A Knowledge-Enhanced Cross-Modal Transformer Network for Sentiment Analysis in Intelligent Interaction

Chunyan Huang, Xinlu Sun*
China Zhejiang Business College, Hangzhou 310053, China

***Abstract*—With the rapid advancement of multimodal emotion recognition technology, sentiment analysis models that integrate heterogeneous information—such as facial expressions and vocal intonation—are driving human–computer interaction and affective computing toward multidimensional, objective, and highly accurate approaches. Conventional emotion recognition methods typically rely on a single-modal input and therefore struggle to capture complex semantic associations and deep emotional features, which in turn undermines the stability of recognition results. A knowledge-enhanced cross-modal Transformer network (KCTN) model was proposed for sentiment analysis, which incorporates a multimodal fusion module and a long-range affective integration module to achieve deep collaborative modeling across text, speech, and facial expression features. This framework substantially enhances the completeness and robustness of emotional semantic representations. Experimental results on the self-built EC-SFED multimodal dataset and the publicly available dataset CMU-MOSI demonstrate that KCTN surpasses several mainstream baseline models in both accuracy and macro-averaged F1 score, validating its superior performance in intelligent interaction and affective computing applications.**

Keywords*—Multimodal sentiment analysis; transformer network; emotion recognition; psychological assessment; intelligent interaction

I. Introduction

Recent research on sentiment analysis within university student populations has primarily evolved along two methodological trajectories: traditional machine learning–based approaches and deep learning–driven models. Traditional methods generally require complex feature engineering on social or textual data. Wang and Zhang [1] designed a topic–sentiment hybrid model to identify students' emotional attitudes toward different discussion themes. Iram [2] proposed an automated framework for detecting affective tendencies in student-generated posts on the Facebook platform. Zhang et al. [3] adopted a sentiment analysis approach based on the Robustly Optimized BERT Pretraining Approach-Whole Word Masking (RoBERTa-WWM) capable of capturing both local semantic features and contextual associations within student discourse.

Deep learning techniques enable the automatic extraction of high-level semantic features and complex associations from multimodal data, and the complementary information across modalities enhances the effectiveness of psychological assessment. In previous studies, Song et al. [4] developed a Multi-level Attention and Multi-task (MAM) fusion model that incorporates an attention mechanism to support multitask multimodal sentiment analysis. Sun et al. [5] designed an efficient multimodal Transformer architecture featuring a dual-level feature restoration mechanism, in which global–local interactive modeling is utilized to improve multimodal fusion performance. Huan et al. [6] proposed a unified multimodal framework composed of a translation module and a prediction module, achieving structured cross-modal integration and semantic alignment. Tsai et al. [7] introduced the Multimodal Transformer (MulT) model, which employs cross-modal attention to process unaligned multimodal data and demonstrated strong performance in comparative experiments. Hazarika et al. [8] designed the Modality-Invariant and -Specific Representations (MISA) framework, which maps each modality into two distinct subspaces to enhance fusion outcomes. Han et al. [9] developed the MultiModal InfoMax (MMIM) framework, where intermodal information is maximized and jointly optimized with the primary task to strengthen modeling performance. Huddar et al. [10] proposed an attention-based multimodal contextual fusion approach that achieved notable improvements in emotion classification. Luo and Zhu [11] introduced the Multi-Dynamic Aware Network (MultiDAN), in which multilayer multi-angle interaction perception is employed to enhance multimodal language sentiment analysis. Yang et al. [12] constructed the CCIN-SA model based on a composite cross-modal interaction network to address challenges associated with multimodal fusion and interactivity, thereby improving classification accuracy. Zhong et al. [13] proposed the KET model, which dynamically integrated commonsense information by computing the cosine similarity and emotional intensity between text and commonsense knowledge features. In addition, the effectiveness of multimodal deep learning for extracting semantic and affect-related features has also been demonstrated in recent applied studies [14, 15].

In multimodal sentiment analysis for psychological assessment, various fusion strategies and deep learning models have been proposed, resulting in notable improvements in emotion recognition accuracy and robustness. However, most evaluation frameworks have primarily emphasized outcome-oriented measures while overlooking behavioral data generated during the assessment process. Intelligent interaction systems, driven by artificial intelligence technologies, create emotional experiences and produce substantial amounts of structurally

*Corresponding author.

complex behavioral process data through visual and speech modalities. These data capture the decision-making trajectory of participants and document interaction outcomes at critical points, thereby addressing the limitations of traditional psychological assessments that rely solely on textual information. This provides valuable insights into individual psychological and behavioral patterns. In this study, a multimodal fusion module and a long-range affective fusion module were integrated. The multimodal fusion module leverages heterogeneous sources, including text, speech, and video. The long-range affective fusion module aggregates affective context from subsequent utterances of the same speaker into the representation of the current utterance, effectively capturing long-distance semantic dependencies. The optimization of deep learning models to strengthen affective computing accuracy constitutes a critical pathway for developing intelligent auxiliary systems for student mental health monitoring.

II. Model Design

A. Task Definition

In the affective computing task for conversational scenarios, a set of multimodal samples is defined as Y. Each dialogue sample contains N successive utterances denoted as $U=\{u_1,u_2,...u_n\}$, and is associated with M speakers represented as $E=\{e_1,e_2,...e_m\}$. For each utterance u_i, multimodal information is composed of three modalities—text (T), audio (A), and visual (V)—such that $u_i=\{u_i^c\}$ and $c\in\{A,V,T\}$. Based on predefined emotion categories, sentiment analysis aims to infer the emotional polarity $p\in$ {negative, neutral, positive} for each speaker by jointly modeling speaker-related characteristics, contextual dependencies, and multimodal cues. Through this process, the emotion label of a new conversational sample can be accurately predicted.

B. Model Architecture

The proposed KCTN model (Fig. 1) for psychological assessment–oriented sentiment analysis incorporates the following three primary methodological contributions:

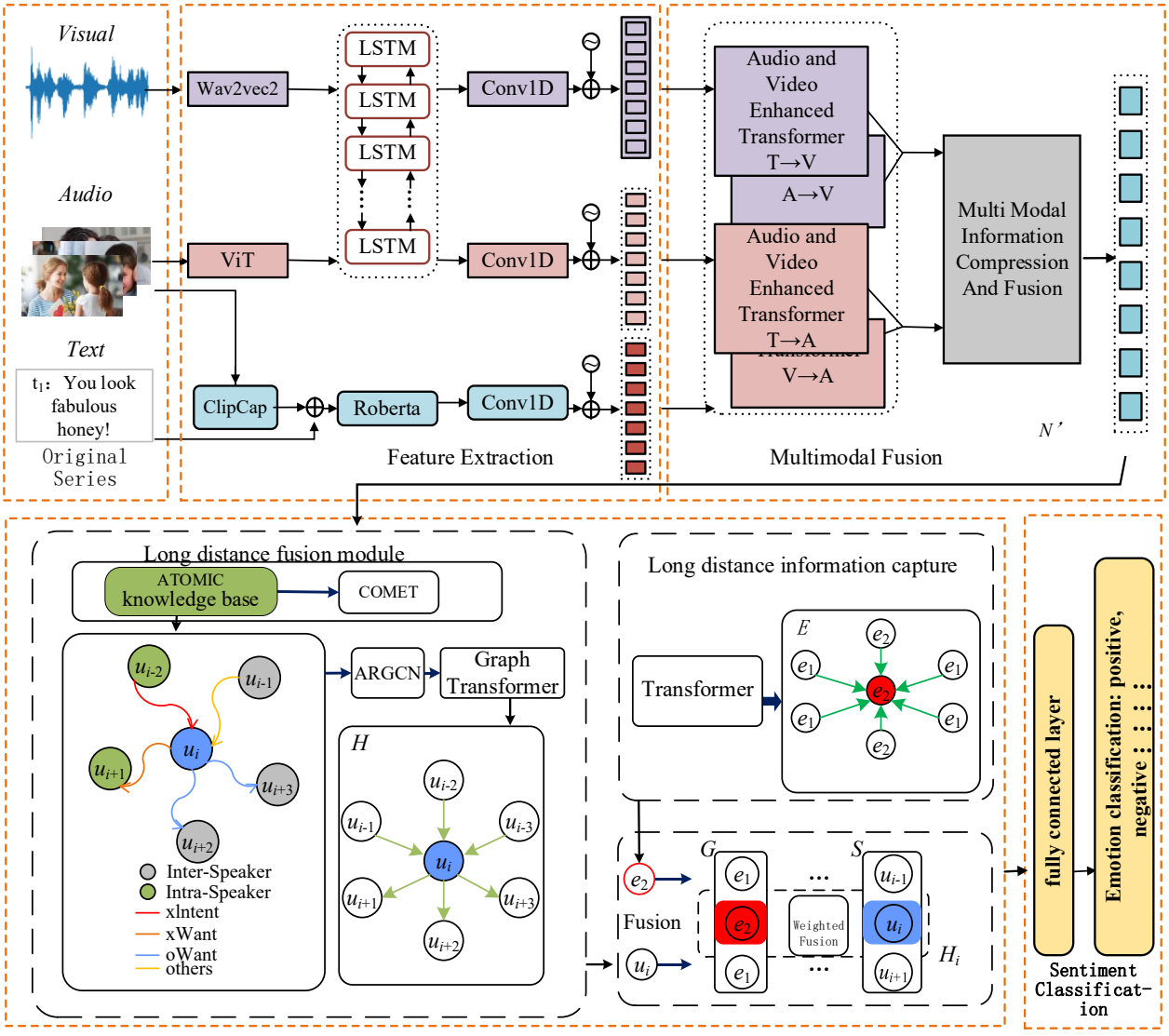

Fig. 1. Overall architecture of the proposed model.

1) In terms of feature extraction, extraction strategies for text, visual, and audio modalities were deeply studied. To address missing aspect-level emotional semantics in text, ClipCap was employed to generate natural language descriptions that align closely with image content, after which textual features were derived using the RoBERTa model. Visual features were extracted using OpenFace 2.0 and a Vision Transformer (ViT) model, while audio features were captured through Wav2Vec 2.0 and OpenSMILE.

2) In the multimodal fusion module, a cross-modal enhanced Transformer fusion network was introduced. Long Short-Term Memory (LSTM) networks, temporal convolutions, and positional embeddings were incorporated to effectively capture temporal dynamics and contextual dependencies within audio–visual modalities. Through an audio–visual enhanced Transformer module and an information compression–fusion module, non-text modalities were strengthened and deep intermodal interactions were achieved.

3) In the long-range affective fusion module, a multi-stage processing architecture was adopted. External knowledge was extracted from textual content using Commonsense Transformers (COMET) [16], and the dialogue structure was modeled through a Graph Neural Network (GNN) that transforms conversational data into graph representations to obtain sentence-level semantic features. Subsequently, the self-attention mechanism of the Transformer architecture was used to capture long-range affective dependencies across speaker utterances. Finally, a feature fusion mechanism was used to integrate sentence-level semantic representations with speaker-level affective features, thereby enhancing the model's affective computing capability.

III. Feature Extraction and Fusion

A. Text Feature Encoding

In the text feature encoding module, features from both text and images are integrated, using visual information to compensate for potentially missing affective semantics in textual data. Inspired by the multimodal processing framework proposed by Khan and Fu [17], the ClipCap image description generator was incorporated. Built upon the advanced Contrastive Language-Image Pretraining (CLIP) cross-modal pretraining framework, ClipCap integrates visual features with the prior knowledge of a language model to generate semantically precise and affectively expressive descriptions. These descriptions preserve the global emotional characteristics of the image and enable effective transfer of visual information into textual semantics.

The original text E and the image description text C were subsequently combined into a unified token sequence X through structured concatenation with special markers. A [SEP] separator was inserted between the two text segments to explicitly define modality boundaries, and a [CLS] token was placed at the beginning of the sequence to aggregate global semantic information. The resulting sequence X can be expressed below. This structured formulation preserves the fine-grained characteristics of the original text while incorporating additional semantic information generated from visual content.

$$X=[CLS]t_{E_1}t_{E_2}\ldots t_{E_{N_E}}[SEP]t_{C_1}t_{C_2}\ldots t_{C_{N_C}}[SEP] \quad (1)$$

where, Ei and Ej denote the tokens of sentences E and C, respectively. The number of tokens in E is represented as N_E, while the number of tokens in C is represented as N_C.

To fully capture the deep semantic features embedded in the textual sequence, RoBERTa-base was employed as the encoder. In the experimental configuration, the encoder was set to a 12-layer architecture with a hidden dimension of 768, and the maximum sequence length was fixed at 64 tokens. Inputs shorter than this threshold were padded with the [PAD] token, whereas overlength sequences were processed through intelligent truncation. For a given sample from the dataset, the encoded representation was obtained as follows:

$$I_t=RoBERTa\left(X;\theta^R\right)\in R^{l_t\times d_t} \quad (2)$$

where, l_t denotes the length of the text sequence, d_t denotes the dimensionality of the textual representation, and θ^R represents the parameters of the RoBERTa model.

B. Audio Feature Encoding

The multimodal emotional information generated during intelligent interaction requires the preservation of temporal continuity and contextual dependencies. However, the raw audio and visual features extracted typically consist of independent frame-wise representations that lack sufficient contextual information. To address this limitation, an LSTM network was employed to encode these modalities. In multimodal sentiment analysis for intelligent interaction, audio features are commonly categorized into three types: spectral features, timbre-related features, and prosodic features. In this study, the wav2vec2-base-960h pretrained model was utilized for feature extraction. This model was pretrained and fine-tuned on the 960-hour LibriSpeech dataset. All audio data were uniformly resampled to a 16 kHz sampling rate. New samples underwent preprocessing, feature encoding, masking, contextual representation, feature quantization, contrastive learning, loss calculation, fine-tuning, and feature extraction, resulting in the acquisition of 768-dimensional audio representations. For a given audio sequence $A=\{a_1,a_2,\ldots a_n\}$, the extracted audio feature representation is defined as:

$$I_a=Wav2vec2.0(A)\in R^{l_a\times d_a} \quad (3)$$

where, l_a denotes the audio sequence length, and d_a denotes the audio sequence dimensionality.

During the audio feature encoding process, OpenSMILE [18] with the IS10 configuration was employed to extract speech segments. Three predefined standard feature sets were included in this configuration: ComParE 2016 (a large-scale paralinguistic feature set), GeMAPS (a basic acoustic parameter set), and its extended version, eGeMAPS. The detailed configuration of these feature sets is presented in Table I.

TABLE I. OVERVIEW OF OPENSMILE FEATURE SETS

Feature Set	Number of Features	Feature Type
ComParE 2016	6373	Static features
GeMAPS	62	High-level statistical function (HSF) features
eGeMAPS	88	Extended features + HSF features

For affective computing tasks, the ComParE 2016 configuration was adopted to extract 6,373 acoustic features from each speech segment. A fully connected network was then used to reduce these features to 1,582 dimensions for the IEMOCAP dataset and to 300 dimensions for the MELD dataset; the resulting unified representation is denoted as u_t^A. To enhance temporal modeling capability, a bidirectional LSTM architecture was employed to capture contextual dependencies within the audio signal. The computation process is expressed as follows:

$$\mathbf{t}_i^{\mathrm{A}},\mathbf{H}_i^{\mathrm{A}}=\left[\overrightarrow{\mathrm{LSTM}}\left(\mathbf{u}_i^{\mathrm{A}},\overrightarrow{\mathbf{H}_{i-1}^{\mathrm{A}}}\right),\overleftarrow{\mathrm{LSTM}}\left(\mathbf{u}_i^{\mathrm{A}},\overleftarrow{\mathbf{H}_{i+1}^{\mathrm{A}}}\right)\right] \quad (4)$$

where, $\boldsymbol{H}_{i-1}^{A}$ represents the (i-1)-th hidden layer state, $\boldsymbol{H}_{i+1}^{A}$ represents the (i+1)-th hidden layer state, and $\boldsymbol{t}_i^{A}$ denotes the context-aware audio encoding.

C. Visual Feature Encoding

During the intelligent interaction process in psychological assessment, facial micro-expression dynamics were captured through video data. Visual feature extraction was performed using the OpenFace 2.0 toolkit and the ViT model. OpenFace 2.0 extracts video features through four stages: frame decomposition, facial detection and landmark localization, facial alignment, and facial feature extraction. The ViT model, built upon the Transformer architecture for image processing, divides each input frame into fixed-size patches and encodes them using a Transformer-based pipeline. The feature extraction process includes video preprocessing, selection of effective frames, image segmentation into patches, patch embedding, ViT-based encoding, and feature extraction. Through multiple layers of Transformer encoders, each token is dynamically updated to feature vectors integrating global contextual semantics with local visual details [19]. A compact representation of each video frame was then obtained by aggregating the encoded tokens. Given a preprocessed video sequence $V=\{v_1,v_2,...v_n\}$, the extracted visual feature representation is defined as:

$$I_v=ViT(V)\in\mathrm{R}^{l_v\times d_v} \quad (5)$$

where, l_v denotes the length of the video sequence, and d_v denotes the dimensionality of the video sequence.

For the video-based emotion extraction task, the raw visual features were initially obtained as independent frame-level representations. Since multimodal sentiment analysis requires the preservation of temporal continuity and contextual dependencies, a 342-dimensional visual feature, denoted as u_t^V, was first extracted. Similar to the audio encoding process, a bidirectional LSTM network was employed to capture contextual cues. The computation is defined as follows:

$$\mathbf{t}_i^{\mathrm{V}},\mathbf{H}_i^{\mathrm{V}}=\left[\overrightarrow{\mathrm{LSTM}}\left(\mathbf{u}_i^{\mathrm{V}},\overrightarrow{\mathbf{H}_{i-1}^{\mathrm{V}}}\right),\overleftarrow{\mathrm{LSTM}}\left(\mathbf{u}_i^{\mathrm{V}},\overleftarrow{\mathbf{H}_{i+1}^{\mathrm{V}}}\right)\right] \quad (6)$$

where, $\boldsymbol{H}_{i-1}^{V}$ and $\boldsymbol{H}_{i+1}^{V}$ represent the (i-1)-th and (i+1)-th hidden layer states, respectively, and $\boldsymbol{t}_i^{V}$ denotes the context-aware video encoding. To ensure that elements within each input sequence adequately capture interactions with adjacent elements, a one-dimensional temporal convolution layer was applied, which transforms the states into a unified dimension to facilitate subsequent processing. The transformation is computed as:

$$\widehat{F}_t=\mathrm{Conv1D}(I_t,\mathrm{kernel}_t)\in\mathrm{R}^{l_t\times d} \quad (7)$$

$$\widehat{F}_{\{a,v\}}=\mathrm{Conv1D}\left(I_{\{a,v\}},\mathrm{kernel}_{\{a,v\}}\right)\in\mathrm{R}^{l_{\{a,v\}}\times d} \quad (8)$$

where, $\mathrm{Kernel}_{\{t,a,v\}}$ denotes the convolution kernels corresponding to the text, audio, and visual modalities, and d represents the modality dimension.

D. Multimodal Fusion

For management units, two remote operation and maintenance modes are provided.

1) Multi-Modal Enhanced Transformer (CMET) module: In sentiment analysis, the textual modality typically plays a dominant role in conveying explicit semantic information, whereas non-textual modalities often provide auxiliary affective cues. To strengthen the emotional representational capacity of non-textual modalities, a CMET module was introduced. This module establishes a bidirectional cross-modal attention mechanism to facilitate complementary information exchange and collaborative optimization across modalities, thereby significantly enhancing the affective expressiveness of non-textual modalities. In the module, and denote the two non-textual modalities, while t denotes the textual modality. Fig. 2 shows the architecture of the audio–visual enhanced Transformer.$\alpha\beta$

The general formulation of $CMET(Q,K,V)$ is given as:

$$\mathrm{CMET}_{\mathrm{mul}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (9)$$

where, Q denotes the queries, K denotes the keys, V denotes the values, and d_k represents the dimensionality of the key or value vectors. The computation of Q, K, and V in the multi-head attention mechanism is expressed below, with the weight matrices defined as $W_{Q_\alpha}\in\mathrm{R}^{\mathrm{d}_\alpha\times d_k}$, $W_{K_\beta}\in\mathrm{R}^{\mathrm{d}_\beta\times d_k}$, and $W_{V_\beta}\in\mathrm{R}^{\mathrm{d}_\beta\times d}$:

$$Q_\alpha=Z_\alpha W_{Q_\alpha}\in\mathrm{R}^{l_\alpha\times d_k} \quad (10)$$

$$K_\beta=Z_\beta W_{K_\beta}\in\mathrm{R}^{l_\beta\times d_k} \quad (11)$$

$$V_\beta=Z_\alpha W_{V_\beta}\in\mathrm{R}^{l_\alpha\times d} \quad (12)$$

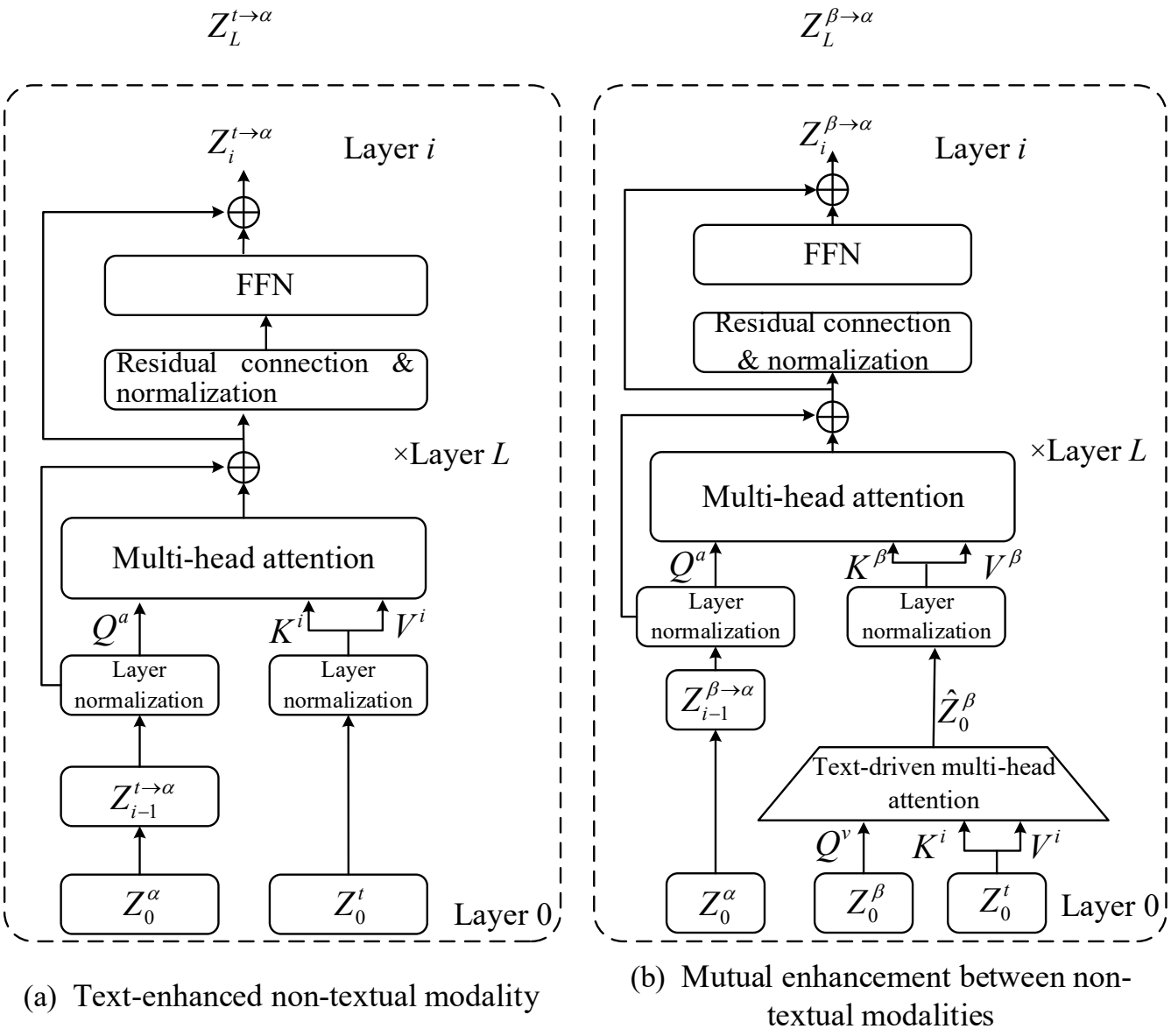

Fig. 2. Architecture of the audio–visual enhanced transformer.

For multimodal psychological testing, a cross-modal interaction encoder was designed that uses the text modality to compensate for underperforming modalities through feature enhancement: *Text→Visual*, *Audio→Visual*, *Text→Audio*, and *Visual→Audio*. This framework fosters deep inter-modal interaction between visual and audio streams while preserving information flow across all modalities. The enhancement of the visual modality is exemplified by the following formula:

$$\tilde{\mathbf{Z}}_i^{T\to V}=\mathrm{CMET}_{(i),\,\mathrm{mul}}^{T\to V}\left(\boldsymbol{t}_{(i\text{-}1)}^{T},\boldsymbol{t}_i^{T},\boldsymbol{t}_i^{V}\right) \tag{13}$$

$$\tilde{\mathbf{Z}}_i^{A\to V}=\mathrm{CMET}_{(i),\,\mathrm{mul}}^{A\to V}\left(\boldsymbol{t}_{(i\text{-}1)}^{A},\boldsymbol{t}_i^{A},\boldsymbol{t}_i^{V}\right) \tag{14}$$

The CMET module stacks multiple layers of cross-modal attention blocks, where the output of each layer serves as the input to the next, progressively enhancing audio and visual modality representations. As an example, for the enhancement of the audio modality via the visual modality (V→A), the audio representation was first initialized as $Z_0^{v\to a}=Z_0^a$, after which textual content was utilized to guide the visual modality in enriching affective information. The enhancement process is expressed as:

$$\hat{Z}_0^v=\mathrm{CMET}_{multi}^{T\to V}=\mathrm{Multihead}\left(Q_v,K_t,V_t,\right) \tag{15}$$

At layer i, the output $Z_{i-1}^{v\to a}$ from the previous layer and the text-enhanced modality information $\hat{Z}_0^v$ are integrated to compute attention weights, capturing inter-modal interactions. This integrated output is normalized to produce $\hat{Z}_i^{v\to a}$, which is then passed through a Feedforward Neural Network (FFN), yielding the final enhanced audio–visual representation $Z_i^{v\to a}$. The calculation is defined by the following formula:

$$\hat{Z}_i^{v\to a}=\mathrm{CMET}_{multi}^{v\to a}\left(\mathrm{LN}(Z_{i-1}^{v\to a}),\mathrm{LN}\left(\hat{Z}_0^v\right)\right)+\mathrm{LN}(Z_{i-1}^{v\to a}) \tag{16}$$

$$Z_i^{v\to a}=\mathrm{FFN}\left(\mathrm{LN}\left(\hat{Z}_i^{v\to a}\right)\right)+\mathrm{LN}\left(\hat{Z}_i^{v\to a}\right) \quad (17)$$

where, LN() denotes layer normalization, and FFN() denotes the feedforward neural network. In standard Transformer architectures, the ReLU activation function is commonly employed. The FFN applies an identical linear transformation and activation function to each element in the sequence independently. The FFN is formulated as:

$$\mathrm{FFN}(X)=\max(0, xW1+b1)W2+b2 \quad (18)$$

where, W_1 and W_2 are weight matrices, b_1 and b_2 are bias terms, and max(0,y) represents the ReLU activation function applied to each element to introduce nonlinearity.

2) Information compression and fusion: Token sequences were utilized as information carriers to enable efficient compression of audio–visual data. The process involves the steps below. First, the audio–visual sequence Z_m^L is concatenated with a dimension-aligned token sequence to obtain Z_{mt}^L . The combined representation incorporates information from both the original audio–visual features and their associated tokens:

$$Z_{mt}^L=\mathrm{Concat}\left(\hat{Z}_m^L,T\right) \quad (19)$$

Subsequently, a Transformer layer is applied to extract and condense the salient information contained within the audio–visual sequence. Drawing inspiration from the ViT architecture, the transformation is expressed as:

$$T_m=\mathrm{Transformer}\left(Z_m^L;\theta_m\right) \quad (20)$$

After information compression, the extracted audio and visual feature vectors are fused into a unified representation that reflects the intrinsic correlation and complementarity between the modalities. The compressed audio feature vector T_a and the compressed visual feature vector T_v are concatenated to form a joint representation T_{av}:

$$T_{av}=\mathrm{Concat}(T_a,T) \quad (21)$$

The fused representation T_{av} is then processed using a Transformer encoder to integrate comprehensive information across audio and visual modalities. The resulting representation is defined as:

$$Z_{av}=\mathrm{Transformer}\left(\hat{T}_{av}\right) \quad (22)$$

The representation Z_{av}, obtained through information compression and fusion, captures comprehensive affective cues from audio–visual data and enhances the model's ability to understand and interpret complex emotional states. After stacking *X* layers of processing, the final textual representation Z_t and the fused audio–visual representation Z_{av} are generated. Because different modality combinations contribute unequally to sentiment analysis, an adaptive weighting module was designed to quantify the contribution of each modality to the final output. The multimodal features are then aggregated through a weighted fusion mechanism to produce the final integrated output. Following a bias-free linear transformation applied to the output vectors Z_t and Z_{av}, a softmax operation is used to compute the output weights. These weights are then applied to each representation to generate the final output $\boldsymbol{N}'$, enabling more effective fusion of multimodal information and improving the accuracy of emotion prediction. The computation is defined as:

$$\boldsymbol{n}_{\mathrm{softmax}}=\mathrm{softmax}([\boldsymbol{W}_1\otimes Z_t,\boldsymbol{W}_2\otimes Z_{av}]) \quad (23)$$

$$\boldsymbol{N}'=\boldsymbol{n}_{\mathrm{softmax}}\odot Z_t+\boldsymbol{n}_{\mathrm{softmax}}\odot Z_{av} \quad (24)$$

E. Long-Range Affective Fusion Module

The long-range affective fusion module employs a hierarchical architecture to integrate emotional information. External knowledge is first extracted using the COMET model, which is pretrained on the ATOMIC knowledge graph. COMET generates nine categories of external knowledge features associated with the input text, including speaker intent (xIntent), speaker need (xNeed), speaker attribute (xAttr), effects on the speaker (xEffect), speaker desire (xWant), speaker emotional reaction (xReact), effects on others (oEffect), others' desire (oWant), and others' emotional reaction (oReact). Next, a graph-structured representation of the dialogue is constructed using a GNN to capture local, sentence-level semantic features. A Transformer encoder is then used to model dependencies and extract affective features across distant utterances. Finally, a feature fusion mechanism is designed to adaptively integrate the local sentence-level representations with global speaker-level emotional information.

Within the long-range fusion module, a directed graph is constructed to represent the interaction flow across dialogue segments. The output from the multimodal fusion module, $\boldsymbol{N}'$, is used as the initial node feature h_i. Knowledge representations generated from COMET, pre-trained on the ATOMIC knowledge graph, are aligned with relation types. The Relational Graph Convolutional Network (RGCN) updates the hidden states of nodes through the neighboring node representations based on edge types [20], while the Attention-based Relational Graph Convolutional Network (ARGCN) extends RGCN by incorporating distance-aware attention mechanisms to further enhance representational capacity [21]. The node calculation process is expressed as:

$$\boldsymbol{h}_i'=\sigma\left(\sum_{r\in R}c_{i,j}\sum_{j\in N_i^r}x_{ij,r}+\boldsymbol{W}_4\boldsymbol{h}_i\right) \quad (25)$$

where, σ denotes the activation function, R represents the set of relation types, $h_i=N'$, N_i^r is the set of neighboring nodes connected to node v_i under relation $r\in R$, W_4 is a learnable parameter matrix, and $C_{i,j}$ denotes the normalization coefficient. To propagate the interaction information aggregated by the ARGCN, a GraphTransformer model consisting of L layers is employed. The updating operation is defined as:

$$\mathbf{h}_i^{(l+1)}=\left(1-\beta_i\right)\left(\sum_{j\in N(i)}\alpha_{i,j}\,m_j\right)+\beta_i\mathbf{W}_6\mathbf{h}_i^{(l)} \quad (26)$$

where, $N(i)$ denotes the set of source nodes connected to node i, m_j represents the message from source nodes, $\alpha_{i,j}$ is the attention score, β_i is the gating parameter for the residual connection, and W_6 denotes the mapping weight. This update rule yields the final representation for all nodes in the dialogue graph. In parallel, a Transformer network is used to capture

self-dependencies across adjacent utterances from the same speaker. The resulting speaker-level contextual representation is denoted as p_i. A function $Y_i=f(L_{past},p_i)$ aggregates information from the preceding n_{past} utterances, where L_{past} denotes the information on n_{past} utterances. For the i-th utterance in the dialogue, its feature representation $\boldsymbol{E}_i$ is computed using ARGCN. In addition, U_λ denotes the information on all utterances in the dialogue.

$$E_i=\text{ARGCN}(\boldsymbol{E}_i,\boldsymbol{E}_j),j\in[i,|U_\lambda|] \tag{27}$$

Finally, a Transformer integrates Y_i with $\boldsymbol{E}_i$. The attention scores are assigned weights, generating the final fused representation H_i:

$$\boldsymbol{H}_i=\text{Transformer}\left(\boldsymbol{Y}_i,E,\sum_{i=1}^{N}\text{sim}(\boldsymbol{Y}_i,\boldsymbol{E}_i)\right) \tag{28}$$

F. Emotion Classification and Model Optimization

During the emotion prediction stage, the composite vector is passed through a fully connected layer for affective classification. The ReLU activation function and the softmax function are used to project the affective features into the emotion category space, producing the classification output:

$$\boldsymbol{H}_i^{'}=\text{ReLU}(\boldsymbol{W}_u\boldsymbol{H}_i+\boldsymbol{b}_1) \tag{29}$$

$$Q_i=\text{Softmax}(\boldsymbol{W}_u\boldsymbol{H}_i^{'}+\boldsymbol{b}_2) \tag{30}$$

where, W_u denotes the weight matrix of the fully connected layer, b_1 and b_2 are bias vectors, and Q_i represents the softmax output. The final predicted emotion category is determined as $\hat{y}_i$. To optimize all parameters in the model, the minimized standard cross-entropy loss function is employed.

$$\hat{y}_i=\text{argmax}(Q_i) \tag{31}$$

$$L(\theta)=-\sum_{i=1}^{N}\sum_{j=1}^{c(i)}\log P_{i,j}\left[y_{i,j}\right] \tag{32}$$

IV. Empirical Results

A. Experimental Environment

The experimental environment was configured using the Ubuntu 22.04 operating system. Python 3.10 was adopted as the programming language, with Anaconda 4.2.0 used for package and environment management. PyTorch 1.12.1 served as the deep learning framework, supported by a 10 GB GPU. Computational efficiency and stability were ensured through the integration of CUDA Toolkit 11.3 and cuDNN v8.3.2 acceleration libraries.

B. Dataset and Evaluation Metrics

1) Self-constructed multimodal dataset: The dataset used in the experimental process was collected from students of the School of E-Commerce at the host institution. The intelligent interaction data were obtained from a representative application case titled AI Heart World, a typical intelligent non-cognitive ability assessment system based on immersive interaction. The system was developed around the core elements of "assessment needs – assessment environment – analytical techniques – information fusion" and integrates sandbox theory, psychometrics, game-based interaction, and artificial intelligence technologies. This interactive environment captures multimodal emotional information generated through scenario-based game interactions, enabling the measurement of an examinee's psychological state, personality traits, and behavioral characteristics in a more naturalistic setting, thus yielding more authentic data [22]. The collected data were processed and organized into a multimodal psychological assessment database (EC-SFED), which provides the foundation for Transformer-based sentiment analysis to effectively represent individual psychological states.

Fig. 3. Selected examples from the EC-SFED dataset.

During the experimental process, audio–visual information was collected from 82 university students using their personal devices. Each sample consisted of text, audio, and video components. Textual data were obtained from students' subjective responses within the intelligent interaction platform, including discussion content, forum exchanges, and instant messaging records. Audio data were recorded through microphones or headset devices during the psychological assessment sessions and captured acoustic characteristics such as intonation, loudness, and speaking rate. Video data were captured in real time via intelligent cameras, recording facial expressions, body posture, gaze dynamics, and other visual cues relevant to emotional analysis. The final dataset contained 720 text entries, 233 dialogues, 408 sentences, and 345 video clips. Each video clip was annotated with an emotion intensity score ranging from –3 to +3 (where –3 indicates highly negative emotion and +3 indicates highly positive emotion). Dialogues involved two or more speakers and were divided into training, validation, and test sets using an 8:1:1 ratio. Fig. 3 provides sample pictures of the participants. For feature extraction, the RoBERTa-base model was employed to generate 1,024-dimensional textual representations. Audio features were extracted using the Waveform-to-Vector 2.0 (Wav2Vec2) pretrained model, yielding 768-dimensional embeddings. Visual features were extracted using the ViT model, resulting in 342-dimensional representations. Multiple rounds of pretraining were conducted, and the final reported performance corresponds to the averaged results of 10 independent random runs on the test set.

2) Publicly available dataset: The CMU-MOSI dataset is a multimodal sentiment analysis dataset comprising three modalities of data - text, audio, and video-extracted from 93 YouTube movie review videos [23]. The text data is generated

through automatic transcription of the videos, while the audio and video data are directly extracted from the original footage. This dataset consists of a total of 2,199 video clips, each annotated with a sentiment intensity score ranging from -3 to 3. These annotations were completed by Amazon Mechanical Turk workers, who took comprehensive multimodal information into consideration.

Based on established findings in psychological assessment research, emotion prediction in this context is typically treated as a regression-oriented task. In the present study, the F1 score (W-F1) and classification accuracy (ACC) were adopted as the primary performance indicators. W-F1 reflects the model's overall classification capability, while ACC provides an intuitive measure of prediction correctness. The evaluation metrics are defined as follows:

$$\text{W-F1}=\frac{\sum_{a=1}^{R} M_a \times F1_a}{\sum_{a=1}^{R} M_a},\ \text{ACC}=\frac{\sum_{a=1}^{R} M_a \times ACC_a}{\sum_{a=1}^{R} M_a} \quad (33)$$

where, R denotes the total number of emotion categories in the dataset, M_a represents the number of samples belonging to category a, $F1_a$ is the F1 score for category a, and ACC_a is the accuracy score for category a.

C. Baseline Models

- MulT [7]: A bidirectional cross-modal attention mechanism is employed, using a modality-translation strategy to construct directed interactions between paired modalities. This design enables effective learning of cross-modal associative features.
- MISA [8]: A dual-space representation learning framework is introduced, in which distributional similarity, orthogonality constraints, reconstruction loss, and task-specific objectives are jointly optimized. The framework enhances cross-modal consistency while preserving modality-specific characteristics.
- CapTrBERT (Caption Transformer BERT) [17]: Semantic understanding is strengthened by generating auxiliary caption-like sentences and integrating them with the original textual input. This approach enables more comprehensive extraction of global semantic information.
- FITE (Face-sensitive Image-to-Emotional-Text Translation) [24]: A face-expression–to-text transformation strategy is adopted, in which facial expression features are converted into descriptive textual representations. These representations are aligned with target aspects at a fine-grained level, and multimodal information is fused through a gating mechanism for emotion classification.
- MIM (Multimodal Interaction Model) [25]: An interactive learning framework based on cross-modal attention is proposed. Modality-specific subspaces and cross-modal information exchange channels are constructed to facilitate efficient multimodal interaction while reducing feature redundancy.
- TIEMFF (Text Information Enhancement and Multimodal Feature Fusion) [26]: A text-dominant hierarchical enhancement architecture is introduced. A dual attention mechanism is used to strengthen textual features and achieve fine-grained alignment of multimodal representations.

D. Comparative Analysis of Experimental Results

Most of the video segments in the dataset were captured using high-definition devices, resulting in favorable lighting conditions and high visual clarity. Fig. 4 presents a comparison of evaluation metrics between the proposed model and the baseline models, including F1 scores and accuracy.

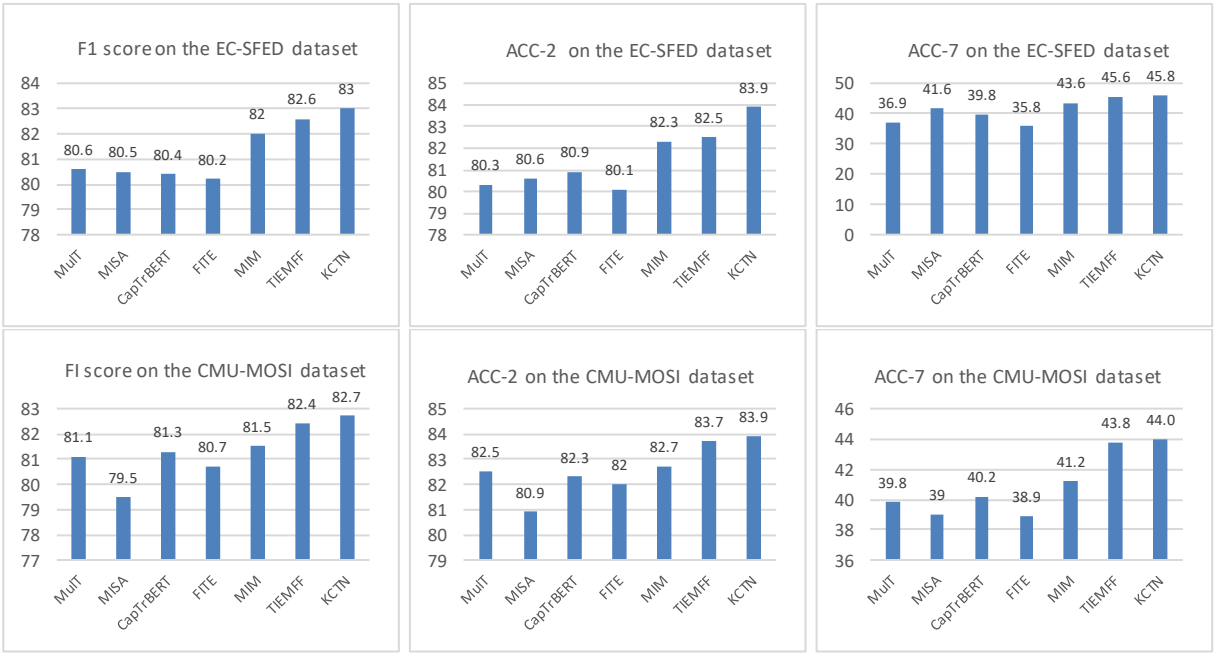

Fig. 4. Evaluation metrics of the proposed model and baseline models on the EC-SFED and the CMU-MOSI datasets.

The comparative analysis of experimental results obtained from the student psychological state dataset indicates that the performance of the FITE model is relatively weak, with both its F1 score and accuracy markedly lower than those of the other models. FITE transforms facial expression features into textual descriptions and subsequently performs multimodal fusion and emotion classification through a gating mechanism. In contrast, the MulT, MISA, and CapTrBERT models demonstrate moderate performance levels across the dataset, with comparable F1 scores and accuracy values. More recent models such as MIM and TIEMFF exhibit notably stronger performance, further validating the effectiveness of cross-modal attention mechanisms in sentiment analysis. The proposed KCTN model achieves the best performance across all evaluation metrics. The results confirm the model's effectiveness in cross-modal interaction and affective feature extraction while highlighting its robustness advantages in student psychological testing across diverse scenarios.

E. Ablation Study

1) Modality ablation: The influence of different modality combinations on the proposed model was examined. Performance was first evaluated using single-modality inputs—text (T), audio (A), and video (V). Dual-modality combinations, including text–audio (T–A), text–video (T–V), and audio–video (A–V), were then assessed. As expected, the ablation results obtained on the EC-SFED and the CMU-MOSI datasets are presented in Table II.

TABLE II. MODALITY ABLATION RESULTS ON THE EC-SFED DATASET

Modality combination	EC-SFED dataset		
ComParE 2016	F1 (%)	ACC-2 (%)	ACC-7 (%)
GeMAPS	80.6	80.3	46.8
	69.2	66.7	37.8
	72.1	71.5	39.8
	82.6	82.7	50.9
	83.3	83.2	51.9
	73.4	72.5	45.6
eGeMAPS	83.8	83.9	52.5

The results demonstrate that the performance of the multimodal affective recognition system is highly sensitive to the choice of modality combinations. In the single-modality experiments, notable disparities were observed across modalities: textual features achieved markedly superior performance (F1 = 80.6%) compared with audio (69.2%) and video (72.1%). The introduction of textual information produced substantial performance gains in dual-modality settings. The text–audio combination increased F1 by 19.36 percentage points relative to audio alone, while the text–video combination improved F1 by 15.53 percentage points. These results confirm the pivotal role of the textual modality in psychological assessment tasks, as its representational capacity surpasses that of auditory and visual modalities. Furthermore, single-modality comparisons revealed that audio conveyed more discriminative information than video, indicating that video data may contain higher levels of noise. The multimodal fusion experiments further demonstrate that combining text, audio, and video yields the highest overall performance. The tri-modal configuration achieved an F1 score of 83.8% on the EC-SFED dataset and 84.1% on the CMU-MOSI dataset, providing strong empirical support for the complementary nature of multimodal integration in psychological assessment scenarios.

2) Module ablation: To examine the contribution and influence of each core module within the overall architecture, a series of module-level ablation experiments was conducted on the two datasets. The notation "w/o" indicates the removal of a specific module during evaluation. Four key modules in the model were ablated sequentially, including the audio–visual enhanced Transformer module (w/o AVET), the information compression and fusion module (w/o ICF), and the long-range emotion fusion module (w/o FFM), enabling analysis of their respective impacts on model performance.

a) Ablation of the audio–visual enhanced transformer module (w/o AVET): In this setting, the audio–visual enhanced Transformer module of the CMET architecture was removed. The non-textual modalities processed by the one-dimensional temporal convolution were fed directly into the information compression and fusion module.

b) Ablation of the information compression and fusion module (w/o ICF): In this configuration, the information compression and fusion module of the CMET architecture was omitted. The modality representations produced by the non-textual enhanced Transformer were fused directly.

3) Ablation of the long-range emotion fusion module (w/o FFM): For this setting, the long-range emotion fusion module was removed. The audio–visual representations obtained from the information compression and fusion module were directly fused with the temporally convolved textual representations.

Table III presents the results of the module ablation experiments conducted on the EC-SFED and the CMU-MOSI dataset.

TABLE III. MODALITY ABLATION RESULTS ON THE EC-SFED AND THE CMU-MOSI DATASET

Model	EC-SFED dataset			CMU-MOSI dataset		
ComParE 2016	F1 (%)	ACC-2 (%)	ACC-7 (%)	F1 (%)	ACC-2 (%)	ACC-7 (%)
GeMAPS	81.3	82.5	50.5	78.0	78.3	47.8
	82.9	82.6	51.4	78.5	78.6	48.2
	82.2	83.3	52.6	80.3	81.2	48.5
	83.8	84.1	53.7	80.8	82.0	50.2

The module ablation results on the EC-SFED dataset demonstrate that the removal of any core module leads to a measurable decline in model performance. Among the examined components, the audio–visual enhanced Transformer module exerts the most substantial influence on overall effectiveness. Its removal reduces the F1 score to 81.3%, decreases ACC-2 to 82.5%, and lowers ACC-7 to 50.5%. This pronounced degradation underscores the central role of the

audio–visual enhanced Transformer in the early stages of multimodal information integration. The removal of the information compression and fusion module and the long-range emotion fusion module also results in significant performance degradation. These reductions highlight their importance in mitigating redundancy within non-textual modalities, enhancing feature robustness, and improving cross-modal discriminability. Ablation experiments conducted on the CMU-MOSI dataset also demonstrate the utility of the three aforementioned modules. Compared with the experimental results on the EC-SFED dataset, removing the long-distance emotion fusion module on the CMU-MOSI dataset leads to a relatively smaller impact on model performance. This suggests that the role of long-range information may be limited on smaller-scale datasets. Conversely, on large-scale datasets, this module can significantly enhance the model's ability to recognize both consistency and divergence in emotional information across different modalities by constructing a greater number of sample pairs, thereby improving the accuracy and effectiveness of sentiment analysis. Collectively, the ablation of any single module results in performance degradation, validating the contribution of each component to the model's overall performance and robustness.

F. Empirical Results and Analysis

To further evaluate the effectiveness of the proposed model in psychological testing, t-SNE was employed to visualize the high-dimensional features extracted from the EC-SFED test set. The results are presented in Fig. 5 and 6. Fig. 5 illustrates the distribution of feature representations before classification. Substantial overlap is observed between different emotional categories, and no clear clustering structure is formed. This indicates that the raw features do not inherently exhibit strong discriminative capability. After processing by the proposed model, the post-classification visualization in Fig. 6 shows that the feature vectors are distinctly clustered into two separable groups corresponding to positive and negative emotional states. The boundary between the clusters becomes pronounced, demonstrating that the model achieves effective multimodal affective feature separation. These empirical findings indicate that the proposed sentiment analysis model exhibits reliable performance when applied to psychological assessment scenarios involving students.

G. Case Analysis of Intelligent Psychological Assessment Interactions

The tasks in AI Heart World are presented within dynamic and immersive interactive scenarios. Participants' responses—including every choice made and action taken within these environments—generate process data with significant reference value for psychological state assessment. Table IV summarizes three representative cases, illustrating how different models handle these multimodal cues. In Case 1, the critical challenge lies in aligning textual information with visual cues from the scenario. The CapTrBERT model combines original inputs with generated auxiliary sentences; however, its architecture does not incorporate an effective mechanism for extracting or aligning emotional features from images. As a result, errors arise in sub-scenarios where textual expressions and visual affective cues are inconsistent. The FITE model establishes an initial degree of text–image integration through its basic cross-modal attention mechanism. In contrast, the proposed model leverages multimodal information captured throughout the intelligent interaction process to accurately identify the test subject's underlying anxious emotional state, which remains implicit beneath a superficially calm textual expression.

Case 2 focuses on a sub-scenario in which the visual information is incomplete, thereby assessing the model's ability to utilize non-facial emotional cues. In this scenario, the participant's face is partially occluded, resulting in the absence of clear facial features. Because the core design logic of the FITE model relies heavily on facial expressions as the primary basis for emotion inference, the absence of key visual features prevents the model from extracting effective supplementary cues from other regions of the image, ultimately leading to a misclassification of the psychological state. In contrast, the proposed model maintains prediction accuracy by leveraging the coordinated effects of the long-range contextual module and the Graph Convolutional Network (GCN), which together enable the extraction of syntactic features.

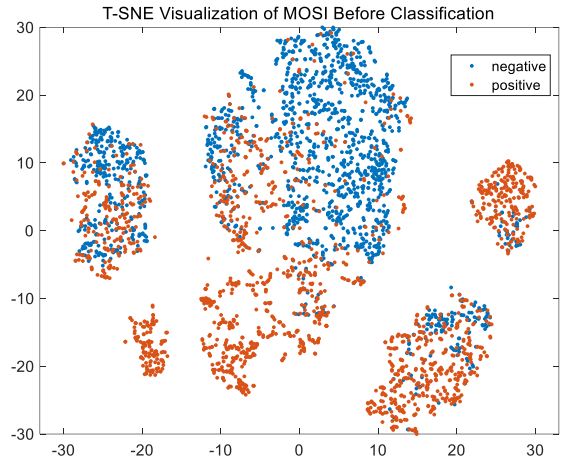

Fig. 5. Visualization of feature vectors before classification.

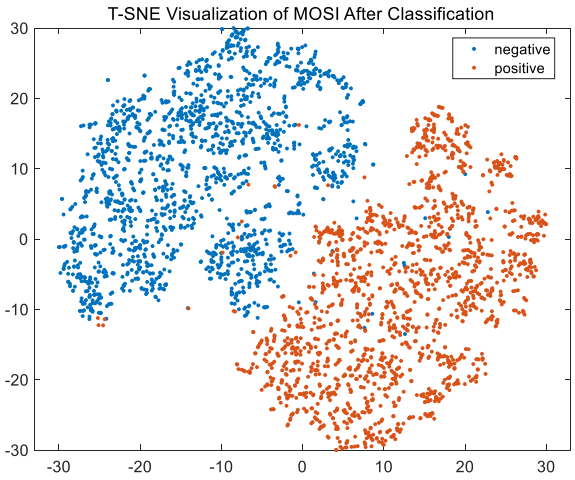

Fig. 6. Visualization of feature vectors after classification.

TABLE IV. CASE ANALYSIS OF PSYCHOLOGICAL ASSESSMENT SCENARIOS

Image				
Feature text		Did I do something wrong? Could you give me a hint?	I don't have enough blocks. They may not be stable.	The whole process was like playing a game. Using the sandbox to create scenes felt very novel, and it was quite interesting how the system analyzed my psychological state based on the 'Bing Dwen Dwen' character I selected.
Long-range text		This is impossible! I don't even know how to assemble the basic structure—forget it, let's move to the next one.	Try to speed up the construction. Use whatever is available first—it'll be finished soon.	Haha, this is kind of fun! Let me try another construction method!
Ground-truth emotion		Negative	Positive	Positive
Model	**FITE**	Negative√	Negative×	Positive√
	CapTrBERT	Positive×	Positive√	Positive√
	Proposed	Negative√	Positive√	Positive√

Case 3 represents an ideal scenario in which multimodal information is highly consistent, designed to validate the effectiveness and rationality of the evaluation benchmark. In this case, the participant's textual response (e.g., "It's quite interesting"), visual indicators (slight smile and relaxed posture), and contextual emotional records exhibit strong coherence, with no conflicting affective cues. Experimental results show that CapTrBERT, FITE, and the proposed model all achieve 100% prediction accuracy in this ideal setting. This outcome not only confirms that the baseline performance of each model meets expectations when sufficient information is provided, but more importantly, it validates the scientific rigor of the selected case set used in this evaluation. By employing a consistent scenario to eliminate external interference, the study successfully pinpoints the genuine performance differences among models under complex conditions, thereby establishing a reliable benchmark for subsequent model optimization.

V. CONCLUSION

A cross-modal fusion network model (KCTN) based on long-range affective integration was proposed for psychological assessment tasks involving university students. Utilizing the "AI Heart World" application case, the model intelligently extracted key information and incorporated a multimodal fusion module enhanced with long-distance knowledge, achieving layer-by-layer integration of textual, audio, and visual features. This approach effectively mined valuable information from weak modalities while suppressing noise interference. In practical psychological assessment tasks, KCTN was capable of handling imbalanced data quality in multimodal inputs. For instance, when audio or video data were partially missing or exhibited a low signal-to-noise ratio, the model maintained stable performance by integrating long-distance knowledge with the textual modality. In complex scenarios such as teacher-student dialogues or psychological counseling sessions, KCTN dynamically adjusted predictions based on contextual and emotional cues of the speaker, thereby improving recognition accuracy. Future research will continue to explore the disparities and complex interactions among different modalities to design more effective multimodal fusion mechanisms. It will also further integrate campus management systems and dialogue information from intelligent interactive platforms to uncover the causal layers underlying students' psychological states, thereby enabling timely detection of adverse mental conditions and contributing to the development of a novel intelligent technology-based paradigm for psychological assessment.

ACKNOWLEDGMENT

This study was funded by Zhejiang Provincial Education Science Planning Project (Research on the Construction of Multi modal Curriculum Resources and Personalized Teaching in Vocational Colleges Empowered by Embodied Intelligence, Grant No.: 2025SCG340); and Zhejiang Provincial Department of Education Project (Research on the Development and Validation of an Intelligent Monitoring System for College Students' Mental Health Integrating DT-IRTI Technology, Grant No.: Y202456020).

REFERENCES

[1] K. Wang, Y. Zhang, "Topic sentiment analysis in online learning community from college students," J. Data Inf. Sci., vol. 2, pp. 33–61, 2020.

[2] A. Iram, "Sentiment analysis of student's facebook posts," in *Int. Conf. Intell. Technol. Appl.*, pp. 86–97, 2018.

[3] M. Zhang, S. Wang, K. Yuan, "Sentiment analysis of barrage text based on ALBERT and multi-channel capsule network," in *Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov.*, pp. 718–726, 2021.

[4] Y. F. Song, G. Ren, Y. Yang, X. C. Fan, "Multimodal sentiment analysis based on hybrid feature fusion of multi-level attention mechanism and multi-task learning," Appl. Res. Comput., vol. 39, no. 3, pp. 716–720, 2022.

[5] L. Sun, Z. Lian, B. Liu, J. Tao, "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," IEEE Trans. Affect. Comput., vol. 15, no. 1, pp. 309–325, 2023.

[6] R. Huan, G. Zhong, P. Chen, R. Liang, "Unimf: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences," IEEE Trans. Multimed., vol. 26, pp. 5753–5768, 2023.

[7] Y. H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, R. Salakhutdinov, "Multimodal transformer for unaligned multimodal

language sequences," in *Proc. Assoc. Comput. Linguist.*, 2019, pp. 6558–6581.

[8] D. Hazarika, R. Zimmermann, S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimed.*, Seattle, WA, USA, pp. 1122–1131, 2020.

[9] W. Han, H. Chen, S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," arXiv preprint arXiv:2109.00412, 2021.

[10] M. G. Huddar, S. S. Sannakki, V. S. Rajpurohit, "Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM," Multimed. Tools Appl., vol. 80, no. 9, pp. 13059–13076, 2021.

[11] J. H. Luo, Y. Zhu, "Multi interaction perception network for sentiment analysis of misaligned multimodal language sequences," Comput. Appl., vol. 44, no. 1, pp. 79–85, 2024.

[12] L. Yang, J. H. Zhong, Y. Zhang, X. Y. Song, "Temporal multimodal sentiment analysis with composite cross modal interaction network," Comput. Sci. Explor., vol. 18, no. 5, pp. 1318–1327, 2024.

[13] P. Zhong, D. Wang, C. Miao, "Knowledge-enriched transformer for emotion detection in textual conversations," arXiv preprint arXiv:1909.10681, 2019.

[14] H. B. U. Haq, W. Akram, M. N. Irshad, A. Kosar, M. Abid, "Enhanced Real-Time Facial Expression Recognition Using Deep Learning," Acadlore Trans. AI Mach. Learn., vol. 3, no. 1, pp. 24–35, 2024.

[15] S. Nadar, D. Gandhi, A. Jawale, S. Pawar, R. Prabhu, "Multimodal Audio Violence Detection: Fusion of Acoustic Signals and Semantics," Acadlore Trans. AI Mach. Learn., vol. 4, no. 4, pp. 301–311, 2025.

[16] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, Y. Choi, "COMET: Commonsense transformers for automatic knowledge graph construction," arXiv preprint arXiv:1906.05317, 2019.

[17] Z. Khan, Y. Fu, "Exploiting BERT for multimodal target sentiment classification through input space translation," in *Proc. 29th ACM Int. Conf. Multimed.*, pp. 3034–3042, 2021.

[18] F. Eyben, M. Wöllmer, B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimed.*, Firenze, Italy, pp. 1459–1462, 2010.

[19] Y. Zhang, *Research on emotion analysis based on cross modal enhanced transformer fusion network*, Chang'an Univ., 2024.

[20] J. Chen, H. Hou, J. Gao, Y. Ji, T. Bai, "RGCN: Recurrent graph convolutional networks for target-dependent sentiment analysis," in *Int. Conf. Knowl. Sci. Eng. Manag.*, pp. 667–675, 2019.

[21] J. Jiang, A. Wang, A. Aizawa, "Attention-based relational graph convolutional network for target-oriented opinion words extraction," in *Proc. 16th Conf. Eur. Chap. Assoc. Comput. Linguist. Main Vol.*, pp. 1986–1997, 2021.

[22] K. Q. Huang, Y. X. Kang, C. X. Yan, et al., "A review of intelligent psychological assessment based on interactive environment," Chin. J. Ment. Health, vol. 39, no. 4, pp. 337–343, 2025.

[23] A. Zadeh, R. Zellers, E. Pincus, L. P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," arXiv preprint arXiv:1606.06259, 2016.

[24] H. Yang, Y. Zhao, B. Qin, "Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis," in *Proc. 2022 Conf. Empir. Methods Nat. Lang. Process.*, pp. 3324–3335, 2022.

[25] Y. Luo, R. Wu, J. Liu, X. Tang, "Balanced sentimental information via multimodal interaction model," Multimed. Syst., vol. 30, no. 1, p. 10, 2024.

[26] Z. Liu, L. Cai, W. Yang, J. Liu, "Sentiment analysis based on text information enhancement and multimodal feature fusion," Pattern Recognit., vol. 156, p. 110847, 2024.