Shiqi Li and Shuhan Xu; methodology – Di Zhang, Shiqi Li and Shuhan Xu; investigation – Di Zhang, Shiqi Li and Shuhan Xu; writing-original draft preparation – Di Zhang, Shiqi Li and Shuhan Xu. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was supported by, This article is the research result of the Basic Research Funds Project for Undergraduate Universities in Heilongjiang Province in 2025, titled "Comparative Study on Audience Cognitive Differences and Emotional Resonance between Virtual Anchors and Real An-chors". Project Number: 2025-KYYWF-RWSK-414

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] J. Ham, S. Li, J. Looi, and M. S. Eastin, "Virtual humans as social actors: Investigating user perceptions of virtual humans' emotional expression on social media," Computers in Human Behavior, vol. 155, Art. no. 108161, 2024, doi: 10.1016/j.chb.2024.108161.

[2] X. Liu and L. Zhang, "Virtual streamer responsiveness and consumer purchase intention in live streaming commerce: Social presence as a mediator," Social Behavior and Personality: An international journal, vol. 52, no. 11, pp. 1-7, 2024, doi: 10.2224/sbp.13708.

[3] Y. Yuan, S. Duo, X. Tong, and Y. Wang, "A multimodal affective interaction architecture integrating BERT-based semantic understanding and VITS-based emotional speech synthesis," Algorithms, vol. 18, no. 8, pp. 513, 2025, doi: 10.3390/a18080513.

[4] R. Das and T. D. Singh, "Multimodal sentiment analysis: A survey of methods, trends, and challenges," ACM Computing Surveys, vol. 55, no. 13s, pp. 1-38, 2023, doi: 10.1145/3586075.

[5] R. Li, X. Yang, J. Lou, and J. Zhang, "A temporal-spectral graph convolutional neural network model for EEG emotion recognition within and across subjects," Brain Informatics, vol. 11, no. 1, pp. 30, 2024, doi: 10.1186/s40708-024-00242-x.

[6] C. Lee, H. Lee, and M. Whang, "Emotion Recognition from rPPG via Physiologically Inspired Temporal Encoding and Attention-Based Curriculum Learning," Sensors, vol. 25, no. 13, pp. 3995, 2025, doi: 10.3390/s25133995.

[7] Z. Zhang, W. Wu, T. Yuan, and G. Feng, "Modality-Enhanced Multimodal Integrated Fusion Attention Model for Sentiment Analysis," Applied Sciences, vol. 15, no. 19, Art. no. 10825, 2025, doi: 10.3390/app151910825.

[8] Z. Li, P. Liu, Y. Pan, J. Yu, W. Liu, H. Chen, et al., "Text-dominant multimodal perception network for sentiment analysis based on cross-modal semantic enhancements," Applied Intelligence, vol. 55, pp. 188, 2025, doi: 10.1007/s10489-024-06150-1.

[9] Y. Song, L. Zhang, R. Zhang, H. Zhan, M. Dai, X. Hu, et al., "MSF-Net: A Data-Driven Multimodal Transformer for Intelligent Behavior Recognition and Financial Risk Reasoning in Virtual Live-Streaming," Electronics, vol. 14, no. 23, pp. 4769, 2025, doi: 10.3390/electronics14234769.

[10] R. Abbas, B. W. Schuller, X. Li, C. Lin, and X. Li, "Emotion recognition in live broadcasting: A multimodal deep learning framework," Multimedia Systems, vol. 31, pp. 253, 2025, doi: 10.1007/s00530-025-01780-y.

[11] U. K. Das, R. S. Ani, N. Datta, I. Fahad, J. Sikder, U. Sara, et al., "Enhancing sentiment analysis accuracy on social media comments using a tuned BERT model," Discover Computing, vol. 28, no. 1, pp. 198, 2025, doi: 10.1007/s10791-025-09599-x.

[12] H. Peng, X. Gu, J. Li, Z. Wang, and H. Xu, "Text-centric multimodal contrastive learning for sentiment analysis," Electronics, vol. 13, no. 6, pp. 1149, 2024, doi: 10.3390/electronics13061149.

[13] S. Bai, L. Cao, and J. Zhou, "Is the Anthropomorphic Virtual Anchor Its Optimal Form? An Exploration of the Impact of Virtual Anchors' Appearance on Consumers' Emotions and Purchase Intention," Journal of Theoretical and Applied Electronic Commerce Research, vol. 20, no. 2, pp. 110, 2025, doi: 10.3390/jtaer20020110.

[14] D. Wang, Y. Peng, L. Haddouk, N. Vayatis, and P. P. Vidal, "Assessing virtual reality presence through physiological measures: A comprehensive review," Frontiers in Virtual Reality, vol. 6, Art. no. 1530770, 2025, doi: 10.3389/frvir.2025.1530770.

[15] O. Galal, A. H. Abdel-Gawad, and M. Farouk, "Rethinking of BERT sentence embedding for text classification," Neural Computing and Applications, vol. 36, pp. 20245-20258, 2024, doi: 10.1007/s00521-024-10212-3.

[16] A. S. Talaat, "Sentiment analysis classification system using hybrid BERT models," Journal of Big Data, vol. 10, no. 1, pp. 110, 2023, doi: 10.1186/s40537-023-00781-w.

[17] X. Wang, H. Ren, and A. Wang, "Smish: A novel activation function for deep learning methods," Electronics, vol. 11, no. 4, pp. 540, 2022, doi: 10.3390/electronics11040540.

[18] Y. R. Youn and J. K. Hong, "Enhancing convolutional neural network performance through optimized sigmoid activation function modeling," Asia-Pacific Journal of Convergent Research Interchange, vol. 9, no. 10, pp. 51-59, 2023, doi: 10.47116/apjcri.2023.10.05.

[19] A. C. Cob-Parro, C. Losada-Gutierrez, and M. Marron-Romera, "Stampede detector based on deep learning models using dense optical flow," Engineering Applications of Artificial Intelligence, vol. 142, Art. no. 109940, 2025, doi: 10.1016/j.engappai.2024.109940.

[20] N. Ranjan, S. Bhandari, Y. Kim, and H. Kim, "Video Frame Prediction by Joint Optimization of Direct Frame Synthesis and Optical-Flow Estimation," Computers, Materials and Continua, vol. 75, no. 2, pp. 2615-2639, 2023, doi: 10.32604/cmc.2023.026086.

[21] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM," Multimedia Tools and Applications, vol. 80, no. 9, pp. 13059-13076, 2021, doi: 10.1007/s11042-020-10285-x.

[22] X. Zhuang, F. Liu, J. Hou, J. Hao, and X. Cai, "Transformer-based interactive multi-modal attention network for video sentiment detection," Neural Processing Letters, vol. 54, pp. 1943-1960, 2022, doi: 10.1007/s11063-021-10713-5.

[23] C. Shi and Y. Zhang, "MMKT: Multimodal Sentiment Analysis Model Based on Knowledge-Enhanced and Text-Guided Learning," Applied Sciences, vol. 15, no. 17, pp. 9815, 2025, doi: 10.3390/app15179815.

[24] J. Hou, N. Omar, S. Tiun, S. Saad, and Q. He, "TCHFN: Multimodal sentiment analysis based on text-centric hierarchical fusion network," Knowledge-Based Systems, vol. 300, no. 1, Art. no. 112220, 2024, doi: 10.1016/j.knosys.2024.112220.

[25] C. Liu, Y. Wang, and J. Yang, "A transformer-encoder-based multimodal multi-attention fusion network for sentiment analysis," Applied Intelli-

gence, vol. 54, no. 17-18, pp. 8415-8441, 2024, doi: 10.1007/s10489-024-05623-7.

[26] N. M. Foumani, C. W. Tan, G. I. Webb, and M. Salehi, "Improving position encoding of transformers for multivariate time series classification," Data Mining and Knowledge Discovery, vol. 38, no. 1, pp. 22-48, 2024, doi: 10.1007/s10618-023-00948-2.

[27] Y. Dogan, "A new global pooling method for deep neural networks: Global average of top-k max-pooling," Traitement du Signal, vol. 40, no. 2, pp. 577-587, 2023, doi: 10.18280/ts.400216.

[28] A. Zafar, M. Aamir, N. Mohd Nawi, A. Arshad, S. Riaz, A. Alruban, et al., "A comparison of pooling methods for convolutional neural networks," Applied Sciences, vol. 12, no. 17, pp. 8643, 2022, doi: 10.3390/app12178643.

[29] K. Zhao, M. Zheng, Q. Li, and J. Liu, "Multimodal Sentiment Analysis—A Comprehensive Survey From a Fusion Methods Perspective," IEEE Access, vol. 13, no. 1, pp. 64556-64583, 2025, doi: 10.1109/ACCESS.2025.3554665.

[30] X. Yan, H. Xue, S. Jiang, and Z. Liu, "Multimodal sentiment analysis using multi-tensor fusion network with crossmodal modeling," Applied Artificial Intelligence, vol. 36, no. 1, Art. no. 2000688, 2022, doi: 10.1080/08839514.2021.2000688.

[31] D. Wang, S. Liu, Q. Wang, Y. Tian, L. He, and X. Gao, "Cross-modal enhancement network for multimodal sentiment analysis," IEEE Transactions on Multimedia, vol. 25, pp. 4909-4921, 2023, doi: 10.1109/TMM.2022.3183830.